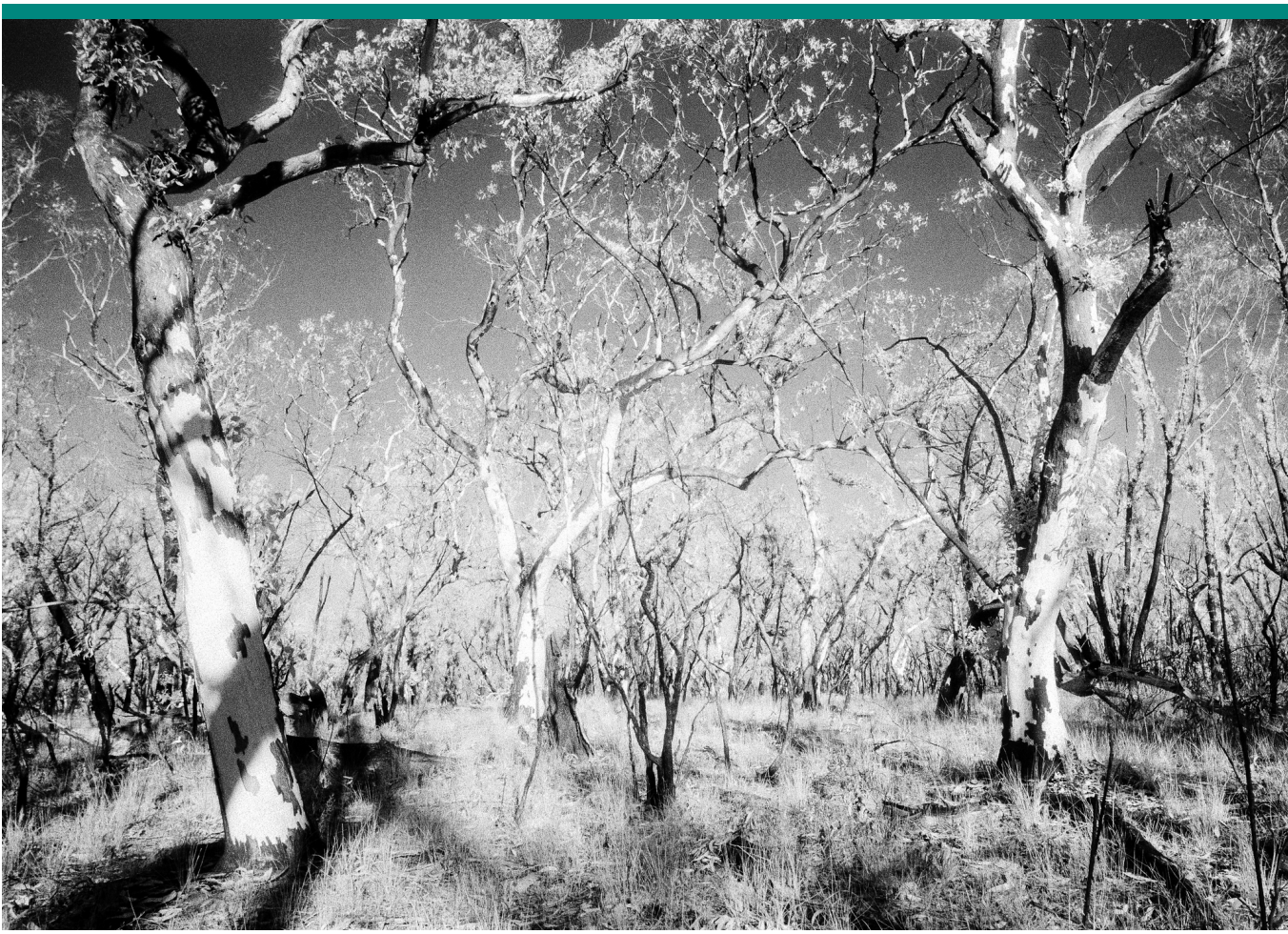
MEANING, DECISION, & NORMS

THEMES FROM THE WORK OF ALLAN GIBBARD



Edited by Billy Dunaway and David Plunkett

Meaning, Decision, and Norms:

Themes from the Work of Allan Gibbard

Meaning, Decision, and Norms:

Themes from the Work of Allan Gibbard

Eds. Billy Dunaway and David Plunkett

Maize Books

Published in the United States of America by
Michigan Publishing
Manufactured in the United States of America

Cover photograph © David Braddon-Mitchell

Cover design by Rachel Dunaway

DOI:http://doi .org/10.3998/mpub.9948199

ISBN978-1- 60785-464-7 (paper)
ISBN978-1- 60785-465-4 (e-book)
ISBN978-1- 60785-707-5 (open access)

An imprint of Michigan Publishing, Maize Books serves the publishing needs of the University of Michigan community by making high- quality scholarship widely available in print and online. It represents a new model for authors seeking to share their work within and beyond the academy, offering streamlined selection, production, and distribution processes. Maize Books is intended as a complement to more formal modes of publication in a wide range of disciplinary areas.

https://www.maizebooks.org

Contents

Contributors

Sara Aronowitz is an Assistant Professor of Philosophy at University of Arizona.

Simon Blackburn is a Fellow of Trinity College, Cambridge; Research Professor of Philosophy at UNC Chapel Hill; and Visiting Professor of Philosophy at the New College of the Humanities.

Paul Boghossian is the Silver Professor of Philosophy at New York University.

David Braddon-Mitchell is a Professor of Philosophy at University of Sydney.

Nate Charlow is an Associate Professor of Philosophy at University of Toronto.

Stephen Darwall is the Andrew Downey Orrick Professor of Philosophy at Yale University and the John Dewey Distinguished University Professor Emeritus at University of Michigan, Ann Arbor.

Jamie Dreier is the Judy C. Lewent and Mark L. Shapiro Professor of Philosophy at Brown University.

Billy Dunaway is an Associate Professor of Philosophy at University of Missouri—St. Louis.

Melissa Fusco is an Assistant Professor of Philosophy at Columbia University.

Sona Ghosh is a Research Analyst at Laurentian University.

Allan Gibbard is the Richard B. Brandt Distinguished University Professor of Philosophy Emeritus at University of Michigan, Ann Arbor.

William L. Harper is a Professor Emeritus of Philosophy at University of Western Ontario.

Paul Horwich is a Professor of Philosophy at New York University.

Zoë Johnson King is an Assistant Professor of Philosophy at Harvard University.

Tristram McPherson is a Professor of Philosophy at The Ohio State University.

Howard Nye is an Associate Professor of Philosophy at University of Alberta.

Lauren Olin is an Assistant Professor of Philosophy at the University of Missouri—St. Louis.

Caleb Perl is a Lecturer in Philosophy at Australian Catholic University.

David Plunkett is an Associate Professor of Philosophy at Dartmouth College.

Peter Railton is the Gregory S. Kavka Distinguished University Professor and the John Stephenson Perrin Professor of Philosophy at University of Michigan, Ann Arbor.

Connie Rosati is a Professor of Philosophy at University of Texas, Austin.

Mark Schroeder is a Professor of Philosophy at University of Southern California.

Alex Silk is an Associate Professor in Philosophy at University of Birmingham.

Daniel J. Singer is an Associate Professor of Philosophy at University of Pennsylvania.

Brian Skyrms is a Distinguished Professor of Logic and Philosophy of Science and Economics at University of California, Irvine, and a Professor of Philosophy at Stanford University.

Seth Yalcin is a Professor of Philosophy at University of California, Berkeley.

Acknowledgments

We first had the idea for a volume of this kind in 2014 on a walk together in England. Since then, our vague idea for a volume has taken concrete shape, and a number of people have helped make this possible. We would like to thank everyone who has helped make it a success. First, we'd like to thank all of the contributors to this volume. Earlier drafts of three papers in this volume (the papers by Simon Blackburn, Paul Boghossian, and Paul Horwich) were originally presented at a conference organized by the University of Michigan Philosophy Department in honor of Allan Gibbard, May 12–13, 2016. So, second, we would like to thank the organizers of that conference for their work. Third, every paper in this volume was peer reviewed. We'd like to thank the (anonymous) twelve philosophers who signed on to refereeing papers for this volume, and for their helpful guidance and feedback on the papers. Fourth, we'd like to thank everyone at Maize Books and Michigan Publishing at the University of Michigan for taking on this project, and helping us publish this volume as an online, open-access volume. Fifth, we'd like to thank the Dartmouth College Ethics Institute, the Dartmouth College Philosophy Department, the John Templeton Foundation, the University of Missouri-St. Louis Philosophy Department, and the University of Michigan Philosophy Department for financial support in helping publish this volume. Sixth, we'd like to thank our research assistants who have helped us during this project: Daniel Gun Lim, Josh Petersen, Ira Richardson, and Michael Tofte. Seventh, we'd like to thank David Braddon-Mitchell for providing a photo for the cover of the volume, and Rachel Dunaway for designing the cover. Finally, we'd like to thank Allan Gibbard for agreeing to be a part of this project, and for his enthusiastic and thoughtful engagement with all of the papers in this volume. Both of us benefited enormously from working with Allan as PhD students at the University of Michigan. This project has been a wonderful way for us to spend more time thinking through his work, and, in the process, creating a volume that we hope will be useful for a wide range of philosophers, across a range of subfields.

Abstracts

1. Decision dynamics and rational choice by William L. Harper, with Appendix 1: *Dynamics Calculations* by Sona Gosh, Appendix 2A: *Death in Damascus in Continuous Time Dynamics* by Brian Skyrms, and Appendix 2B: *Death in Damascus in Tempered Discrete Time Dynamics* by Brian Skyrms.

My contribution to this volume is a review of arguments and issues raised by examples where causal decision theory results in decision instability. Appendix (1) by Sona Ghosh gives some results from applications of computer resources to discrete decision dynamics for such problems. These include the surprising result that the utilities for choosing shoot, for choosing don't shoot, and for choosing the ratifiable mixed strategy M, at the deliberational fixed point probabilities for shoot and don't shoot in Joyce's version of Egan's murder lesion example generate choosing the ratifiable mixed strategy M, even though all three have equal causal utilities. Appendix (2A) by Brian Skyrms gives a general convergence theorem for applications of continuous decision dynamics to such problems. Appendix (2B) by Brian Skyrms gives a convergence result for tempered discrete time dynamics for such problems.

2. Counterfactuals and the Gibbard-Harper Collapse Lemma by Melissa Fusco.

Gibbard & Harper (1978) provides a classic statement of Causal Decision Theory ("CDT"), which uses counterfactual conditionals to express the causal relationships that are, according to CDT, of particular relevance to rational decision-making. Classic CDT—in particular, the Gibbard-Harper formulation of it—has been challenged by an influential 2007 paper by Andy Egan (Egan 2007), which presents several counterexamples to the theory. On Egan's telling, causal decision theorists adhere to the motto "do whatever has the best expected outcome, holding fixed your initial views about the likely causal structure of the world" (96). However, Egan argues, there are cases where agents should not hold such initial views fixed as they act. In such cases, agents should use their anticipated future causal views instead, taking into account what they expect to learn by performing the very act in question.

In this paper, I focus on the dialectic from the CDTer's point of view, with an eye to a formal result pointed out by Gibbard & Harper in the third section of their classic paper. There, they show that if an agent's credences are probabilistically coherent, and the semantics for counterfactuals obeys Strong Centering—roughly, the view that each possible world is counterfactually closest to itself—then the probability of (a counterfactual conditional on its antecedent) simplifies to the probability of (its consequent, given its antecedent). This has the eyebrow-raising consequence that "Eganized" causal decision theory, the view on which agents anticipate their future causal views, recommends an act just in case classical evidential decision theory does. The "collapse", as I call it, complicates the traditional way of glossing the relationship between EDT, CDT, and diachronic coherence norms.

3. Who's Afraid of Normative Externalism? by Zoë Johnson King.

This paper discusses a puzzle for agents who are not only uncertain which first-order moral theory is true but also uncertain about the true view of how one should act when one is uncertain which first-order moral theory is true. Many in the literature on moral uncertainty—myself included—find plausible the view that uncertain agents should act so as to maximize expected objective moral value. And, helpfully, this view can apply to higher-order as well as to first-order uncertainty. However, a puzzle arises for expected-objective-value-maximizers who have some non-zero credence in *normative externalism*: a view that holds that moral uncertainty is morally irrelevant and that the way one should act when one is uncertain which first-order moral theory is true is simply to perform the action that is, in fact, required by the true first-order moral theory. If a first-order-uncertain expected-objective-value-maximizer has any non-zero credence in normative externalism, then she is in a pickle, as she cannot fill out the corresponding column of her higher-order decision table and thus can no longer calculate her action's higher-order expected objective value. It is disingenuous to respond by simply ignoring the possibility that normative externalism is true. But, I argue, expected-objective-value-maximizers can do better than that. For there is a strategy we can use—what I call the *repartitioning strategy*—to take account of all of the possibilities that we can fathom about how normative externalism might evaluate our available acts. This strategy generates determinate rankings and choice sets even for higher-order uncertain agents with some credence in normative externalism. I then show that, curiously but fortunately, an uncertain agent who follows the repartitioning strategy can shift any amount of credence back and forth between normative externalism and the view that uncertain agents should act so as to maximize expected objective moral value, without her choice set, her actions' rankings, or even the higher-order expected objective value of any of her available acts changing at all. This is good news for expected value maximizers: on principled grounds, we can pretty much just ignore—and certainly need not be afraid of—normative externalism.

4. What Epistemic Reasons Are For: Against the Belief–Sandwich Distinction by Daniel J. Singer and Sara Aronowitz.

The standard view says that epistemic normativity is normativity of belief. If you're an evidentialist, for example, you'll think that all epistemic reasons are reasons to believe what your evidence supports. Here we present a line of argument that pushes back against this standard view. If the argument is right, there are epistemic reasons for things other than belief. The argument starts with evidentialist commitments and proceeds by a series of cases, each containing a reason. As the cases progress, the reasons change from counting in favor of things like having a belief to things like performing ordinary actions. We argue that each of those reasons is epistemic. If the argument succeeds, we should think there are epistemic reasons to consider hypotheses, conduct thought and physical experiments, extend one's evidence, and perform mundane tasks like eating a sandwich, just as there are epistemic reasons to believe what one's evidence supports.

5. Assessing Feelings by Simon Blackburn.

In my paper I attempt to allay Allan Gibbard's doubts about whether the phenomena of 'normative governance' can be explained on what I call a Humean basis. I argue that the kinds of worry that indeed beset us when we doubt the justifiability of our own emotional reactions to things are more widely spread across many forms of reaction other than the strictly normative, and that they can be understood in terms of a self-consciousness that is explicable by our social natures, without invoking any sui generis normative governance.

6. A Gibbardian Account of (Narrow) Moral Concepts by Stephen Darwall.

Allan Gibbard distinguishes between narrow and wide conceptions of morality in a way that tracks Bernard Williams's distinction between morality and ethics. Like Williams, Gibbard ties the narrower conception—what we might call *deontic morality*—to accountability, specifically, to justified attitudes of guilt and blame through which we hold one another and ourselves accountable for complying with what morality requires. But not all notions of deontic morality are explicitly deontic in this way. This chapter explores how moral notions that do not seem explicitly deontic—like the notions of moral choiceworthiness, moral reasons, and moral goodness in the sense of being a suitable object of moral esteem, might best be accounted for within a Gibbardian conceptual framework.

7. Morality and the Bearing of Apt Feelings on Wise Choices by Howard Nye.

It is often assumed that the best explanation of why we should be moral must involve a substantive account of what there is reason to do and how this is related to what morality requires and recommends. In this paper I argue to the contrary that the best explanation

of why we should be moral is neutral about the content of morality, and does not invoke an independent substantive account of what there is practical reason to do. I contend that an act's deontic status as recommended or required by morality is best understood as its being fitting for us to feel obligated to perform it, which essentially involves motivation to perform it. I argue, moreover, that our having reason to do something is a matter of its being fitting for us to be motivated to do it. Since an act's being favored by morality conceptually entails the fittingness of our being motivated to perform it, and the fittingness of this motivation conceptually entails that there is reason to perform it, it is actually a conceptual truth that there are reasons to do what morality requires and recommends, whatever that turns out to be. I contend, finally, that this kind of account best explains why, although moral considerations are not always overriding, we necessarily have conclusive reasons to do what morality requires. I argue that an act counts as morally required only if the reasons to feel obligated to perform it are conclusive, which entails that it is unfitting to fail to be most strongly motivated to perform it. This, together with my account of the connection between fitting motives and practical reasons, entails that whatever considerations are weighty enough to make the act morally required are conclusive reasons to perform it. I believe that this conceptual account of reasons to be moral is important, because it removes the explanation of why we should be moral as a desideratum on normative ethical theories, which may significantly decrease the attractions of some and increase the attractions of others.

8. Expressivism without Minimalism by Tristram McPherson.

Many of the most recently influential expressivists in metaethics have embraced minimalist interpretations of certain seemingly metaphysically significant locutions such as 'truth,' 'fact,' and 'belief.' And others have argued that this marriage to minimalism is crucial to the overall plausibility of expressivism as an interpretation of our actual normative thought and talk. This paper argues against the idea that metaethical expressivists need to commit themselves to minimalism. I show that the dialectical costs to the expressivist of marriage to minimalism are higher, and the costs of divorce lower, than has typically been appreciated, and sketch a theory of error to account for the considerations that appear to force the expressivist toward minimalism.

9. Metasemantic Quandaries by Nate Charlow.

This paper advocates a generalized form of Expressivism, as a strategy for resolving certain metasemantic puzzles about identifying the semantic value of a context-sensitive expression in context. According to this form of Expressivism, speakers express properties of semantic parameters, and they do so in order to proffer those properties for cognitive adoption (acceptance) by their addressees. Puzzles arising from the pressure to say what a putatively context-sensitive expression refers to or denotes in contexts that do not seem to specify

a referent or denotation dissolve, once we appreciate that such attempts were ill-placed to begin with. Gibbard's Norm Expressivism, according to which speakers express properties of planning states or normative systems, is a branch of more general theory (although, I will argue, Gibbard takes on commitments—optional for the Expressivist—that make it a bit hard to see how to distinguish his theory from a nuanced form of Subjectivism).

10. Weak and Strong Necessity Modals: On Linguistic Means of Expressing "A Primitive Concept OUGHT" by Alex Silk.

This paper develops an account of the meaning of 'ought' and the distinction between weak necessity modals ('ought', 'should') and strong necessity modals ('must', 'have to'). I argue that there is nothing specially "strong" about strong necessity modals per se. Uses of 'Must ϕ' predicate the (deontic/epistemic/etc.) necessity of the prejacent ϕ of the actual world (evaluation world). The apparent weakness of weak necessity modals derives from their bracketing whether the necessity of the prejacent is verified in the actual world. 'Ought ϕ' can be accepted without needing to settle that the relevant considerations (norms, expectations, etc.) that actually apply verify the necessity of ϕ. I call the basic account a *modal-past approach* to the weak/strong necessity modal distinction (for reasons that become evident). Several ways of implementing the approach in the formal semantics/pragmatics are critically examined. The account systematizes a wide range of linguistic phenomena: it generalizes across flavors of modality; it elucidates a special role that weak necessity modals play in discourse and planning; it captures contrasting logical, expressive, and illocutionary properties of weak and strong necessity modals; and it sheds light on how a notion of 'ought' is often expressed in other languages. These phenomena have resisted systematic explanation. In closing I briefly consider how linguistic inquiry into differences among necessity modals may improve theorizing on broader philosophical issues.

11. How to Outfox Sly Pete: A Picture of the Pragmatics of Indicatives by Caleb Perl.

This paper develops a novel account of a central puzzle about indicative conditionals. Allan Gibbard noted that different speakers are sometimes perfectly entitled to accept indicatives with the same antecedent and incompatible consequents. Standard contextualist semantics for indicatives (like Robert Stalnaker's or Angelika Kratzer's) seem to make implausible predictions about those cases. Gibbard used those problems to motivate an expressivist account of indicatives, where indicatives semantically express the speaker's state of mind. Others have used this puzzle to motivate relativist accounts, where indicatives express propositions whose truth-conditions are relative to novel points of assessment. This puzzle is interesting in large part because of its significance for foundational questions in the philosophy of language: as evidence for an expressivist approach, or a relativist approach, or some other heterodox approach.

This paper argues that Gibbard's puzzle is not a license for optimism for the expressivist or the relativist. It has three central goals. The first central goal is to show that Gibbard's puzzle is even harder to answer than most philosophers think. The second central goal is to introduce a novel account of Gibbard's puzzle, which posits new presuppositions, and to show that this account captures all facets of the puzzle. I intend this presuppositional account as a defense of traditional contextualism. The presuppositional account shows how a traditional contextualist can get the puzzle just right. And the third central goal is to explore which heterodox semantic theorists can give an equally good explanation of the puzzle.

12. Modeling with Hyperplans by Seth Yalcin.

I explore Gibbard's idea of a hyperplan. First I give my take on how Gibbard puts hyperplans to work—how he takes them to help in modeling normative states expressivistically, and how he situates hyperplans philosophically in a theory of content. Gibbard models normative content with the help of hyperplans. He also explains his plan-laden content with the help of primitive appeal to normative mental states. I consider the path of modeling with Gibbard-like hyperplans while eschewing Gibbard's style of explanation of them. Down this road I find expressivisms that can get what is prototypically wanted out of the view, but which otherwise cohere with broadly representationalist conceptions of the mental.

13. Convergence in Plan by Mark Schroeder.

When judgments are plan-laden, that can help to explain why reasonable and intelligent people can still diverge over those judgments. For in such cases, nothing ultimate decides between two alternative plans except for the planners themselves. In metaethics, that can help us to understand why competent speakers and thinkers can deeply disagree over moral questions, in the theory of rationality it can explain why the meaning of 'rational' does not answer for us how to regard Savage's "sure thing" principle, and in the theory of linguistic meaning, it can explain how linguistic meaning runs beyond the facts that ground it. Yet in some cases we observe and think that it is reasonable to expect wide divergences in plan-laden thoughts, while in other domains we observe and expect relatively little. This paper pursues the question of what to make of the expressivist explanation of disagreement if plan-laden thought diverges less than the initial prospects for expressivism would lead us to believe.

14. Comic Disagreement by Lauren Olin.

Some disputes about the funny are, apparently, faultless; sometimes when people's comic judgments conflict, "we are not inclined to say that anybody is in error" (Egan 2014: 74). At the same time, some disputes about the funny seem genuine: jokes can and do serve as loci

for serious disagreements, and people are on occasion severely censured for their attempts at humor, or their expressions of amusement. While normative disagreement has received substantial philosophical attention in moral, epistemic, and semantic domains, comic disagreement has been neglected. This is unfortunate, since it is plausible to suppose that facts about comic disagreement are relevant to understanding normative disagreement more generally. This essay first argues that comic judgments, like moral, epistemic, and semantic judgments, are normative. It then articulates and defends an account of comic normativity and disagreement that might be usefully deployed to analyze disagreement in other normative domains. In conclusion it sketches an associated semantic framework, and suggests that thinking about apparently faultless comic disputes can help Gibbard in his efforts to reduce properly normative beliefs to states of planning.

15. Expressivism and Objectivity by Peter Railton.

The idea that normative judgments have as their primary meaning the *expression* of attitudes, rather than the *statement* of propositions capable of ordinary truth or falsity, received its first systematic statements in the 20th century in the form of "emotivism", as developed by A. J. Ayer and C. L. Stevenson. Originally seen as a form of "non-cognitivism", this general approach to normative judgment has evolved to encompass more complex—and more cognitive—attitudes, including the acceptance of norms or plans, as developed by Allan Gibbard. Even truth—at least, in a minimal sense—has found its way back into "expressivism", as the view is now called. Throughout this evolution, however, there has been a tendency on the part of critics and advocates alike to see expressivism as a close cousin to *subjectivism*, and well-suited for those who would deny that *objectivity* is possible in normative matters. Yet far from being natural kin, expressivism constitutes a *denial* of orthodox subjectivism, and is an interpretation of normative judgment especially well-suited to the staunch objectivist about value, rationality, or morality. Gibbard has appreciated this, and his case for expressivism does not depend at all upon subjectivist doctrine or spirit. Instead, he has insisted that we should not be intimidated out of our aspirations to objectivity about what kinds of lives are most worth living.

16. The Metaphysical Conception of Realism by Billy Dunaway.

This paper presents an account of realism in metaphysical terms. The central concept is that of *relative metaphysical fundamentality*: realist views, in general, hold that their subject matter is more fundamental than competing views of the subject matter. I argue that both by considering prominent examples of realism from the philosophical tradition, and by accounting for the structural features of realism, the relative fundamentality view fares better than the usual characterizations of realism in the literature. This discussion sheds light on the *quasi-realist* program developed by Allan Gibbard and others: quasi-realists show

that expressivist tools are sufficient to capture many of the trappings of realism. But quasi-realism does not entail that the domain in question is highly fundamental, and this leaves an additional challenge for those quasi-realists who wish to claim that there is no difference between quasi-realism and genuine realism.

17. The Normativity of Meaning Revisited by Paul Boghossian.

Kripke has argued that the relation between the meaning of a word and its future applications is normative not descriptive. This paper revisits the question in what sense Kripke's thesis is true, in light of Gibbard's recent attempts to defend a particular interpretation of it. It begins by clarifying my own position as laid out in previous papers: while meaning is plausibly normative in the relatively lightweight sense of being inextricably tied to some notion of correct application, it's not plausibly normative in the full-throated sense of being analytically subjective ought-entailing. Against Gibbard's argument, that a subjective ought claim does follow from the meaning of "nothing," it contends that the ought claim in question is much more plausibly regarded as flowing from the normativity of belief, rather than from that of either meaning or mental content. Furthermore, it argues that Gibbard's argument relies on the independently implausible theory of belief as consisting in the acceptance of sentences of one's public language. Returning to the lightweight correctness-based version of the normativity thesis, it argues that Kripke has conflated it with the distinct thesis that grasp of the meaning of an expression can justify its use. The problem of explaining how meaning can have such a justificatory role is challenging and may well require for its solution an appeal to the notion of intuition.

18. Obligations of Meaning by Paul Horwich.

Is the word "meaning" a *normative* term, one (like "ought", "evil", "beautiful", and "delicious") with a prescriptive or evaluative function? Or is it a purely *naturalistic* term (like "electron", "big", "kill", and "red")? And, regarding the *phenomenon* of a given word possessing its particular meaning: is such a fact normative 'all the way down'? Or is it ultimately and entirely constituted by the word's naturalistic properties – including, perhaps, that the word tends to be used in conformity to a certain distinctive, naturalistically characterized regularity? The following discussion of these related questions will focus on Allan Gibbard's answers to the—answers he began to develop in the early 1980s and which he elaborates with characteristic insight, subtlety and power in his *Meaning and Normativity*. The paper proceeds by:

- articulating the core of Gibbard's position (that "meaning" is a normative term)
- addressing the question of what it is for a concept (and the word expressing it) to be *normative*

- questioning the plausibility of his view that we can infer the normativity of "meaning" from the fact that any attribution of meaning to a term entails a proposition concerning how the term *ought* to be used
- spelling out some uncontroversial constraints on an adequate theory of how the meaning-properties of words are constituted at an underlying level
- arguing that these adequacy conditions can be met only if the underlying properties are fully naturalistic tendencies (or dispositions, or *ceteris paribus* laws) of word-use
- and resisting Gibbard's contention that *naturalistic* conceptions of how meaning is grounded lead to a rampant indeterminism, which requires us, in the case of many words, to admit that their meanings are indeterminate, even when we have the compelling intuition that they are certainly not.

19. The Normative Explanation of Normativity by Jamie Dreier.

What does it mean for metaethics if meaning is normative? That is the question this chapter explores. Mostly it asks what happens to *expressivism* if meaning is normative. In the first section I try to say in a theoretically neutral way what it means for meaning to be normative, and what reasons there might be to think that it is. Then until section 6 I proceed on the assumption of the normativity of meaning, and ask what follows; only in that last section do I return to the question of whether it is true.

20. Gibbard on Reconciling Our Aims by Connie S. Rosati.

In his Tanner Lectures, *Reconciling Our Aims*, Allan Gibbard's claim that his metaethical account of our moral inquiry, as presented in *Thinking How to Live*, may make some answers to our normative questions "more plausible than others" (33). He claims, more specifically, that it may make utilitarian answers more plausible than non-utilitarian answers. This chapter explores a number of important ideas from Gibbard's lectures. In particular, it explores and raises puzzles for Gibbard's account of moral inquiry and his attempt to effectuate a connection between his metaethics and the kind of utilitarian normative theory that he has long favored. As the chapter explains, Gibbard's account doesn't seem to capture either the phenomenology or the normativity of moral inquiry, and even if his account of moral inquiry were correct, Gibbard's story about how that account might favor utilitarian answers to our normative questions may only be persuasive to those who are already inclined to accept utilitarianism.

21. Freedom and Direct Binding Consequentialism by David Braddon-Mitchell.

Alan Gibbard argues that it follows from various of his views, that a kind of act consequentialism is the right normative ethic. For this to be good news for much of his philosophy,

act consequentialism needs to be independently acceptable, lest this be used as a *modus tollens* against those views. This paper develops a new way of understanding direct act consequentialism that will provide the same evaluations of the rightness of acts as indirect disposition, motive, or character consequentialism, thus reconciling the coherence of direct consequentialism with the plausible results-in-cases of indirect consequentialisms. This is achieved by seeing that adopting certain kinds of moral dispositions causally constrains our future acts and limits our future freedom so that the maximizing acts that are ruled out by the disposition can no longer be chosen, and we do the right thing in doing the best we can.

Introduction to Meaning, Decision, and Norms*:*

Themes from the Work of Allan Gibbard

Billy Dunaway and David Plunkett

It is not an exaggeration to say that Allan Gibbard is one of the most significant contributors to philosophy over the last five decades. We intend this volume both as a tribute to this work and as a cutting-edge work in the field that engages with it. In putting this volume together, we have aimed to reflect the scope and significance of Gibbard's contributions. The scope of Gibbard's work is evident from the sections in this volume, which we summarize below. As we discuss, Gibbard's work covers an impressive number of subfields within philosophy, including ethics, philosophy of language, decision theory, epistemology, and metaphysics. It also engages with, and makes significant contributions to, work from the natural and social sciences (e.g., evolutionary psychology and economics). The significance of Gibbard's work is reflected in a second aspect of the present volume. The philosophers who have agreed to publish their work in this volume range from some of the most influential senior philosophers in the field (many of whom have long been interlocutors for Gibbard) to younger philosophers who are just beginning their promising careers. There is a final aspect of this volume that speaks to the significance of Gibbard's work as well. This volume is not a collection of artifacts from past decades of philosophy. Instead, it is a collection of essays that each make a significant contribution to contemporary work in philosophy. This reflects the fact that Gibbard's work has not only had a massive influence on past discussion in philosophy but also continues to influence new directions of philosophical research.

The sections of this volume are: I. Norms in Decision and Belief; II. Warranted Feelings; III. Expressivism, Normative Language, and Semantics; IV. Disagreement, Objectivity, and Realism; V. The Normativity of Meaning; and VI. Consequentialism. In some cases, there

are multiple topics covered by a section heading, which we will point out as we elaborate in more detail on Gibbard's work in each area below.

The papers in section I build upon Gibbard's work in decision theory and epistemology. Gibbard's most discussed paper in decision theory, coauthored in 1978 with Bill Harper, is "Counterfactuals and Two Kinds of Expected Utility". Bill Harper, along with Brian Skyrms and Sona Ghosh, expands on the core ideas from this classic paper, while taking cues from Gibbard's 1980 paper "Two Recent Theories of Conditionals" plus "Weakly Self-Ratifying Strategies", published in 1992. Subsequent literature has classified the Gibbard-Harper view put forth in the paper as a formative statement of "causal" decision theory, though (as Gibbard emphasizes in his replies at the end of this volume) the foundation of the theory is a certain kind of counterfactual about possible outcomes. Regardless of the proper name for the theory, much has been written on its costs and benefits compared to alternatives. The most frequently cited alternative is "evidential" decision theory, and Melissa Fusco critically engages with Gibbard's contributions in the ongoing debate between causal and evidential decision theorists.

Section I also includes papers that deal with the normative question not of what one should do but of what one should *believe*. Gibbard has had much to say on this normative topic in epistemology over the course over his career, including discussions of what it is to engage in thought and talk about this normative question. Zoë Johnson King brings Gibbard's work on these topics to bear on a recent issue that has received significant attention, namely the norms governing higher-order uncertainty about one's own normative beliefs. Gibbard's 2007 paper "Rational Credence and the Value of Truth" aims to clarify the slogan "belief aims at truth". The framework he develops is an extension of the expected utility framework from decision theory, with truth as the central value. But true belief, in the view Gibbard arrives at, is not valuable for its own sake but rather for the guidance it provides. Daniel Singer and Sara Aronowitz explore this idea by developing the conclusion that the popular distinction between reasons for action and reasons for belief is not a deep one.

Gibbard's first book *Wise Choices, Apt Feelings* (published in 1990) develops a systematic expressivist theory of normative judgment. Roughly, on this kind of view, normative judgments are (at the most explanatorily basic level) explained as consisting in "noncognitive" mental states (e.g., desires, planning states, emotions, etc., as opposed to beliefs, traditionally understood). In turn, normative talk is understood in terms of the "expression" of those mental states. *Wise Choices* aims to systematically develop and defend this basic kind of expressivist view in metanormative inquiry, improving on the earlier versions of it (defended by the likes of Ayer, Stevenson, etc.). Along the way, Gibbard draws insightfully on evolutionary theory and game theory, and develops views in philosophy of language, philosophy of mind, and many other areas. The papers in section II engage with some of the themes from Gibbard's work that are a key part of his work from *Wise Choices*. Simon Blackburn considers the ways in which he and Gibbard, while sharing a broadly expressivistic framework, depart on their understanding of which attitudes (e.g., desires or the state of "norm

acceptance") and which notions (e.g., "warrant") the expressivist needs to treat as basic. Another central component of *Wise Choices* is Gibbard's theory of moral judgment and the role of moral emotions in human life. Stephen Darwall and Howard Nye are sympathetic to Gibbard's basic approach to the relationship between judging that an act is morally wrong and judging that blame for acting is warranted. However, both suggest that components of the view should be revised, in order to explain a wider range of moral judgments than Gibbard's previous view covered (e.g., judgments involving different moral concepts than the ones Gibbard focused on).

Gibbard's 2003 book *Thinking How to Live* raises and refines a host of issues in metaethics that continue to shape much of the discussion in that area today. Indeed, its influence expands well beyond metaethics (or metanormative inquiry, more broadly construed). This is in part because of how it (as well as *Wise Choices*) raises questions in the philosophy of language throughout. For example, in *Thinking How to Live*, Gibbard makes heavy use of "minimalist" accounts of truth and facthood, develops an expressivist semantics for normative terms, and integrates these linguistic tools with a theory of mind which gives central importance to "planning states" in addition to prosaic belief. These themes in the philosophy of language are taken up in section III, where authors take up the importance of minimalism to the expressivist picture (Tristram McPherson), explore the expressivist semantic theory with the tools of formal semantics (Nate Charlow and Alex Silk), raise questions about the relationship between planning states and normative judgment (Seth Yalcin), and extend expressivist ideas beyond normative language (Caleb Perl).

While Gibbard's work in metaethics (and metanormative inquiry, more broadly construed) has focused on developing his own expressivistic program, he indirectly raises difficult questions for those pursuing other metanormative research programs, especially those inclined to more metaphysically robust forms of realism. The final two papers in section IV, by Peter Railton and Billy Dunaway, take up the "quasi-realism" from *Thinking How to Live*. The expressivist does not start with an inventory of normative properties to explain normative thought and talk. However, as Gibbard points out in chapter 9 of the book, he agrees with much of what typical "realists" will say about the existence of normative properties, and about the objectivity of morality. This puts pressure on realists to say what more traditional, nonexpressivist forms of "realism" gain over Gibbardian expressivism. This is a project that many continue to undertake.

Just as beliefs can conflict with each other by being inconsistent, Gibbard's notion of planning states allows for an analogous relation by virtue of being capable of *disagreement*. Disagreement in plan is central to Gibbard's development of a logic for normative judgments, and allows him to recover much of what traditional realists say about normativity. To take one example: Gibbard's expressivism avoids the (what many take to be undesirable) consequences of subjectivism because, on Gibbard's view, normative judgments by people with different normative sensibilities express plans that disagree with each other. The paper by Mark Schroeder in section IV takes up the explanation for the absence of convergence in

plans (or the persistence of disagreement), and Lauren Olin explores the work the resource of disagreement can do in other areas of thought and talk (in particular, thought and talk about humor).

Gibbard makes a number of additional interesting claims in *Wise Choices* and *Thinking How to Live*, and each deserves its own paper. For example, more comprehensive versions of the previous sections would include commentary on Gibbard's discussion of the Frege-Geach problem, in addition to work on supervenience, truth, and the Open Question Argument from G. E. Moore.

Gibbard's most recent book, *Meaning and Normativity*, applies his expressivist theory to the claim that "meaning is normative". Although Gibbardian expressivism has found applications to many areas outside the subject matter of metaethics, the concerns of *Meaning and Normativity* are especially interesting because expressivism is a theory of meaning itself, as it purports to explain what normative terms mean. The immensely rich discussion of the book, then, occurs within a theory that holds that expressivism explains what normative claims mean. If meaning is normative, expressivism is itself a normative claim. Gibbard's book, which aims to unpack the slogan "meaning is normative", introduces and explores the complexities that arise for expressivists who accept this idea. This inspires a new literature at the intersection of the philosophy of language and metaethics. The papers in section V by Paul Boghossian, Paul Horwich, and Jamie Dreier take up Gibbard's rich discussion of these issues.

Gibbard's PhD dissertation at Harvard (which stemmed from an undergraduate paper he wrote at Swarthmore) concerned consequentialism in ethics. He continued to write about broadly consequentialist themes throughout his career. His first published article is "Rule Utilitarianism: Merely an Illusory Alternative?", which appeared in 1965. In 2008 he published his U. C. Berkeley Tanner Lectures on Human Values as *Reconciling Our Aims: In Search of Bases for Ethics*, which involves a defense of a form of consequentialism. The papers by David Braddon-Mitchell and Connie Rosati in section VI take up these themes by exploring the tenability of act utilitarianism, consequentialism in general, and the methodology that underlies discussion in normative ethics.

As our brief discussion above underscores, the six sections in this volume cover a wide range of topics. And so, we hope, the papers in this volume will give the reader a good sense of the range of contributions Gibbard has made over the years. But we have left a good amount out as well. For example, many philosophy graduate students today who take a seminar in metaphysics will still read about Gibbard's views about the statute and the clay in "Contingent Identity", which Gibbard published in 1975. Or, to take another example, in social choice theory, a theorem bears his name as the Gibbard-Satterwaithe Theorem, which Gibbard published in "Manipulation of Voting Schemes: A General Result" in 1973.[1] Gibbard continued to regularly publish on social choice and economic modeling throughout his career. This volume

1 The Satterwaithe paper, which appears in 1975, contains the same result.

would, ideally, give the reader a chance to see how Gibbard's contributions in these areas (and others) have unfolded in recent philosophy. Each of these contributions is notable in its own right, and would merit discussion in a commemorative volume. Instead, we point the reader to Gibbard's discussion of these papers in his reply to authors at the end of this volume. Their absence from the main volume will have to be another form of praise of Gibbard's impressive philosophical career, as he simply produced too much to be covered in one volume.

As we hope our brief discussion here underscores, Gibbard's work involves a combination of two intellectual traits that are often not combined in such a fruitful way. On the one hand, much of Gibbard's work is systematic and aimed at the development of an overall "expressivist" view. On the other hand, his work is wide-ranging, covering a wide array of different subareas of philosophy, as well as ongoing engagement with work from the natural and social sciences. This has led to an impressive body of work that we (both as coeditors of this volume and as former PhD students of Gibbard's) have gained a lot from carefully engaging with over the years. We hope that this volume will help others engage with Gibbard's work in similarly productive ways in the years to come.

I

NORMS IN DECISION AND BELIEF

1

DECISION DYNAMICS AND RATIONAL CHOICE

William L. Harper

I would like to begin with a story that testifies to Allan Gibbard's philosophical acumen, and the benefit afforded by the opportunity to interact and discuss philosophy with him. When I was visiting at Pittsburgh in 1974–1975, Rich Thomason had me over for brunch where he introduced me to Gibbard. I remember that he very enthusiastically said that Gibbard was someone I would really benefit from interacting with. I had actually met Gibbard when he gave a talk on conditionals and probability at Rochester when I was a graduate student; but, Pittsburgh was the first place we had an opportunity to interact at length. Within a few weeks, it was clear that Thomason was right. Gibbard soon became one of my best friends. We spent a lot of time together. Our many discussions, often at dinner with John Haugeland, made it clear that Gibbard's talent for philosophy, including his grasp of technical issues, was extraordinary. I am very grateful for having had the opportunity to benefit from interacting and working with him.

I also want to include a story giving some background about the coming to be of our well-known joint paper. At a conference, on foundations and applications of decision theory, in spring of 1975, Gibbard was among several who were staying at the home of my wife Susan and me. One evening I had been at a talk by Dick Jeffrey, where I had brought up David Lewis's triviality result for conditioning on conditional probability conditionals and Stalnaker's response.[1] Anatol Rapaport said he would revise the conference schedule to have a session on this. When I got back home, I told Gibbard about it. We began a long discussion that ended with a draft for a joint presentation. Gibbard and I worked it up into the Gibbard and Harper paper for inclusion in the conference proceedings. I vividly remember

1 See Lewis (1976) and Stalnaker (1980a). Lewis first presented his trivialization result and Stalnaker responded in a session at the 1972 meeting of the Canadian Philosophical Association in Montreal.

how impressed I was when Gibbard brought the draft he had completed to a meeting with me. He deserves far more of the credit for that paper than I do.

Finally, before getting to my contribution I want to mention that Gibbard's 1980 paper, "Two Recent Theories of Conditionals," and Stalnaker's 1980b paper, "Indicative Conditionals," are two of the most seminal treatments of indicative conditionals that have ever been produced. Both of them were delivered at a 1978 workshop on pragmatics and conditionals.[2]

My contribution to this volume is a review of arguments and issues raised by examples where causal decision theory results in decision instability. In addition, this paper includes an appendix by Sona Ghosh with some results from applications of computer resources to discrete decision dynamics for such problems. It also has an appendix (2a) by Brian Skyrms giving a general convergence theorem for applications of continuous decision dynamics to such problems and a second appendix (2b) by Brian Skyrms giving a convergence result for tempered discrete time dynamics for such problems.

1. Death in Damascus and Ahmed's Challenge

Gibbard's *Death in Damascus*:

> Consider the story of the man who met death in Damascus. Death looked surprised, but then recovered his ghastly composure and said, "I'm coming for you tomorrow." The terrified man that night bought a camel and rode to Aleppo. The next day death knocked on the door of the room where he was hiding and said "I have come for you."
>
> "But I thought you would be looking for me in Damascus" said the man.
>
> "Not at all" said death "that is why I was surprised to see you yesterday. I knew that today I was to find you in Aleppo."
>
> Now suppose the man knows the following. Death works by an appointment book which states time and place; a person dies if and only if the book correctly states in what city he will be at the stated time. The book is made up weeks in advance on the basis of highly reliable predictions. An appointment for the next day has been inscribed for him. Suppose, on this basis, the man would take his being in Damascus the next day as strong evidence that his appointment with death is in Damascus, and would take his being in Aleppo the next day as strong evidence that his appointment is in Aleppo. (Gibbard and Harper 1978, 157–58)

If we set utilities and epistemic probabilities appropriately we have,

$$U(D) > U(A) \text{ whenever } P(A) > \tfrac{1}{2}, \text{ and } U(A) > U(D) \text{ whenever } P(D) > \tfrac{1}{2},$$

2 These papers, the other papers at that conference, together with a number of significant earlier papers, including Lewis's trivialization paper and Stalnaker's letter responding to it, are available in Harper et al. (1980).

where U is the causal utility of the acts of staying in Damascus, D, and going to Aleppo, A.

The mixed strategy (½ A, ½ D), which assigns equal chances of ½ to each of these alternative acts, is the only strategy that is *self-ratifying*. Learning that one was committed to choosing it would not give one grounds for choosing an alternative instead.

Suppose I am faced with this choice and I have a coin in my pocket, so I can implement this mixed strategy M = (½ A, ½ D). Choosing M does seem as good as one can do in this terrible situation. The Death in Damascus example, also, has the virtue that the probabilities recommended in M are relatively easy to implement.[3]

Suppose we consider the classic, win-lose, zero-sum game that would represent the man as a player against death as his opponent. By allowing mixed strategies, every two-person zero-sum game has at least one equilibrium pair, and when there are several they all have the same payoffs and the equilibrium strategies in them are interchangeable.[4] Having each player commit to their part of one of these equilibrium pairs is counted as the solution game theory recommends for a pair of rational players in such a game. In this game, there is only one equilibrium. This recommended solution would be the pairing of the two corresponding (½ A, ½ D) mixed strategies against one another. Having either player's choice known ahead of time by the other player would not give that other player grounds for altering their choice.

In his "Weakly self-ratifying strategies: Comments on McClennen," Gibbard argues that for a zero-sum game with a mixed strategy solution for Dick and Jane,

> What matters for this equilibrium is that Dick have the right subjective probabilities for what Jane will do, not that these be the actual probabilities with which she will do these things.
>
> . . . I don't know how to work out a theory of this kind so as to be fully satisfactory. But it does have a virtue: It does not issue in the strange requirement that in a situation in which Jane finds acts A and B to offer equal prospects, she must, to be rational, choose some particular probability mixture of the two. It allows that any probability mixture would be rational—even A for sure and B for sure.
>
> (1992, 224)

According to the classic game theoretic solution for this game, you should accept that Death is implementing his mixed strategy (½ A, ½ D). This makes you sure that any probability

---

3 Implementing chances for mixed strategies generally is more difficult. Gibbard points out that implementing mixed strategy probabilities with pseudo-randomizers leads to the problem that they can perform determinate calculations that pretty well mimic randomization. But in one way they do not mimic true randomization: a sufficiently powerful observer could predict the outcome of the process. (1992, 223)

One significant new development is that there are now available quantum computing programs that provide genuine randomizers (see http://qrng.anu.edu.au/index.php for the ANU quantum random number generator website).

Having access to chance devices affords many advantages to Bayesian decision theory (Harper et al. 2012).

4 See for example Luce and Raiffa (1957, 86)

mixture between *A* and *D*—even *A* for sure and *D* for sure—would give you the same expected utility. Such certainty about Death's choice makes the requirement that you choose your part of the mixed strategy solution one that endorses implementing this particular strategy while you also judge that all your alternatives to this choice have equal expected utility. There certainly is a tension between individual rational choice theory and what Gibbard has identified as this "strange" requirement imposed by game theory.

Suppose you are in this Death in Damascus situation without a coin and are offered the opportunity to use a fair coin for a price. In a recent paper, "Dicing with Death," Arif Ahmed (2014) has challenged causal decision theory by showing that it would recommend rejecting the offer, however small the price. He calls this "absurd advice" (Ahmed 2014, 590–91).

2. Introducing Causal and Evidential Decision Theory Using Examples from Egan and Gibbard

In his 2007 "Some Counterexamples to Causal Decision Theory," Andy Egan began with a version of the, widely discussed, medical Newcomb problem as a counterexample to evidential decision theory.

> *The Smoking Lesion*
> Susan is debating whether or not to smoke. She believes that smoking is strongly correlated with lung cancer, but only because there is a common cause—a condition that tends to cause both smoking and cancer. Once we fix the presence or absence of this condition, there is no additional correlation between smoking and cancer. Susan prefers smoking to not smoking without cancer, and she prefers smoking with cancer to not smoking with cancer. Should Susan smoke? It seems clear that she should. (Set aside your theoretical commitments and put yourself in Susan's situation. Would you smoke? Would you take yourself to be irrational for doing so?)
> (2007, 94)

For evidential decision theory, having Susan's belief that whether or not she has cancer is fixed by whether or not she has the common-cause-condition should be irrelevant to your choice. Where *C* is your having cancer and *S* is your smoking, evidential decision theory evaluates your options by

$$Eu\,(S) = Cr\,(C \mid S)\,\mathrm{u}(S\&C) + Cr\,(\text{-}C \mid S)\,\mathrm{u}(S\&\text{-}C)$$

$$Eu\,(\text{-}S) = Cr\,(C \mid \text{-}S)\,\mathrm{u}(\text{-}S\&C) + Cr\,(\text{-C} \mid —S)\,\mathrm{u}(\text{-}S\&\text{-}C),$$

where $Cr\,(C \mid S)$ is your epistemic conditional credence in having cancer given that you smoke and $Cr\,(C \mid \text{-}S)$ is your epistemic conditional credence in having cancer given that you don't smoke, u (*S*&C) represents your utility for smoking with cancer and u (*-S* & *C*)

the utility representing your preference for not smoking with cancer, and so forth. Given that your epistemic conditional credence $Cr\,(C \mid S)$ is sufficiently higher than $Cr\,(C \mid \text{-}S)$, evidential decision theory will recommend not smoking, even though your utility for smoking with cancer is significantly higher than your utility for not smoking with cancer and your utility for smoking without cancer is, also, significantly higher than your utility for not smoking without cancer.

Take yourself as, like Susan, accepting that the correlation between smoking and cancer is only because there is a common condition *K* that tends to cause both smoking and cancer. Causal decision theory would evaluate your choosing to smoke by

$$Cu\,(S) = Cr\,(S \mathbin{\square\!\!\rightarrow} C)\ \mathrm{u}\ (S\&C) + Cr\,(S \mathbin{\square\!\!\rightarrow} \text{-}C)\ \mathrm{u}\ (S\&\text{-}C),$$

where $(S \mathbin{\square\!\!\rightarrow} C)$ and $(S \mathbin{\square\!\!\rightarrow} \text{-}C)$ are, respectively, the subjunctive conditionals.

If I were to smoke I would have cancer

and

If I were to smoke I would not have cancer.

Similarly, causal utility theory would evaluate your choosing the alternative of not smoking by

$$Cu\,(\text{-}S) = Cr\,(\text{-}S \mathbin{\square\!\!\rightarrow} C)\ u(\text{-}S\&C) + Cr(\text{-}S \mathbin{\square\!\!\rightarrow} \text{-}C)\ \mathrm{u}(\text{-}S\ \&\ \text{-}C).$$

If the condition *K* does obtain, then $(S \mathbin{\square\!\!\rightarrow} C)$ and $(\text{-}S \mathbin{\square\!\!\rightarrow} C)$ are both true and $(S \mathbin{\square\!\!\rightarrow} \text{-}C)$ and $(\text{-}S \mathbin{\square\!\!\rightarrow} \text{-}C)$ are both false. You have cancer whether or not you decide to smoke or to not smoke. Similarly, if condition *K* does not obtain, then $(S \mathbin{\square\!\!\rightarrow} C)$ and $(\text{-}S \mathbin{\square\!\!\rightarrow} C)$ are both false and $(S \mathbin{\square\!\!\rightarrow} \text{-}C)$ and $(\text{-}S \mathbin{\square\!\!\rightarrow} \text{-}C)$ are both true. In either case, whether or not you have cancer is not influenced by whether or not you smoke. In this example, smoking does not increase your chance of having lung cancer and not smoking doesn't make you any less likely to have lung cancer.

That subjunctive conditionals differ significantly from indicative conditionals is clear from examples like Adams's (1970) comparing of

If Oswald didn't shoot Kennedy, then someone else did

and

If Oswald hadn't shot Kennedy, then someone else would have.

Very likely, you clearly accept the indicative conditional. It is evaluated by Ramsey's test. You hypothetically add the antecedent to what you accept and then evaluate the consequent. Your knowledge that Kennedy was shot supports the resulting conclusion. Even more likely, you are not that confident in the corresponding subjunctive conditional. It is not acceptable without assuming significant details about the relevant causal circumstances.

In our example decision, the appropriateness of the Ramsey test evaluation of the indicative conditional would make your evaluation of

If I smoke I have cancer

equivalent to

Cr (I have cancer | I smoke),

your epistemic conditional credence $Cr\,(C \mid S)$.

As we have noted, this evaluation makes your knowledge of Susan's beliefs about the causal circumstances irrelevant to what evidential decision theory recommends that she chooses.[5]

Causal decision theory makes Susan's beliefs about the relevance of the presence or absence of the common cause *K* support the rationality of her choosing to smoke. Consider the following decision matrix.

	K	**-K**
S	*u*(S&C)	*u*(S&-C)
-S	*u*(-S&C)	*u*(-S&-C)

Given that $u(S\&C) > u(\text{-}S\&C)$ and that $u(S\&\text{-}C) > u(\text{-}S\&\text{-}C)$, your choice to smoke dominates your choice to not smoke. Causal decision theory endorses dominance when causal independence obtains, because $Cr\,(S \mathbin{\square\!\!\rightarrow} C) = \text{Cr}\,(S \mathbin{\square\!\!\rightarrow} \text{-}C)$ and $Cr(\text{-}S \mathbin{\square\!\!\rightarrow} C) = \text{Cr}\,(\text{-}S \mathbin{\square\!\!\rightarrow} \text{-}C)$. Using these subjunctive conditionals lets causal decision theory recover your rationally endorsing choosing smoking over not smoking, when you take yourself as sharing Susan's preferences and her beliefs about the relevant causal structure.

5 John Cantwell (2010) has argued that future tensed indicative conditionals with acts under consideration for choice as antecedents and outcomes of choosing as consequents are appropriately evaluated as subjunctives. This may appear to support Stalnaker's treatment of indicative conditionals.

It would be very interesting to see what Gibbard has to say about this and about Cantwell's application of it to counter Egan's proposed counterexample to causal decision theory.

The prisoner's dilemma is a classic game-theoretic example, in which the causal independence of your choice and that of your partner supports the rationality of the dominant choice to confess; provided that, all you care about is how long you stay in jail. Evidential decision theory would endorse cooperation if you took your decision to cooperate as strong enough evidence that your partner will cooperate. This is not enough to rationally support cooperative outcomes in a one-shot prisoners' dilemma situation.[6]

Where the evidential decision theorist relies on well-understood epistemic conditional credences to represent the resources needed to explicate rational choice, causal decision theorists argue that more is required to adequately represent an agent's relevant beliefs about the causal circumstances. The subjunctive conditionals used in Gibbard and Harper bring in what might be regarded as metaphysical baggage that the evidential decision theorists seek to avoid. The truth semantics for such conditionals is a difficult subject that has not yet resulted in widespread agreement among philosophers and linguists.

The *K* partition for Susan yields

$$Cr\,(S \mathbin{\square\!\!\rightarrow} C) = Cr\,(K)\;Cr\,(C \mid S \,\&\, K) + Cr\,(\text{-}K)\;Cr\,(C \mid S \,\&\, \text{-}K)$$

One feature of the epistemic evaluation of such conditionals, which is widely agreed upon, is that a rational agent's credence in such a conditional ought to agree with her estimate of the conditional chance of the consequent on the antecedent. In Susan's case, we have the *K*'s counting as the relevant alternative chance setups so that $Cr\,(C \mid S \,\&\, K)$ counts as her estimate of the conditional chance $Ch\,(C \mid S)$ given *K* and $Cr\,(C \mid S \,\&\, \text{-}K)$ counts as her estimate of the conditional chance $Ch\,(C \mid S)$ given *-K*. As Brian Skyrms (2013) has argued, for core cases, the evaluations of the conditional chances corresponding to an appropriate *K* partition of the relevant alternative chance setups yield evaluations that agree with evaluations of the subjunctive conditionals. This allows conditional chances to do the work needed for causal decision theory without requiring any more detailed assumptions about the semantics and metaphysical commitments of subjunctive conditionals.

---

6 I remember a conversation I had with Gibbard in which I mentioned that American Korean war prisoners were much more likely to choose the dominant noncooperative act when put into prisoners' dilemma situations by their North Korean captors than were Turkish prisoners. I suggested that this might be explained by the sort of very strict discipline I had observed on a Turkish ship when I was a U.S. naval officer. Alan responded by pointing out that U.S. Marines were also resistant to the prisoners' dilemma temptations offered by their North Korean captors. These examples strongly suggest that the utilities of these agents are not limited to what would correspond to their times in jail. For such agents, the noncooperative act would not dominate.

See Hofstadter (1983) and discussion in Harper (1993) for evidence that just expecting the others to do what you do is not enough to rationally support cooperation in a one-shot prisoners' dilemma.

3. Decision Instability

Egan also claims:

> It's easy to modify *The Smoking Lesion* in order to make it a counterexample to CDT rather than EDT. We just have to change the case in the following way: Rather than letting Susan believe that the lesion (a) causes one to smoke, and (b) causes one to get cancer, let her believe that the lesion (a) causes one to smoke, and (b) causes one's lungs to be vulnerable to cigarette smoke, such that smoking causes cancer in those with the lesion, but not in those without.
>
> (2007, 103)

Such an example had already been introduced by Gibbard, who, in the 1992 paper we have been discussing, counts it as one that exhibits problems for rationally choosing to adopt a mixed strategy.

Gibbard's smoking gene example.

> Imagine this: Smoking it turns out, does cause cancer, but only in people with a certain gene. This gene, indeed, has two effects: First, as I said, it makes one susceptible to smoking's causing cancer. Second, though, it makes one reckless in situations precisely like this, so that in this situation, if one enjoys smoking, one will smoke despite the danger. Knowing all this, the question is, what is it rational to do if you somewhat enjoy smoking, and very much don't want to get cancer? If you recklessly smoke, that indicates you have the gene, so that with you, smoking causes cancer. If you cautiously refrain from smoking, that indicates you don't have the gene, so that you could enjoy smoking and still not get cancer.
>
> You might adopt a mixed strategy, and give yourself some intermediate chance of smoking. For this case, we have to elaborate the story. Imagine in addition, that the more reckless you are—the higher a chance you give yourself of smoking—the more likely you are to have the gene. Suppose the probability of having the gene is a continuous function of how recklessly you gamble. Then there will be a probability of smoking such that, if you know that it is the probability you are adopting, the prospects you rationally envisage are equal whether you smoke or not. This strategy is self-ratifying, but only weakly. It recommends itself, but equally recommends any other probability mixture of smoking or not smoking.
>
> (1992, 218)

Gibbard introduces this individual-decision-theoretic example to exhibit the basic problem he finds with choosing the mixed strategy, recommended as your part of the solution for a zero-sum game, as the unique option for you to adopt. The recommended mixed strategy is self-ratifying:

> Your choosing it would not give you evidence that you would be better off choosing one of your other options instead.

The problem is that it is only weakly self-ratifying. Your choosing it gives you evidence that your utility would be the same whether you committed to it, or either of the pure strategies in it, or any other probability mixture of them.

> If you know that it is the strategy you are adopting, then you will think that it is the strategy that holds out best prospects. But, it is only weakly self-ratifying: If you adopt it, you won't think it is the unique strategy that holds out best prospects. It recommends itself, but it recommends alternative strategies too. Adopting your equilibrium strategy, you will think that these alternative strategies hold out equally good prospects. But these alternative strategies are not self-ratifying: for any of them, if you had known that it was the strategy you were adopting, then you would not have thought it held out best prospects. Only one strategy recommends itself, but it recommends other strategies equally. The strategy is uniquely but weakly self-ratifying.
>
> (Gibbard 1992, 224)

Gibbard argues that committing to such a mixed strategy would raise a conflict with expected utility as your rational decision-making guide.

4. Ratifiability

Where Gibbard focused on problems for choosing mixed strategies, as the classic solution for decision instability in game theory, Egan put forward *The Murder Lesion* and *The Psychopath Button* as counterexamples to causal decision theory.

> *The Murder Lesion* (Egan)
> Mary is debating whether to shoot her rival, Alfred. If she shoots and hits, things will be very good for her. If she shoots and misses, things will be very bad. (Alfred always finds out about unsuccessful assassination attempts and he is sensitive about such things) If she doesn't shoot things will go on in the usual, okay-but-not-great kind of way. Though Mary is fairly confident that she will not actually shoot, she has, just to keep her options open, been preparing for this moment by honing her skills at the shooting range. Her rifle is accurate and well maintained. In view of this, she thinks that if she were to shoot, then she would hit. So far so good. But Mary also knows there is a certain sort of brain lesion that tends to cause both murder attempts and bad aim at the critical moment. If she has this lesion, all of her training will do her no good—her hand is almost certain to shake as she squeezes the trigger. Happily for most of us, but not so happily for Mary, most shooters have this lesion, and so most shooters

> miss. Should Mary shoot? (Set aside your theoretical commitments and put yourself in Mary's situation. Would *you* shoot? Would you take yourself to be irrational for not doing so?)
>
> *The Psychopath Button*
> Paul is debating whether to press the "kill all psychopaths" button. It would, he thinks, be much better to live in a world with no psychopaths. Unfortunately, Paul is quite confident that only a psychopath would press such a button. Paul very strongly prefers living in a world *with* psychopaths to dying. Should Paul press the button? (Set aside your theoretical commitments and put yourself in Paul's situation. Would *you* press the button? Would you take yourself to be irrational for not doing so?)
>
> (Egan 2007, 97)

He admits that some people lack the clear intuition of the irrationality of shooting for Mary; but, he claims that pretty much everyone seems to have the requisite intuition in the case of Paul pressing the button (2007, 97, fn3)

Egan also discusses a three-option smoking lesion example due to Anil Gupta, which convinced him to give up his initial tentative acceptance of ratifiability as a constraint that would yield the correct decisions in his proposed counterexamples to causal decision theory.

> *The Three-Option Smoking Lesion*
> Samantha is deciding whether to smoke. But her situation is slightly more complicated than Susan's. Samantha has three options: Smoke cigars, smoke cigarettes, or refrain from smoking altogether. Call these options **CIGAR, CIGARETTE**, and **NO SMOKE**. Due to the ways that various lesions tend to be distributed, it turns out that cigar smokers tend to be worse off than they would be if they were smoking cigarettes, but better off than they would be if they refrained from smoking altogether. Similarly, cigarette smokers tend to be worse off than they would be smoking cigars, but better off than they would be refraining from smoking altogether. Finally, nonsmokers tend to be best off refraining from smoking.
>
> (2007, 112)

CIGAR is unratifiable because choosing to smoke cigars is good evidence that you would be better off smoking cigarettes and **CIGARETTE** is unratifiable because smoking cigarettes is good evidence that you would be better off smoking cigars. **NO SMOKE** is the only ratifiable option. Egan argues:

> In this sort of case, we can understand someone who finds herself deciding on option 1 or 2 rethinking and doing some vacillating *between options 1 and 2*. What seems clearly *irrational* is for the person who finds herself deciding on either 1 or 2 to perform 3 on grounds of its

> ratifiability. If she finds herself deciding on 1 or 2, she has excellent reason to think that 3 would be the *worst* thing to choose.
>
> (2007, 112)

Suppose that her vacillation between 1 and 2 reached a probability mixture between 1 and 2 on which she would regard their expected values as equal. The corresponding mixed strategy between 1 and 2 would be ratifiable and would, clearly, be preferable to the only ratifiable pure option of not smoking.

In his 2008, "No Regrets, or: Edith Piaf Revamps Decision Theory," Frank Arntzenius applies ratifiability to a version of Egan's *Psychopath Button* example. He takes the *K* partition for his agent Johnny to be given by 'Johnny is a psycho' and 'Johnny is not a psycho' and puts in utilities and probabilities that will lead causal decision theory to recommend that he push. He then points out that, given that Johnny's conditional credence $Cr(\text{I am a psycho} \mid \text{I push}) = 0.9$, his updated credences on the assumption that he pushes will make the causal utility of not pushing significantly higher than that of pushing. So pushing would violate Piaf's maxim that a rational person should not be able to foresee that she will regret her decisions. He suggests applying Brian Skyrms' dynamics of rational deliberation (Skyrms 1990).

> The basic idea is very simple. Johnny starts his deliberation with certain credences that he is a psycho and certain credences that he will push. He then does a causal utility calculation. He finds the causal utility of pushing is higher than that of not pushing. But rather than that he immediately becomes certain that he will push, he merely increases his credence that he will push. In light of these new credences he recalculates causal utilities. Armed with the result he resets his credences in pushing the button. And so on, until he reaches credences which are at equilibrium.
>
> (2008, 293)

Johnny's credence about his choice in such an equilibrium would not lead to violations of Piaf's maxim.

Arntzenius does not, however, argue that Johnny should commit to the corresponding ratifiable mixed strategy, in which you turn over your choice to the outcome of chance device with the equilibrium probabilities for your acts. Instead he, tentatively, offers an alternative view of mixed decisions.

> My tentative view is that to make a certain mixed decision is just to have certain credences in one's acts at the end of a rational deliberation. On this view, mixed decisions are not decisions to *perform* certain acts with certain probabilities. Now, my tentative view is admittedly an odd view. For on this view rationality constrains what *credences* one should have at the end of

> deliberation. Decision theory, on this view, does not evaluate the rationality of actions. Rather it evaluates the rationality of credences in actions.
>
> (2008, 292)

Arntzenius avoids commitment to ratifiable mixed strategies by limiting his appeal to ratifiability to evaluations of credences in actions, rather than evaluations of decisions of which act to choose.

5. Response by Joyce to Egan

In his 2012 "Regret and Instability in Causal Decision theory," Joyce offers the following case, as a version of Egan's *Murder Lesion*.

> *Murder Lesion* (Joyce)
> Life in your country would be better if you killed the despot Alfred. You have a gun aimed at his head and are deciding whether to shoot. You have no moral qualms about killing; your sole concern is whether shooting Alfred would leave your fellow citizens better off. Of course, not everyone has the nerve to pull the trigger, and even those who do sometimes miss. By shooting and missing you would anger Alfred and cause him to make life in your country much worse. But, if you shoot and aim true the Crown Prince will ascend to the throne and life in your country will improve. Your situation is complicated by the fact that you are a random member of a population in which 20% of people have a brain lesion that both fortifies their nerve and causes their hands to tremble when they shoot. Eight in ten people who have the lesion can bring themselves to shoot but they invariably miss. Those who lack the lesion shoot only one time in ten, but always hit their targets.
>
> (2012, 124)

Joyce gives utility numbers that correspond to the following decision matrix.[7]

	L	**-L**
S	-30	10
-S	0	0

His initial epistemic probabilities are:[8]

7 Joyce (2012, 124).
8 Ibid.

$$P_0\,(L) = .2,\ P_0\,(S\&L) = .16,\ P_0\,(\text{-}S\ \&\ L) = .04,\ P_0\,(S\ \&\ \text{-}L) = .08,\ P_0\,(\text{-}S\ \&\ \text{-}L) = .72$$

These give the following initial causal utilities:

$$U\,(S) = .2\,(\text{-}30) + .8\,(10) = 2, \quad U(\text{-}S) = 0.$$

These basic utilities and initial probabilities give shooting a higher causal utility than not shooting. But, they also make

$$P_0\,(L \mid U(S) = .2) > P_0\,(L)$$

so you are now more sure that you have the lesion. This affords new information about the causal circumstances relevant to your choice. As Joyce informatively points out, to act on your initial utility calculation, without taking such potentially relevant freely available information into account, would be to act like a blackjack player who takes a hit without bothering to look at her hole card. If you are rational, you should take this information into account by using your new credence for L,

$$P_1(L) = P_0\,(L \mid U(S) = .2)$$

to calculate your new utilities $U_1\,(S)$ and $U_1\,(\text{-}S)$ before committing to choose either option.

Joyce argues that, as the blackjack example illustrates, the right constraint on commitment to act is:

> *Full Information*. You should act on your time-t utility assessments only if those assessments are based on beliefs that incorporate all the evidence that is both freely available to you at t and relevant to what your acts are likely to cause.
>
> (2012, 127)

This requires that your sequence of recalculations should not be terminated by your committing to a choice until you reach a stage t such that conditioning on the outcome of your utility calculation does not differ from your credence in L.

Joyce imposes a condition to reflect the feature of *Murder Lesion* examples that learning that you are definitely inclined toward or against shooting provides you with evidence about the presence of the lesion. He points out that on any of the wide range of update rules that can have this feature it is only possible to obtain an equilibrium in his example when $U_t\,(S) = 0$ and $P_t\,(L) = 0.25$.

Like Arntzenius, Joyce does not endorse adopting the corresponding mixed strategy. He does, however, go beyond Arntzenius by using the information gained to guide your choice.

> Given full information, this is *the* state that you should use when assessing the causal expected utilities for purposes of action. So, you must base your choice in Murder Lesion on the assessments of causal expected utility characterized by $U(S) = U(\text{-S})$. It follows that CDT, rather than recommending either action alone, tells you to be entirely indifferent between shooting and not shooting. Once you have processed all the available information about what your acts might cause, you can rationally choose to shoot, to refrain from shooting, or to perform any "mixed act" that leads to shooting with probability p and to refraining with probability $1—p$. All these choices are on a par with respect to their ability to cause desirable results since each maximizes causal expected utility given full information about the causal properties of your acts.
>
> (2012, 134)

Like Gibbard, Joyce takes the equality of the expected utilities to fix the equal rationality of choosing each of all these alternative options.

Joyce defends this conclusion from two objections.

> First, it just might just seem intuitively clear that shooting cannot be permitted. Indeed, Egan notes that many people have this intuition, and he takes this as a reason to regard shooting as irrational. I am less inclined to give such intuitions weight. Many philosophers seem to accept something like this: a strong and commonly held intuition that a given action is rational/irrational is powerful evidence for concluding that the act is rational/irrational, especially when the intuition persists in intelligent people under reflection. I think this is wrong. We have more than fifty years of research by cognitive psychologists showing that people make a wide range of predictable and systematic *errors* when evaluating acts, and that these errors often persist under reflection. Here is a partial list (see Shafir and Tversky (1995) and Gilovich et al. (2002) for more).
>
> (2012, 134)

One thing about Egan's examples is that any moral qualms one has against unjustifiable killing will make it hard to put one's self in Mary's, or in Paul's, situation. Your moral qualms would endorse not-shooting, and not-pushing, independently of any of the issues about decision instability that have been the focus of the debate about whether or not these, now famous, examples pose any difficulties for causal decision theory. Choosing not to shoot, or not to push, because you have such moral qualms would make your choice fail to count as any problem for causal decision theory. Perhaps, the merely implicit influence of such qualms helps bias your intuitions against shooting and pushing the button. By assuming life in your country would be better if you killed the despot, Joyce makes the preferences he asks you to imagine, perhaps, less morally objectionable than those in Egan's version. Nevertheless, your implicit bias against murder may also influence your intuitions in his version, even though you are told: "You have no moral qualms about

killing; your sole concern is whether shooting Alfred will leave your fellow citizens better off." Can these moral qualms be dismissed as irrelevant to the rationality of the choice, the way Shafir and Tversky (1995), Gilovich et. al. (2002), or Tversky and Kahneman (1974) treat heuristics and biases?

The second objection Joyce considers is that neither of the pure strategies is ratifiable.

In a version of Bayesian decision dynamics that we applied to Joyce's *Murder Lesion* (with utilities $U(S\&\text{-}L) = 40$, $U(\text{-}S) = 30$, and $U(S\&L) = 0$), the deliberational fixed point, where $P_f(L) = .25$, was reached with $P_f(S)$ at about .32.[9] The corresponding mixed strategy M would use a random device to generate the choice of S with that probability. Suppose, as seems reasonable, that reaching the commitment to choose S by this method does not change your $P_f(L)$ so that $P_f(L \mid M \text{ (with } S)) = P_f(L \mid M \text{ (with } \text{-}S)) = P_f(L) = .25$. This makes this mixed strategy ratifiable. You would not regard implementing M as giving you evidence that you would be better off implementing an alternative instead. Neither of your pure strategies is ratifiable;[10] but, at the fixed point $U_f(S) = U_f(\text{-}S) = U_f(M)$, as does $U_f(M')$ for any alternative mixed strategy M'.

Joyce argues that, at this deliberation fixed point, you now rightly regard yourself as having appropriately taken into account all the available information about what your acts might cause. Your $P_f(L) = .25$ is the result supported by this.

> Since you know that your decisions cannot influence the presence or absence of the lesion, and since the regrets you will come to have upon choosing S will be based on the belief, caused by the decision, that your chance of having the lesion is $prob(L \mid S) > .25$, it follows that you do not currently fully trust the accuracy of the future beliefs on which your regrets about shooting will be based!
>
> (2012, 140)

Joyce offers an additional argument to back up this claim.

> Once you have achieved the equilibrium in which all available information about the effects of your acts has been taken into account, your unconditional causal expected utilities incorporate all relevant ratifiability considerations!
>
> (2012, 138)

Where we denote the decision do A by δA, he argues:

---

9 See appendix 1.

10 Given that the $P_t(L)$ goes by what is essentially Jeffrey rule updating, where $P(L \mid S)$ and $P(L \mid \text{-}S)$ are fixed at their initial values, we have:

$$U_f(S \mid S) = 8,\ U_f(\text{-}S \mid S) = 30, \text{ and } U_f(\text{-}S \mid \text{-}S) = 30, \text{ while } U_f(S \mid \text{-}S) = 37.896.$$

Specifically, the fact that you know you will regret any act but M is reflected in the unconditional $\boldsymbol{U}$- values of your acts. To see this, suppose for simplicity that S, $\text{-}S$, and M are your only options. We can write their unconditional expected utilities as:

$$\boldsymbol{U}(S) = prob(\delta S)\boldsymbol{U}(S \mid \delta S) + prob(\delta M)\boldsymbol{U}(S \mid \delta M) + prob(\delta\text{-}S)\boldsymbol{U}(S \mid \delta\text{-}S)$$
$$\boldsymbol{U}(M) = prob(\delta S)\boldsymbol{U}(M \mid \delta S) + prob(\delta M)\boldsymbol{U}(M \mid \delta M) + prob(\delta\text{-}S)\boldsymbol{U}(M \mid \delta\text{-}S)$$
$$\boldsymbol{U}(\text{-}S) = prob(\delta S)\boldsymbol{U}(\text{-S} \mid \delta S) + prob(\delta M)\boldsymbol{U}(\text{-}S \mid \delta M) + prob(\delta\text{-}S)\boldsymbol{U}(\text{-}S \mid \delta\text{-}S)$$
$$\ldots$$

The values of $\boldsymbol{U}(S)$, $\boldsymbol{U}(M)$ and $\boldsymbol{U}(\text{-}S)$ thus already reflect not only the bare fact that S is unratifiable, but also the *extent* of its unratifiability (as measured by $\boldsymbol{U}(\text{-}S \mid \delta S)$—$\boldsymbol{U}(S \mid \delta S)$ and $\boldsymbol{U}(M \mid \delta S)$—$\boldsymbol{U}(S \mid \delta S)$). (2012, 139)

The three conditions Joyce imposes mathematically generate game theory's strange constraint on the probabilities for choosing the acts.[11]

6. Responding to Ahmed's Challenge

You are faced with Death in Damascus against a perfect predictor Omega. As Ahmed correctly points out, causal decision theory endorses refusing to pay any price for the opportunity to let the toss of fair coin provide a randomizing alternative to your two predictable basic acts, *Alep* and *Dam*. Joyce defends causal decision theory's assignment of equal utilities to *Dam*, *Alep*, and a free coin option so that $Coin^{-\Delta}$ (paying Δ for the coin option) would be paying to buy what you already have.

In virtue of Omega's reliability you are justifiably certain of this hypothesis:

H If I choose *Alep* or *Dam*, then choosing that act will certainly cause my death.

H seems to entail that you should not choose *Alep* or *Dam* when you have an option, like $Coin^{-\Delta}$, that lowers your mortality probability at a reasonable cost. This is wrong. In *H* the phrase "that act" *rigidly* designates the act you *actually* choose. *H* says nothing about the other act. In fact, choosing the other act, whatever it may be, will certainly cause your survival. This makes the other act a much better choice than $Coin^{-\Delta}$. Unfortunately, you will not know what "that act" and "the other act" refer to until you actually make your choice. Since you (unlike Omega) will not know which act you will pick until you pick it, you lack the crucial piece of evidence needed to determine which of your options will surely cause your death (until it is too late to matter). At the time you pick, you are constrained by the evidence you have then, . . .

11 See section 7 for result and appendix 1 for details.

> So, your credence for "I live" conditional on any of the three acts is one-half. In terms of estimated survival probability, all three offer you the same thing, and paying for *Coin* would be paying to buy what you already have.

At my deliberation equilibrium, I already assign equal epistemic probabilities of ½ to Omega has predicted my picking *Alep* and ½ to Omega has predicted my picking *Dam*. So my information at this point supports assigning to picking each of them a higher expected utility than to paying to be able to pick *Coin*.

This, together with the other arguments discussed, is enough to convince me, a long-time defender of ratifiability as a constraint on rational choice, to consider backing off from that commitment, though I still have reservations about doing so.[12] I do think that Gibbard's worry about the incompatibility between individual rationality and mixed-strategy solutions in zero-sum games motivates issues for game theory as we now understand it.

Before concluding this section, I want to remark on two other contributions in Joyce's forthcoming paper. The first is Joyce's informative defense of causal decision theory's requirement that, at an equilibrium point, your decision will have to be to just *pick* one of the equally desirable alternatives.

> The fact that you are picking means that you are *not* obliged to see your act as optimal conditional on being picked. You must see it as being optimal *in light of the information you have when you pick*.

The term "picking" is a useful way to designate the equilibrium point decision where it is equally rational to choose any of the alternatives. As we have seen, the extra virtue to be conferred on the mixed strategy *M* by its ratifiability would be based upon the additional information you would regard your picking it to provide.[13]

That paper also, very informatively, offers a problem in which paying for the fair coin clearly is the right choice.

> Suppose you are slated to face *DD* with a perfect predictor *later on*. But, you are now offered the opportunity to *avoid* that choice by paying Δ and going to Aleppo/Damascus iff a fair coin lands heads/tails. Omega predicted whether you will take this deal. If he guessed you would

---

12 One is the surprising result mathematically following from the conditions Joyce imposes. See section 7 below. Another is that the assumption of Omega's perfect reliability would still make me unwilling to pick either *Alep* or *Dam*.

13 The case made for accepting "you do not currently fully trust the accuracy of the future beliefs on which your regrets about shooting will be based!" in the shooting example does not support regarding the freely available evidence afforded by considering your conditional credences on *Alep*, on *Dam*, and on $Coin^{-\Delta}$ as irrelevant to what your acts are likely to cause. This makes it hard for me to dismiss the virtue afforded to $Coin^{-\Delta}$ by its ratifiability.

> take it he rolled a fair die and sent assassins to Alleppo/Damascus iff even/odd came up. If not, he executed his standard *DD* protocol.

As Joyce points out,

> Instead of deciding among *Alep*, *Dam* and *Coin*$^{-\Delta}$, you are deciding between F = [face *DD* later, keep Δ] and $-F$ = [avoid *DD* later, pay Δ, have a 50% chance of survival]. From this perspective, facing-DD-and-choosing-*Alep* or facing-DD-and-choosing-*Dam* are not *options*, but acts-*in-prospect* that lie causally downstream from your choice.

Here causal utility will agree with counting the ratifiabiliy of $-F$ as grounds for choosing it over its unratifiable alternative F.

I have long favored the extensive-form tree-structure representation of game theoretic interactions. I expect that investigating game theory issues would be facilitated by more focus on extensive-form reasoning, which includes planning how to make the decisions one reaches as one proceeds down the branches over time.[14]

7. Computations in Deliberation Dynamics[15]

The philosophical discussions about equilibrium endpoint rationality I have been presenting have appealed to acceptance of general results that are claimed to hold for a wide variety of deliberation dynamics. I wanted to see what could be learned from actually carrying out the details of specific applications to these decision problems. I asked Sona Ghosh for help and computer resources that could carry out the details of such applications. The informative results she has obtained are presented in appendix 1.

Bayesian dynamics does not handle negative utilities. So, we had to revise the utilities Joyce used in his example to apply it. We also ran the problem with the decision dynamics Nash used to generate his important results for game theory. This dynamic gave the same initial updates for Joyce's utilities as for our revised utilities to make them nonnegative; but, instead of converging to a single pair of equilibrium values for $P(S)$ and $P(L)$ it began jumping back and forth, over what we expected would be the outcome of dynamic deliberation. This made the Nash dynamic fail to deliver the unique epistemic probability fixed point Joyce argued for. Brian Skyrms suggested we try a version of this dynamic with tempering added to it.[16] We found that this version did converge.

---

14 This does not endorse evidential decision theory for game theoretic reasoning. See, for example, the extensive-form treatment of the prisoner's dilemma in Harper (1993).

15 Harper and Gosh thank Rotman Institute member and Western University applied mathematics professor Robert Corless for generous help with understanding mathematical details.

16 See Skyrms' appendix 2B, p. 34-37, for an account and application of tempering, and p. 31-33 for our application of his suggestion.

We applied the three conditions Joyce offers (in his additional argument for his claim that once you have achieved equilibrium your unconditional causal utilities incorporate all relevant ratifiability considerations) to the equilibrium resulting from applying the Bayes dynamic with our revised nonnegative utilities. The surprising result was that they mathematically entailed picking the mixed strategy over the two alternatives, even though they all have the same expected utility. I believe this result affords grounds for further exploring the issues raised for considering ratifiability as a constraint to add to basic causal decision theory as a guide to rational choice.

I asked Skyrms what he would be able to provide, in addition to his valuable suggestion that we try adding tempering to our application of the Nash deliberation dynamic. He responded by proving the results contained in appendices 2A and 2B. In appendix 2A, he proves the general result that continuous deliberation dynamics will converge to a ratifiable mixed equilibrium in Death in Damascus-type problems. In appendix 2B, he applies tempering to extend this result to a wide range of discrete time dynamics, such as those used in the examples discussed above.

APPENDIX 1

DYNAMICS CALCULATIONS

Sona Ghosh

The calculations described below were all carried out in the open-source mathematical software system SageMath using code written in Python, and often rerun in Maple.

1. Basic Result for Joyce's *Murder Lesion* Problem

We ran the problem, using Joyce's Bayesian decision dynamics for our $P(S)$ update rule.[17] This Bayes update rule is:

$$P(S)_{\text{new}} = P(S)_{\text{Old}} * (U(S)/U(\text{StatusQuo})),$$

where $U(\text{StatusQuo})$ is given by

$$U(SQ) = P(S)_{\text{old}} * U(S) + P(\text{-}S)_{\text{old}} * U(\text{-}S)$$

17 This decision dynamic rule is equivalent to the update rule for discrete replicator dynamics in evolutionary dynamics. See, for example, Maynard Smith (1982, 183).

This rule requires nonnegative utilities. So we ran it with utilities that were all 30 units higher than those stated above to avoid negative utilities (i.e., for initial input values $U(S,L) = 0$, $U(S,\text{-}L) = 40$, $U(\text{-}S,L) = 30$, $U(\text{-}S,\text{-}L) = 30$, and $P(L)_{\text{initial}} = 0.2$). When we did this, $P(S) = 0.321428571428571$ and $P(L) = 0.25$, *at* the fixed point, which was obtained at iteration 164.

We reran the dynamics over alternative initial $P(L)$'s: For initial $P(L) = 0.25$, $P(S) = 0.275$ at the fixed point, obtained immediately. For initial $P(L) = 0.4$, $P(S) = 0.169642857142857$ at fixed point was obtained at iteration 232. For initial $P(L) = 0.6$, there was extremely slow convergence toward $P(L) = 0.25$ and $P(S) = 0$. For initial $P(L) = 0.7$, results appeared to converge quickly again, to $P(S) = 0$, $P(L) = 0.341463414634148$ starting at iteration 239; for initial $P(L) = 0.8$, even more quickly, to $P(S) = 0$, $P(L) = 0.470588235294118$ starting at iteration 93; and for initial $P(L) = 0.9$ to $P(S) = 0$, $P(L) = 0.666666666666667$ more quickly still at iteration 40.

It seems that for varying initial values of $P(L)$, the fixed-point value obtained for $P(L)$ may consistently be 0.25, approaching convergence evermore slowly, as initial $P(L)$ is increased, until some point between 0.6 and 0.7, after which the fixed-point value for $P(L)$ slowly moves toward 2/3 as the initial value of $P(L)$ moves toward 1 and as convergence becomes quicker. And this is just for our initial utilities, again, of $U(S,L) = 0$, $U(S,\text{-}L) = 40$, $U(\text{-}S,L) = 30$, $U(\text{-}S,L) = 30$. Using a wide variety of other values for utilities, convergence was also consistently achieved.

2. Nash Dynamic

We also reran the problem using an alternate decision dynamic: We used the Nash formula for our dynamical update rule. This decision dynamic resulted in the same initial updates regardless of whether we used the original utilities given in Joyce's statement of the problem or our revised nonnegative utilities. It is, apparently, able to handle negative utilities.

The Nash rule is:

$$P(S)_{\text{new}} = (P(\text{S})_{\text{old}} + \text{cov}(S))/(1 + \text{cov}(S) + \text{cov}(\text{-}S)),$$

where cov (A), the covetability of an act A (which in our case is either S or $\text{-}S$), is given by:

$$\text{cov}(A) = \max\{U(A) - U(SQ), 0\},$$

where $U(SQ)$ is given (as in the Bayes case) by:

$$U(SQ) = P(S) * U(S) + P(\text{-}S) * U(\text{-}S)$$

Using, for our initial inputs, $U(S,L) = 0$, $U(S,\text{-}L) = 40$, $U(\text{-}S,L) = 30$, $\text{U}(\text{-}S,L) = 30$, and $\text{P}_0(L) = 0.2$ with this update rule, we found that starting at iteration numbers 24 and 25, we obtained the following convergence-only-on-alternate iterations results:

Odd iteration number:

$P(S) = 0.0692172300957866$

$P(L) = 0.0951333869009216$

Even iteration number:

$P(S) = 0.862430045118893$

$P(L) = 0.582193887353706$

This update rule gave separate convergences for odd and even iterations.

We then tested the dynamics for a wide variety of initial inputs for utilities and initial probability of lesion and continued to get the same type of result, namely separate convergences on the odd versus even iterations, for almost all input values (for nonextremal initial probabilities). In particular, for our initial inputs for utilities as given above, and when initial $P(L) = 0.1$, 0.2, or 0.35, we obtained alternating convergence fairly quickly; when $P(L) = 0.5$, we obtained complete convergence fairly quickly; when initial $P(L) = 0.85$ or 0.9, we obtained extremely slow approach to convergence (likely convergence, but the computer could not handle more iterations); and when initial probability of lesion was set to $P(L) = 0.25$, we obtained immediate convergence to $P(S) = 0.275$. It would be useful to explore whether a general result can be obtained regarding when we achieve no convergence (not even alternating convergence), only alternating convergence, and absolute convergence, and how quickly convergence or alternating convergence is achieved, for the Murder Lesion problem using the Nash update rule, for all utilities and nonextremal probabilities.

Brian Skyrms suggested we try a tempered version of the Nash update rule such that initially dampening 1/2 (move only half way), then to move only 1/4 of the way suggested by the Nash rule, then 1/8, and so forth. This tempered Nash did converge.

3. The Strange Consequence of Joyce's Three Conditions for Choosing among δS, δ-S, and δM

For the Murder Lesion problem that uses $U(S,L) = 0$, $U(S,\text{-}L) = 40$, $U(\text{-}S,L) = 30$, $U(\text{-}S,\text{-}L) = 30$, $P(L) = 0.2$, $P(S|L) = 0.8$, and $P(S|\text{-}L) = 0.1$ as its initial values, and inserting the corresponding fixed-point values we obtained for $P(S)$ and $P(L)$ using the Bayes rule, Joyce's three conditions become the following:[18]

---

18 The details of the application of Joyce's constraints proceed as follows:

a. Calculate $U(S)$, $U(M)$, and $U(\text{-}S)$ as follows:

$U(S) = P(L) * U(S\&L) + P(\text{-}L) * U(S\&\text{-}L)$ (i)

$U(M) = P(L) * U(M\&L) + P(\text{-}L) * U(M\&\text{-}L)$ (ii)

$U(\text{-}S) = P(L) * U(\text{-}S\&L) + P(\text{-}L) * U(\text{-}S\&\text{-}L)$ (iii),

where (1) $P(L)$ = the fixed-point value we obtained in Part A (e.g., 0.25); (2) utilities are the ones we used in Part A (e.g., $U(S\&\text{-}L) = 40$, $U(\text{-}S) = 30$, and $U(S\&L) = 0$; and (3) $U(M\&L) = P(S) * U(S\&L) + P(\text{-}S) *$

$$30.0000000000000 = 40/3 * P(\delta S) + 30.0000000000000 * P(\delta M) + 37.8947368421053 * P(\delta\text{-}S) \quad \text{(i")}$$

$$30.0000000000000 = 24.6428571428572 * P(\delta S) + 30.0000000000000 * P(\delta M) + 32.5375939849624 * P(\delta\text{-}S) \quad \text{(ii")}$$

$$30.0000000000000 = 30 * P(\delta S) + 30.0000000000000 * P(\delta M) + 30.0000000000000 * P(\delta\text{-}S) \quad \text{(iii")}$$

Applying SageMath yields that the solution to the system of three equations (i"), (ii"), and (iii") just above is:

$$P(\delta M) = 1, P(\delta S) = 0, P(\delta\text{-}S) = 0.$$

This result is also yielded by an application of the Maple program. The same results obtained for a wide variety of alternate choices of $P(L)$, $P(S|L)$, and $P(S|\text{-}L)$. We continued to get

$$P(\delta M) = 1, P(\delta S) = 0, P(\delta\text{-}S) = 0,$$

for the alternatives we explored.

These results show that, for the conditions assumed, Joyce's constraints imply picking this mixed strategy *M*, rather than *S* or -*S*, even though all of them have the same expected utility.[19]

$U(\text{-}S\&L)$ (with $P(S)$ = the fixed-point value we obtained in Part A (e.g., 0.321428571428571) and $U(M\&\text{-}L)$ is obtained analogously).

b. Calculate $U(S \mid \delta S) = P(L|S) * U(S\&L) + P(\text{-}L|S) * U(S\&\text{-}L)$, $U(S \mid \delta M) = P(L|M) * U(S\&L) + P(\text{-}L|M) * U(S\&\text{-}L)$, and $U(S \mid \delta\text{-}S) = P(L|\text{-}S) * U(S\&L) + P(\text{-}L|\text{-}S) * U(S\&\text{-}L)$ using the utilities from Step 1a and $P(L)$ and $P(L|S)$ as obtained in Part A, and insert into the following equation:
$U(S) = P(\delta S) * U(S \mid \delta S) + P(\delta M) * U(S \mid \delta M) + P(\delta\text{-}S) * U(S \mid \delta\text{-}S)$ (i')
Do analogously for the following two equations:
$U(M) = P(\delta S) * U(M \mid \delta S) + P(\delta M) * U(M \mid \delta M) + P(\delta\text{-}S) * U(M \mid \delta\text{-}S)$ (ii')
$U(\text{-}S) = P(\delta S) * U(\text{-}S \mid \delta S) + P(\delta M) * U(\text{-}S \mid \delta M) + P(\delta\text{-}S) * U(\text{-}S \mid \delta\text{-}S)$ (iii')

c. Set the value obtained in 1a(i) to the right side of (i') and analogously for (ii) and (iii).

19 For SageMath, the results are extremely sensitive to rounding error: We found that changing the coefficient in its thirteenth decimal place in one of the formulas used generating $P(\delta M) = 1$ in the basic result gives an alternative solution in which $P(\delta M) = 0$. This did not happen with Maple.

Here are the equations resulting from the decrease of a single digit in the thirteenth decimal place of the factor in the first term of the second equation in the above note (see (ii") above):

$$30.0000000000000 = 40/3 * P(\delta S) + 30.000000000000 * P(\delta M) + 37.8947368421053 * P(\delta\text{-}S)$$

APPENDIX 2A

DEATH IN DAMASCUS: CONTINUOUS TIME DYNAMICS

Brian Skyrms

Gibbard and Harper's "Death in Damascus" and Egan's "Murder Lesion" are examples of a special class of decision instability problems. In these problems there are two possible pure acts and two relevant states of the world. Deliberation generates information about the state of the world, which feeds back into deliberation and alters the relative attractiveness of the acts. In these examples, neither pure act is an equilibrium of this process. In Harper's terminology, neither pure act is ratifiable. Under quite mild assumptions about deliberation aiming at doing the best thing, that is, "seeking the good" in Skyrms' deliberational dynamics, the unique deliberational equilibrium in these examples is a mixed state that gives each pure act some positive probability. This can be viewed as a state of indecision, or alternatively as a ratifiable mixed act. One is then naturally led to ask, "Does the process of deliberation get one to such a deliberational equilibrium?" In the decision instability literature in philosophy, as in the larger game theory literature on game dynamics, this is a complex question whose general answer is: "It depends on a lot of things." But for Death in Damascus problems, because of their special qualitative structure, we can give a clean and very general answer. Put informally, it says that for any reasonable continuous dynamics of deliberation, deliberation leads to the ratifiable mixed strategy. This result supports the robustness of the position taken by Joyce. We now state this more precisely, and prove it.

1. Death-in-Damascus-Type Problems

There are two possible pure acts. We are interested in dynamics of the decision maker's degree of belief, x, that she will eventually do act 2. This can take any value from zero to one. We can take the state space here to be the open interval, (0,1). The decision maker's degree of belief that she will eventually do act 2 is stipulated to give her information that

---

30.0000000000000 = 24.642857142857**1** * P(δS) + 30.0000000000000 * P(δ*M*) + 32.5375939849624 * P(δ-*S*)

30.0000000000000 = 30 * P(δS) + 30.0000000000000 * P(δ*M*) + 30.0000000000000 * P(δ-*S*)

These give the following alternate solution:

P(δ*M*) = -28/9 * P(δS) + 1 and P(δ -*S*) = 19/9 * P(δS).

Inserting our fixed point of P(δS) = 0.321428571428571 in this solution results in:

P(δM) = 0 and P(δ-S) = 0.678571428571428.

determines her degree of belief in one of the two states of nature. It is usually assumed that the probabilities of states of nature conditional on act that will be done are fixed so that this determination is by using these conditional probabilities in that calculation of total probabilities of states. Probabilities of states, in turn, determine her expected utility of each of the two acts, E_1, E_2, according to some specified payoffs.

The conditional probabilities and payoffs are specified such that if the decision maker is very confident that she will do act 2, act 1 has higher expected utility. If she is very confident that she will do act 1, act 2 has higher expected utility. There is a unique interior point, x^*, where the expected utilities of acts 1 and 2 are equal.

$$\text{When } x < x^*, \text{ EU (Act 2)} > \text{EU (Act 1)}$$

$$\text{When } x > x^*, \text{ EU (Act 2)} < \text{EU (Act 1)}$$

$$\text{When } x = x^*, \text{ EU (Act 2)} = \text{EU (Act 1)}$$

This is all we need for a qualitative specification of Death-in-Damascus-type problems.

2. Deliberational Dynamics

In these problems there is a feedback process, degrees of belief about the act that will be done influence degrees of belief about the true state of nature, which influences degrees of belief about the act that will be done. We model this as a continuous process. We only need qualitative features of the dynamics, so our analysis covers a broad class of dynamics.

We consider autonomous dynamics. That is to say a point, x, determines the rate of change of x, dx/dt, at x, via this feedback process. There is a function, f, such that:

$$dx/dt = f(x).$$

In order to rule out pathology, we assume that f is smooth. Then our dynamics is a flow. Any point in our space generates an orbit tracing out where the dynamics goes from that initial condition.

We assume that the probability of the more attractive act increases and if neither is more attractive probabilities do not change. All the dynamics of interest have this property. This is all we need to qualitatively specify flows that "seek the good" in our problems.

3. Convergence to the Ratifiable Mixed Equilibrium

In Death-in Damascus-type problems, any flow that seeks the good converges to x^* from every point in (0,1).

Proof: It is convenient to change the state variable to y, where $y = x - x^*$, so our equilibrium is at $y = 0$. From the two previous sections we have as qualitative properties of any flow that seeks the good in any Death-in-Damascus-type problem.

$$\text{If } y < 0, \text{ then } dy/dt > 0.$$

$$\text{If } y > 0, \text{ then } dy/dt < 0.$$

$$\text{If } y = 0, \text{ then } dy/dt = 0.$$

The function y^2 is a strict global Lyapunov function.[20]

(i) It is smooth.
(ii) It is zero at $y = 0$ and positive everywhere else (positive definite).
(iii) Its time derivative is $2y * dy/dt$. By the foregoing, this is 0 when $y = 0$ and negative elsewhere (negative definite).

The equilibrium at $y = 0$ is globally asymptotically stable by Lyapunov's theorem.[21]

APPENDIX 2B

DEATH IN DAMASCUS: TEMPERED DISCRETE TIME DYNAMICS

Brian Skyrms

Let the probability of doing Act 2 be x. We are interested in discrete time dynamics that maps x to x'. Our state space will be the interval from zero to one. For one class of dynamics, it will be the open interval; for another, it can be the closed interval.

There is a single interior equilibrium at some point x^*. An adaptive dynamics moves in the direction of the act that currently has higher expected utility. Thus:

$$\text{At any } x > x^*, \text{ the dynamics jumps to an } x' < x.$$

$$\text{At any } x < x^* \text{ the dynamics jumps to } x' > x.$$

20 Radial unboundedness is not relevant because of the nature of the state space.
21 See Hirsch et al. (2004).

$$\text{At } x = x^*,\ x' = x^*.$$

In contrast to continuous time dynamics, the discrete time dynamics introduces an additional complication. Discrete time dynamics can jump over x^*, and possibly get trapped in cycles. Convergence to x^* is not assured.

For an extreme example, consider the best-response dynamics, where x' is state that has the highest expected utility from the standpoint of x. The equilibrium, x^*, is a stationary point of this dynamics. Starting there generates the sequence

$$x^*, x^*, x^*, x^*, \ldots$$

But starting at some $x > x^*$, best-response dynamics generates a sequence:

$$x, 0,1,0,1,0,1,0,1, \ldots$$

and starting at some $x < x^*$ it generates the sequence:

$$x, 1,0,1,0,1,0,1,0, \ldots$$

It would seem strange if a deliberator would not notice such a cycle, and would just continue on and on. We may, then, be led to consider a modified dynamics that tempers jumps in response to encountering a cycle. To begin with, we consider a more or less natural tempering of best response that leads to convergence to the equilibrium. The same tempering leads to convergence in a large class of adaptive dynamics for Death-in-Damascus-type problems. (Nothing is claimed here for other problems.)

We modify our dynamics by letting our deliberator have a little memory, notice she is in a qualitative cycle, and respond with a tempered response.

We have a *qualitative cycle* when we jump from one side of the equilibrium to the other and then jump back over to the original side. The perceived optimal act goes from Act 1 to Act 2 and back to Act 1, or from Act 2 to Act 1 back to Act 2.

We modify our base dynamics by taking some fraction in (0,1) of the previous jump size to be the new jump size when we have gone through a qualitative cycle. Call the fraction the *tempering factor, a*. After one qualitative cycle, if best response moves delta x, we now move a smaller distance, *a* delta x. After n cycles, we move a^n delta x.

For an example, consider tempered best-response dynamics; let the equilibrium be $x^* = .8$, and let the tempering factor $a = 1/2$.

We start at $x = .5$ (but could start anywhere).

	Value of x	Best response
t1	.5	1
t2	1	0
t3	0	1 (Cycle) (move ½ way toward 1)
t4	.5 (half)	1————— (move ½ way toward 1)
t5	.75 (half)	1————— (move ½ way toward 1)
t6	.875 (half)	0 (Cycle) (move 1/4 way toward 0)

and so forth.

We now prove that this tempered best-response dynamics will converge to x* for any tempering constant, *a*, in (0,1).

Suppose there are only a finite number of jumps. Then you have hit x*, since it is the only rest point.

Suppose there are an infinite number of jumps.

Then either:

(1) after some finite number of jumps, all subsequent points are on one side of x*

or

(2) the dynamics jumps over x* an infinite number of times.

Suppose (1). Consider the case where, after some time, all points, x, are greater than x*. This gives a monotone decreasing sequence of points, since for each point best response is 0. Each jump must move closer to 0, and thus closer to x*. By hypothesis, the sequence has x* as a lower bound. Thus, the sequence must converge to something, either x* or something greater than x*.

It cannot converge to any x# > x*. The tempering factor remains constant, because we are not crossing over x*. x'-x is negative at x#, since x# > x*. The map is continuous at x#, since tempered best response is continuous everywhere except at x*. Thus, x'-x is negative and bounded away from zero in some neighborhood of x#. The sequence can't converge to x#.

It must converge to something and it cannot converge to anything but x*.

Likewise, if after some time all points are less than x*, the sequence must converge to x*.

Suppose (2). We jump over x* an infinite number of times. Then there are an infinite number of qualitative cycles, shrinking the distance of the jumps, |x'-x|, to zero. Since we continue jumping over x* with arbitrarily small jumps, we must converge to x*.

This convergence result for Death in Damascus problems generalizes to a large class of tempered discrete-time dynamics. These include, as base dynamics, well-known dynamics such as Nash dynamics on the closed interval [0,1] and Replicator dynamics on the open interval (0,1). (Note: Replicator is not adaptive at the endpoints.)

Consider any discrete time dynamics that maps any x in the domain to x' such that:

(i) It is adaptive, that is:
If x is not a best response to itself, then x' is closer to the best response to x than x is.
If x is a best response to itself, x' = x.
(ii) It is continuous, except possibly at the mixed equilibrium x*.

Then any tempered version of that dynamics starting at any point in the domain converges to x*.

Proof: The tempered version inherits continuity from the original version. Then the proof proceeds just as in the foregoing.

References

Adams, E. (1970). "Subjunctive and Indicative Conditionals." *Foundations of Language* 6: 39–94.

Ahmed, A. (2014). "Dicing with Death." *Analysis* 74, no. 4: 587–92.

———, ed. (Forthcoming). *Newcomb's Problem*. Oxford: Oxford University Press.

Arntzenius, F. (2008). "No Regrets, or Edith Piaf Revamps Decision Theory." *Erkenntnis* 68: 277–97.

Cantwell, J. (2010). "On an Alleged Counter-Example to Causal Decision Theory." *Synthese* 173: 127–52.

Egan, A. (2007). "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116, no. 1: 93–114.

Gibbard, A. (1980). "Two Recent Theories of Conditionals." In *Ifs*, edited by W. L. Harper, R. Stalnaker, and G. Pearce, 211–47. Dordrecht: Reidel.

——— (1992). "Weakly Self-Ratifying Strategies: Comments on McClennen." *Philosophical Studies* 65: 217–25.

Gibbard, A., and W. L. Harper (1978). "Counterfactuals and Two Kinds of Expected Utility Theory." In *Foundations and Applications of Decision Theory*, edited by C. Hooker, J. Leach, and E. McClennen, 125–62. Dordrecht: Reidel.

Gilovich, T., D. Griffin, and D. Kahneman (2002). *Heuristics and Biases: The Psychology of Intuitive Judgement*. Cambridge: Cambridge University Press.

Harper, W. L. (1993). "Causal and Evidential Expectations in Strategic Settings." *Philosophical Topics* 21, no. 1: 79–97.

Harper, W. L., R. Stalnaker, and G. Pearce, eds. (1980). *Ifs*. Dordrecht: Reidel.

Harper, W. L., S. J. Chow, and G. Murray (2012). Bayesian Chance. *Synthese* 186: 447–74.

Hirsch, M. W., S. Smale, and R. L. Devaney (2004). *Differential Equations, Dynamical Systems and an Introduction to Chaos*. San Diego, CA: Academic Press/Elsevier.

Hofstadter, D. R. (1983). "Metamagical Themas: The Calculus of Cooperation Is Tested through a Lottery." *Scientific American* 248: 6, 122–31, 154.

Joyce, J. M. (2012). "Regret and Instability in causal decision theory." *Synthese* 187: 123–45.

——— (Forthcoming). "Deliberation and Stability in Newcomb Problems and Pseudo-Newcomb problems," in *Newcomb's Problem*, Arif Ahmed (ed.), Cambridge University Press, 2018, pp. 134–59.

Lewis, D. (1976). "Probabilities of Conditionals and Conditional Probabilities." *Philosophical Review* LXXXV, no. 3: 297–315.

Luce, R. D., and H. Raiffa (1957). *Games and Decisions*. New York: John Wiley & Sons.

Maynard Smith, J. (1982). *Evolution and the Theory of Games*. Cambridge: Cambridge University Press.

Skyrms, B. (1990). *The Dynamics of Rational Deliberation*. Cambridge: Harvard University Press.

———. (2013). "The Core Theory of Subjunctive Conditionals." *Synthese* 190: 923–28.

Shafir, E., and A. Tversky. (1995). "Decision Making." In *An Invitation to Cognitive Science, Second Edition* (*Volume 3: Thinking*), edited by E. E. Smith, E. E. and Osherson, 77–100. Cambridge: MIT Press.

Stalnaker, R. (1980a). "Letter to David Lewis May 21, 1972." In Harper et al., eds., 151–52.

———. (1980b). "Indicative Conditionals." In Harper et al., eds., 193–210.

Tversky, A., and D. Kahneman. (1974). "Judgment under Uncertainty: Heuristics and Biases." *Science*, New Series 185, no. 4157: 1124–31.

2

COUNTERFACTUALS AND THE GIBBARD–HARPER COLLAPSE LEMMA

Melissa Fusco

COLUMBIA UNIVERSITY

Gibbard and Harper (1978) provides a classic statement of Causal Decision Theory ("CDT"), which uses counterfactual conditionals to express the causal relationships that are, according to CDT, of particular relevance to rational decision-making. The account builds a bridge between decision theory and the semantics of natural language counterfactuals, active at the time in the work of Lewis and Stalnaker and still vibrant today.[1]

CDT's rival in the dialectic in which Gibbard and Harper are situated is Evidential Decision Theory ("EDT"). While EDT, like CDT, holds that any choiceworthy act is one which maximizes expected utility, EDT employs act-conditionalized probabilities in the calculation of expected utility. This is typically conceived of as an attitude of austerity towards the causation-correlation distinction: while it may be perfectly real, it has no important direct role to play in a theory of decision.

Classic CDT—in particular, the Gibbard-Harper formulation of it—has enjoyed wide acceptance. Many in the recent literature, however, hold that tides are turning. One factor in the sea-change is an influential 2007 paper by Andy Egan (Egan 2007), which presents several counterexamples to the theory. On Egan's telling, causal decision theorists—and he does have in mind those who appeal to the counterfactual formulation of the theory[2]—adhere

1 Gibbard and Harper cite, in particular, Lewis (1973) and Stalnaker (1968). For recent work in this tradition, see, *inter alia*, Ahmed (2013) and Kment (2019).

2 See Egan (2007, 95).

to the motto "do whatever has the best expected outcome, holding fixed your initial views about the likely causal structure of the world" (96). However, Egan argues, there are cases where agents should *not* hold such initial views fixed as they act. In such cases, agents should use their anticipated future causal views instead, taking into account what they expect to *learn* by performing the very act in question.

In this paper, I focus on the dialectic from the CDTer's point of view, with an eye to a formal result pointed out by Gibbard and Harper in the third section of their classic paper. There, they show that if an agent's credences are probabilistically coherent, and the semantics for counterfactuals obeys Strong Centering—roughly, the view that each possible world is counterfactually closest to itself—then the probability of (a counterfactual conditional on its antecedent) simplifies to the probability of (its consequent, given its antecedent). This has the eyebrow-raising consequence that "Eganized" causal decision theory, the view on which agents anticipate their future causal views, recommends an act just in case classical *evidential* decision theory does.

The "collapse", as I call it, complicates the traditional way of glossing the relationship between EDT, CDT, and diachronic coherence norms. I canvas three takes on Gibbard and Harper's discussion of the result in §4 below, arguing in favor of one which emphasizes the peculiarity of predicaments involving choosing one's own evidence. This peculiarity raises doubts about whether update in Egan-type cases can aspire to the status of probabilistic *knowledge*, in the sense of Moss (2013a, 2018). I suggest that it does not, and explain why this consideration both points the way to understanding the true significance of the collapse, and functions as a defense of CDT against Egan's counterexamples.

EDT versus CDT: an overview

Decision Problems

Both classical EDT and classical CDT begin with the thought that the value of each of a set of available acts (call them the *a*'s) can be calculated by identifying a set of states which fix one's welfare (call them the *s*'s) and then multiplying the utility of each state-act conjunction by one's subjective probability, or *credence*, that that state obtains.

For example, suppose that Otto's tennis match is today. Calliope is offered a bet on his winning at even odds for a dollar: she can either *bet on Otto* (henceforth B) or decline the bet (= $\neg B$). The payoff of the bet depends on whether *Otto wins* (= W) or not (= $\neg W$). In matrix form, her possible payoffs look like this:

	***Otto wins* (W)**	***Otto loses* ($\neg W$)**
Bet (B)	\$1	-\$1
Don't bet ($\neg B$)	\$0	\$0

According to both decision theories, Calliope is facing a *decision problem* in which her goal is *maximize expected utility*. Her noninstrumental desires equip her with a value function $Val(\cdot)$ over states called *outcomes*, which following standard idealization I will assume does not change over time, and is such that $Val(\$nk) = kVal(\$n)$. Her decision problem at a time t can be represented as a triple $\langle Cr^t, \mathscr{A}, \mathscr{S}\rangle$, where

> $Cr^t(\cdot)$ is a credence function over the state space W, here representing Calliope's subjective confidence in a variety of propositions at t;
> $\mathscr{A} = \{a_1, \ldots, a_n\}$ is a partition of W into Calliope's available *acts* at t, and
> $\mathscr{S} = \{s_1, \ldots, s_m\}$ is a partition of W into admissible *states of nature*, where a partition is admissible only if each act-state conjunction $(a_i \wedge s_j)$ determines some number $Val(a_i \wedge s_j)$ (unique up to positive affine transformation) under the value function.[3]

Epistemologists of varying stripes will take the first element of the decision problem, the agent's credence function $Cr^t(\cdot)$, to be subject to a variety of *epistemic norms*. Of particular note are, first, **Probabilism**: $Cr^t(\cdot)$ should be a probability function. Second, $Cr(\cdot)$ is commonly taken to be governed by **Conditionalization**, a diachronic norm which concerns how the agent should respond to new information. Conditionalization states that an agent who, between times t and t^+, learns exactly E, should adopt the posterior credence function $Cr^{t^+}(\cdot) = Cr^t(\cdot|E)$, where this is defined.[4]

For our purposes, a more general norm is also worth mentioning, which extends Conditionalization to cases where a learning experience does not result in an agent's becoming *certain* of any proposition E. **Jeffrey Conditionalization** states that an agent who, between times t and t^+, undergoes a learning experience that directly alters her credences in members of the partition $\{E_i\}$ from $Cr^t(E_i)$ to $Cr^{t+}(E_i)$ should adopt the posterior credence function $Cr^{t+}(\cdot) = \sum_i Cr^{t+}(E_i)Cr^t(\cdot \mid E_i)$. (Jeffrey Conditionalization will become relevant in §4.2 below.) There may be—and discussions in the literature often presuppose that there are—further, less purely subjective norms governing credence functions, such as that they are based on reasonable priors, sufficiently sensitive to the observed frequencies of events, and so on.

---

3 In what follows, I will speak of the members of $\mathscr{A}$, $\mathscr{S}$, and the domain of $Cr^t(\cdot)$ alike as *propositions*. This differs from Savage's original picture of the relevant primitives, on which acts are *functions* from states to outcomes. See Joyce (1999, Ch. 2) for discussion of this shift.

4 Conditional probabilities of the form $Pr(A \mid B)$ are customarily defined, via the "Ratio Formula", to be $Pr(A \wedge B) / Pr(B)$. On primitive treatments of conditional probability, such as Spohn (1986), the Ratio Formula equality does not *define* conditional probability, but holds whenever $Pr(B) \neq 0$.

The EDT Branch

EDT and CDT's common starting point is Savage's (1972) notion of expected utility, which is simply the generic statistical notion of expected value, applied to the value function.[5] His theory says: *in an uncertain world, estimate the value of act a_j by taking the sum of its value in each state of nature, weighted by one's current estimate that that state obtains.*

Equation 1 (Savage Expected Utility). $SEU^t(a_j) = \sum_i Cr^t(s_i)Val(a_j \wedge s_i)$.

Savage's decision rule: maximize expected utility at t by choosing an act $a_j \in A$ such that $SEU^t(a)$ is maximal.[6]

Savage's theory entails the validity of *dominance arguments* in favor of a particular acts. An act a_j dominates all other acts $a_i \in \mathscr{A}$ when, for all states s, $Val(s \wedge a_j) > Val(s \wedge a_i)$. The payoff of a_j is thus greater than the payoff of *any* other a_i in *any* state, and Savage's theory will recommend a_j.

While granting that Savage's norm works for some decision problems, both CDT and EDT move away from it as a general decision rule. One way to see why is to observe that there are cases in which the agent believes the likelihood of a state W (Otto's winning) *depends* on whether she performs an act like B (taking the bet).

Recall that, in Calliope's case, she should take the bet on Otto at t just in case the expected utility of B exceeds the expected utility of $\neg B$. According to Savage's theory, this happens just in case $SEU^t(B) > SEU^t(\neg B)$, which happens in this instance just in case $Cr^t(win) > Cr^t(\neg win)$.[7] Suppose Calliope knows that Otto's confidence increases whenever he sees her betting on him: based on her data, he has a 70% chance of winning if she accepts B, but only a 40% chance otherwise. Assuming Probabilism, Calliope's current best estimate of Otto's likelihood of winning, $Cr^t(W)$, is the weighted average of the probability that he wins, *given that she bets on him*, and the probability that he wins, *given that she doesn't*, where the "weights" are her unconditional credences in her own acts B and $\neg B$, respectively:

$$Cr^t(W) = Cr^t(W \wedge B) + Cr^t(W \wedge \neg B)$$

---

5 Where F is a function, the generic notion of expected value says that $E[F] = \sum_i f_i Pr(F = f_i)$. Here, our probability function is the subjective credence function $Cr^t(\cdot)$, and the function F is $Val(\cdot)$ across act-state pairs, where the act is held fixed.

6 That is, such that $SEU^t(a_j) \geq SEU^t(a_i)$, for any $a_i \in \mathscr{A}$.

7 Calculation:

$$\begin{aligned}
& SEU^t(B) > SEU^t(\neg B) \\
\textit{iff} \quad & SEU^t(B) > 0 \ (\textit{no money changes hands if Calliope declines to bet}) \\
\textit{iff} \quad & \sum_i Cr^t(s_i)Val(B \wedge s_i) > 0 \\
\textit{iff} \quad & Cr^t(W) \times 1 + Cr^t(\neg W) \times -1 > 0 \\
\textit{iff} \quad & Cr^t(W) > Cr^t(\neg W).
\end{aligned}$$

$$= Cr(W \mid B)Cr^t(B) + Cr^t(W \mid \neg B)Cr^t(\neg B)$$

$$= .7 \times Cr^t(B) + .4 \times Cr^t(\neg B)$$

But Calliope doesn't usually place bets, so her initial credence in B is only 20%. This attaches a small weight to the highish probability she assigns to Otto's winning conditional on her bet—and a *large* weight to the lowish probability she assigns to Otto's winning conditional on her *not* betting. A flatfooted application of Savage's theory will therefore recommend against B. This seems obviously wrong. In using Savage's equation to assign expected utility to B, Calliope is improperly diluting the probability assigned to Otto's winning by including in her calculations worlds where she doesn't take the bet. Calliope wants to know what the expected utility of *the act of betting on Otto* is; in that case, though, it is certainly true—with probability 1, not probability .2—that B occurs.

Savage's own response to this problem was to require that decision problems be formulated with a special partition $\mathscr{S}^*$ of states that are known to be *independent* of the agent's contemplated acts.[8] But it isn't always clear that such an $\mathscr{S}^*$ can be found. The evidential decision theorist takes a different path: when assigning expected utility to an option a, use any partition $\mathscr{S}$ you like, but do not use your current credence in $s_i \in \mathscr{S}$ rather, use your credence in s_i, *conditional on* a.

Equation 2 (Evidential Expected Utility). $\mathcal{V}^t(a_j) = \sum_i Cr^t(s_i \mid a_j)Val(a_j \wedge s_i)$.

Conditional decision rule: maximize expected utility by choosing an act $a \in \mathscr{A}$ such that $V^t(a)$ is maximal.

With respect to the example above, this change removes the problematic weighting by Calliope's (low) current credence in B. The factor by which $Cr^t(W \mid B)$ is now weighted is simply 1. Given this way of calculating expected utility, the result for B is positive, so EDT recommends that Calliope take the bet.

The CDT Branch

In the case of Calliope and Otto, it is natural to assume that B (Calliope's bet) and W (Otto's winning) are probabilistically correlated because Calliope's bet on Otto *causally conduces* to his winning (perhaps by increasing his confidence). But not all correlation is causation. CDT diverges from EDT by insisting that causal information be represented *separately* from evidential support, and appealing to act-conditioned probabilities *only* when evidential support is also causal.

---

8 Is the relevant type of independence *causal*, or *evidential*? At stake is, once again, the difference between EDT and CDT. My understanding is that there is controversy over which form of independence Savage intended (see e.g. Jeffrey (1983, pgs, 21-22) and Joyce (1999, Ch. 4.1).

Here is the sort of case where the two theories diverge. Suppose Adeimantus is hoping to get an REI jacket from his mother for Christmas (= *J*). He begins to contemplate *leaving REI catalogues around the house* (= *C*), on the grounds that, statistically, houses full of REI catalogues are more likely to have REI jackets inside. If buying catalogues and planting them in the house costs 1 utile, Adeimantus's outcomes look like this:

	Jacket (*J*)	**No jacket (¬*J*)**
Catalogues (*C*)	9	0
No catalogues (¬*C*)	10	1

$Cr^t(J \mid C) > Cr^t(J)$

A calculation by the conditional decision norm will recommend that Adeimantus see to it that there are catalogues around the house, so long as *C* raises the statistical probability of *J* by more than 1/9.[9]

Suppose, however, that Adeimantus believes his mother has *already* purchased the gift at time *t*. In this case, it isn't clear that EDT makes the right recommendation. The causal decision theorist will grant that, since $Cr^t(J \mid C) > Cr^t(J)$, it would be good for Adeimantus to spontaneously *discover*, or *receive the news*, that there are REI catalogues around the house. That is why the CDTer calls the EDTer's decision-making quantity, $\mathcal{V}^t(\cdot)$, a "news value" function.[10] The problem, the CDTer says, is that *receiving the news* that there are REI catalogues around the house should be treated differently than *making it the case* that the very same thing obtains. What matters for a decision problem is how likely an available act *a* is

9 Calculation:

$$\begin{aligned} \mathcal{V}^t(C) &> \mathcal{V}^t(\neg C) \text{ iff} \\ \sum_i Cr^t(k_i \mid C) Val(k_i \wedge C) &> \sum_i Cr^t(k_i \mid \neg C) Val(k_i \wedge \neg C) \text{ iff} \\ [Cr^t(J \mid C) \times 9 + Cr^t(\neg J \mid C) \times 0] &> [Cr^t(J \mid \neg C) \times 10 + Cr^t(\neg J \mid \neg C) \times 1] \text{ iff} \\ [Cr^t(J \mid C) \times 9] &> [Cr^t(J \mid \neg C) \times 10 + Cr^t(\neg J \mid \neg C)] \end{aligned}$$

Since $Cr^t(\cdot \mid C)$ and $Cr^t(\cdot \mid \neg C)$ are themselves probability functions, this is equivalent to

$$> [Cr^t(J \mid \neg C) \times 10 + (1 - Cr^t(J \mid \neg C))]$$

Letting $n = Cr^t(J \mid C)$ and $m = Cr^t(J \mid \neg C)$, this obtains just in case

$9n > 10m + (1 - m)$ iff
$n > m + 1/9$.

10 For use of the term "news value" to describe $\mathcal{V}(a)$, see e.g. Lewis (1981), Gibbard and Harper (1978), and Joyce (1999).

to *cause* some state such as *J* that the agent desires. And as Adeimantus himself believes, it cannot cause that state, since all gifts have already been purchased.

In light of cases like this, Gibbard and Harper advance a different utility-maximizing equation, wherein the relevant subjective probability is $Cr(a_j \rightarrow s_k)$.

I will call the object of the agent's credence here the *act-counterfactual* $a_j \rightarrow s_k$, and follow Gibbard and Harper in reading it as the subjective probability that *if act* a_i *were performed, state* s_k *would obtain*:

Equation 3 (Causal Expected Utility). $\mathscr{U}^t\left(a_j\right) = \sum_i Cr^t\left(a_j \rightarrow s_i\right) Val\left(a_j \wedge s_i\right)$

Causal decision rule: maximize expected utility by choosing an act $a \in \mathscr{A}$ such that $\mathscr{U}^t(a)$ is maximal.

In our example, Adeimantus's belief that leaving REI catalogues around the house at *t* has no causal influence over whether *J* obtains is reflected in the fact that his credences satisfy what we will call the *Counterfactual Independence Criterion*:

Definition 1 (Counterfactual Independence Criterion). An agent believes, at *t*, that act *a* has no causal power over state *s* iff $Cr^t(s) = Cr^t(a \rightarrow s) = Cr^t(\neg a \rightarrow s)$.[11]

In the case at hand, Adeimantus's credences are such that $Cr^t(J) = Cr^t(C \rightarrow J) = Cr^t(\neg C \rightarrow J)$. Likewise, $Cr^t(\neg J) = Cr^t(C \rightarrow \neg J) = Cr^t(\neg C \rightarrow \neg J)$.

The CDTer will thus accept a Savage-style dominance argument to the effect that, no matter what Adeimantus's time-*t* credence in *J* is, he should refrain from planting catalogues in the house.[12] He should save himself the one-utile effort of bringing about *C*.

Newcomb Problems and NC Problems

We are now in a position to describe *NC Problems*, of interest to us because they are the simplest cases in which EDT and CDT disagree "on the ground." For our purposes, an NC problem will be a problem with two acts $\{a, \neg a\}$ and two states $\{s, \neg s\}$ for which the Counterfactual Independence Criterion holds with respect to credence function $Cr(\cdot)$ at time *t*. Nonetheless, the agent's relevant conditional credences—perhaps because she defers to reliability—are such that *a* is a very good indication of *s*: $Cr^t(s \mid a) \gg Cr^t(s) \gg Cr^t(s \mid \neg a)$. In such a situation, it is easy to engineer the stakes so that CDT and EDT come apart. It is

11 See Gibbard and Harper (1978, pg. 136, paragraph 2). Note that for *conditional* probabilities, that $Cr^t(s) = Cr^t(s \mid a)$ entails that $Cr^t(s) = Cr^t(s \mid \neg a)$. The analogous entailment holds for counterfactuals given Gibbard and Harper's Axiom 2 (op. cit., pg. 128), but fails easily on Lewis (1973)'s more general treatment of the semantics for counterfactuals.

12 This follows by Equation 3, Counterfactual Independence Criterion, and the dominance structure of the payoff matrix.

sufficient, for example, to "sweeten" $\neg a$, making it slightly better ($+\Delta$) than a both in s and in $\neg s$, while making state s considerably more valuable ($+\Theta$) than $\neg s$, no matter whether a holds:

	s	$\neg s$
a	Θ	0
$\neg a$	$\Theta + \Delta$	Δ

The high conditional probability of s, given a, gives the EDTer decisive reason to choose a in this problem, the lack of causal influence between a and s notwithstanding. On the other hand, the sweetener Δ and the causal independence of $\neg a$ from s gives the CDTer decisive reason to choose $\neg a$, the statistical support $\neg a$ lends to $\neg s$ notwithstanding.

The classic statement an NC Problem, of course, involves a wizardly predictor:

> *Newcomb's Puzzle*. There are two boxes before you, a large opaque box and a small clear box containing $1,000. You may take both boxes (= 2*B*), or take just the opaque box (= 1*B*), keeping whatever is inside the box(es) you take. But: an uncannily accurate predictor has put either $1 million or $0 in the opaque box. She has put the million in the opaque box just in case she predicted you would take one box (= *P*1), and withheld the million just in case she predicted you would take both boxes (= *P*2). (Nozick, 1969; Gibbard and Harper 1978, Sec. 10)

	***P*1**	***P*2**
1*B*	$1 million	$0 million
2*B*	$1.001 million	$0.001 million

$$Cr^t(P1 \mid 1B) \gg Cr^t(P1) \gg Cr^t(P1 \mid 2B)$$

Egan's Case, and its diachronic bite

While Egan argues that CDT is the wrong theory of decision, he concedes that it delivers the right verdict in Newcomb's puzzle (Egan, 2007, pg. 94): in light of the fact that the million dollars has already been distributed (or withheld), it is better to take both boxes. However, he disagrees with similar reasoning in other cases. His counterexample to Gibbard and Harper-style CDT goes as follows:[13]

> *Murder Lesion*. Mary is deliberating about whether to shoot the tyrant Alfred. She would prefer to shoot him, but only if she will hit him, rather than miss him. Mary has good evidence that a certain kind of brain lesion, which she may or may not have, causes murderous

13 This section, and the next, reproduce at slightly greater length the arguments given in Fusco (2017).

tendencies but also causes shooters to have bad aim. Mary currently has high credence that she has good aim. But (like Calliope in §1) she assigns low credence to the proposition that she will act.

	Don't hit ($\neg H$)	**Hit (H)**
Shoot (S)	terrible	good
Don't shoot ($\neg S$)	neutral	*(impossible)*

In the Murder Lesion case, the available acts are: shoot or don't (S,$\neg S$) and the basic states are: hit or don't (H,$\neg H$). However, Mary's knowledge includes information about causal influence: she has conditional and unconditional subjective probabilities on well-formed formulas like $S \rightarrow H$—*if I were to shoot Alfred, I would hit him*. Egan suggests that the set of states used to calculate Mary's expected utility should be

$$\{S \rightarrow \neg H, S \rightarrow H\}$$

rather than $\{H, \neg H\}$. This gives rise to a second matrix:

	$S \rightarrow \neg H$	$S \rightarrow H$
S	terrible	good
$\neg S$	neutral	neutral

However, the second matrix has a striking feature: although Mary takes the causal relationships across the top to be causally independent of her acts, her credences in them will change drastically if she actually shoots. Egan argues on this basis that the act-conditional probability relevant to the calculation of expected utility is given by the more complex formula (*):

$$(*)\ Cr\left(\left(a_i \rightarrow s_k\right)|a_i\right)$$

In (*), a_i appears twice, and both subjunctive and evidential probability are invoked. (*) should be read as: *the subjective probability that (e.g.) if Mary were to shoot Alfred, she would hit him, given that she shoots.* I will henceforth call (*) "the Egan credence on s_k in a_i." Calculating the expected utility of an act a_i with Egan credences in state-act pairs yields a quantity we can call $\mathcal{U}_{\text{Egan}}(a_i)$, or $U_{\text{E}}(a_i)$ for short:

Equation 4. $U_E^t\left(a_i\right) = \sum_k Cr^t\left(a_i \rightarrow s|a_i\right)Val\left(s_k \wedge a\right)$

Eganized causal decision rule: Maximize expected utility by choosing an act $a \in \mathcal{A}$ such that $U_E^t\left(a\right)$ is maximal.

Egan argues that, intuitively, Mary should *not* shoot in *Murder Lesion*, and that the equation for U_E delivers this result. As Mary confronts her decision, $Cr(S)$, by description of the case, is low. Therefore, $Cr(S \rightarrow H)$ is high, since she is relatively confident she does not have the brain lesion. Finally, the Egan credence $(\star) = Cr\big((S \rightarrow H)|S\big)$ is low, since once Mary conditionalizes on *shoot*, she is quite confident she has the lesion, and the lesion causes bad aim. Applying the Eganized Causal Credence norm, we get Egan's favored answer, which is that shooting has a low expected utility. By contrast, classical CDT uses the unconditional probability of $S \rightarrow H$, which—wrongly, Egan says—predicts the expected utility of shooting to be high.

Having presented these examples, it is worth briefly reviewing the dialectic. Egan *follows* Gibbard and Harper in holding that counterfactuals have a role to play in describing the subjective probability relevant to the expected value of an act a given a partition $\mathscr{S}$. Although he frames his case as a counterexample to classical CDT, the case involves an agent who has credences concerning causal influence—credences that are appealed to in deliberation.

What Egan's considerations add, though, is the diachronic norm of Conditionalization, previously mentioned but hitherto not explicitly invoked. Conditionalization entails that an agent's conditional credences remain constant over the course of a learning experience on E, a feature known as *rigidity*:

Fact 1. **(Conditional Rigidity)**. As an ideal agent undergoes a learning experience on E between any t and $t^+ \geq t$,

$$Cr^t(\phi \mid E) = Cr^{t^+}(\phi \mid E)$$

By contrast, and as Gibbard and Harper emphasize, the CDTer's proxy for credence in the counterfactual, ⌜if I *were* to do a, then it *would* be the case that s⌝ is *non-rigid*: it varies with $Cr(s)$. This is so even if, intuitively, the agent keeps her views on causal relations constant—not explicitly changing her mind, that is, about *what causes what*.[14]

This is essential to the role that act counterfactuals play in the Gibbard-Harper formulation of CDT. Recall that according to the Counterfactual Independence Criterion, $Cr^t(a \rightarrow s)$ and $Cr^t(\neg a \rightarrow s)$ are set equal to the prior $Cr^t(s)$ *in virtue of the fact* that the agent takes a to have no causal influence over whether s obtains. Because the act-counterfactuals at a time

14 For example, take their case of King Solomon and Bathsheba (op cit., pg 135). Solomon believes doing unjust things, like sending for Bathsheba (B), would indicate (though not cause) an underlying state, lack of charisma, that foretells revolt (R). They write that "[s]ince [Solomon] knows that B's holding would in no way tend to bring about R's holding, he always ascribes the same probability to $B \rightarrow R$ as to R" (136). However, $Pr(B \rightarrow R \mid B) > Pr(B \rightarrow R)$ (pg. 136): that sending for Bathsheba (B) would result in a revolt (R) is more likely if you *do* send for her.

t are in this way set equal to $Cr(s)$ at time t, they are vulnerable to fluctuations via update in $Cr(s)$. In NC Problems $Cr(s)$ is *itself* probabilistically tied to $Cr(a)$, through Conditional Rigidity and the fact that $Cr^t(s \mid a) \gg Cr^t(s)$. The upshot is that rigidity *for counterfactuals* fails over learning experiences on acts:

Fact 2. (Counterfactual non-Rigidity). It is not in general the case that, as an ideal agent undergoes a learning experience on E between any t and $t+ \geq t$,

$$Cr^t(E \rightarrow \phi) = Cr^{t+}(E \rightarrow \phi)$$

This contrast distills the real motivation behind Egan's point against classical CDT. The norm of Conditionalization seems to entail:

(C^*) In decision problems, one should anticipate rationally updating act counterfactuals $a_i \rightarrow s_k$ by conditionalization on one's chosen act.

If (C*) is correct, there would seem to be a strong argument in favor of "Eganized" CDT over classical CDT. The force of Egan's argument comes from the thought that, in situations where an agent expects to get more information as time passes, she should regard her evidence-conditioned credences as better-informed than her current ones. Egan, in effect, asks: should this not be the case for our future credences *in act-counterfactual propositions*, as well as everything else? Assuming an agent in a decision problem is generally self-aware, she can anticipate what her future credence in $a \rightarrow s_k$ should be, given that she undertakes a.[15] By Conditionalization, this more informed credence is just the *current* Egan credence on s_k in a.[16] Thus reaching for Egan credences, instead of her current act-counterfactual credences, in assessing the utility of acts seems like common sense: an application of Jeffrey's appealing claim that a decision-maker should, in general, "choose for the person [she] expect[s] to be when [she has] chosen" (Jeffrey, 1983, pg. 16).

---

15 More carefully, by an agent's being "self-aware", we are assuming that *if the agent performs a at t^+, she becomes certain of it*: $Cr^{t^+}(a) = 1$.

16 Argument: the agent expects that if she brings about a, she will learn (viz., come to have time-t^+ credence 1 that) a. Hence by Conditionalization, her future credence function should be

$$Cr^{t^+}(\cdot) = Cr^t(\cdot \mid a)$$

Plugging in any act counterfactual of the form $a \rightarrow s_k$, we conclude that

$$Cr^{t+}(a \rightarrow s_k) = Cr^t(a \rightarrow s_k \mid a)$$

. . . the right-hand side is just the current Egan credence on s_k in a.

The Collapse Lemma

The foregoing is an opinionated—albeit, I think, accurate—account of the state of play vis-a-vis the challenge Egan presents to classical CDT. But there is a serious dialectical puzzle facing that challenge, which becomes clear when we look more closely at the semantics of the counterfactual.

Gibbard and Harper appeal to just two principles governing the semantics of the counterfactual connective $\rightarrow$: Modus Ponens and the Conditional Excluded Middle ("CEM") (Stalnaker, 1968):

$$\text{(CEM) } (A \rightarrow S) \vee (A \rightarrow \neg S)$$

It is worth a quick aside to explain why CEM, in particular, is both *controversial from a truth-conditional perspective*—the original context of the Lewis-Stalnaker debate over the truth-conditions of counterfactuals—and *desirable from a probabilistic perspective*, such as Gibbard and Harper's.

First, the controversy. One basis for resistance to CEM is "indeterminate" pairs like

(1) If Bizet and Verdi had been compatriots, Bizet would have been Italian.
$B \rightarrow I$

(2) If Bizet and Verdi had been compatriots, Bizet would *not* have been Italian.
$B \rightarrow \neg I$

In light of there being two "equally good" ways for the antecedent to be true—one in which both composers are Italian and one in which both composers are French—some semantic accounts, including the prominent account of Lewis (1973), classify both (1) and (2) as (completely) false.[17] Hence both instantiated disjuncts of CEM are false, and CEM itself is no axiom.

But when we move to subjective probability—scoping $a \rightarrow s$ under a credence function $Cr^t(\cdot)$ for the purposes of calculating $\mathscr{U}(a)$—CEM contributes a nonnegotiable feature: namely, it secures that $Cr^t(a \rightarrow \cdot)$ is an additive probability function given that $Cr^t(\cdot)$ is. From this it follows that $Cr(a \rightarrow \neg s)$ goes up *in proportion* to $Cr(a \rightarrow s)$'s going down.

$$(**)\ Cr(a \rightarrow \neg s) = 1 - Cr(a \rightarrow s)$$

17 These Bizet-Verdi conditionals are based on a examples from Quine (1950), which are much-discussed in Lewis (1973).

This is incompatible with the CEM-rejecting attitude towards (1)-(2) above: the "both (completely) false" view is one on which the agent's credal views concerning *what would happen if she were to perform a* are *sub-additive*: an unacceptable violation of Probabilism.

We return, then, to accepting CEM—at least, for the sake of cashing out CDT, if not for the the sake of cashing out the semantics of natural language counterfactuals.[18] Together, Modus Ponens and CEM entail a principle which Gibbard and Harper call (Consequence 1), and take as the characterizing axiom of the counterfactual:

> ***Consequence 1***: $A \supset [(A \rightarrow B) \equiv B]$
>
> (Gibbard and Harper, 1978, 127-128).

We are now in a position to articulate the Collapse Lemma from which this paper takes its title. The lemma states that the Egan credence on *s* in *a* collapses into the conditional credence in *s* given *a*, with the result that for any decision problem and any option *a*, $\mathscr{U}_{\text{Egan}}(a) = V(a)$.

Proof. by the Ratio Formula,

$$Cr_{Egan}(a, s) = Cr(a \rightarrow s \mid a)$$

$$= Cr\Big(\big(a \rightarrow s\big) \wedge a\Big) / Cr\big(a\big)$$

By Consequence 1, applying the biconditional from right to left:

$$Cr\Big(\big(a \rightarrow s\big) \wedge a\Big) / Cr\big(a\big) = Cr\big(s \wedge a\big) / Cr\big(a\big)$$

$$= Cr(s \mid a).$$

QED.

Eganized Causal Decision Theory is thus *equivalent* to Evidential Decision Theory: on the leading model-theoretic implementation the $\rightarrow$ connective, future-directed CDT collapses into classical EDT, a theory that eschews representing causal relationships altogether.

Amongst the immediate dialectical consequences of this result is that it becomes obscure how Egan himself can get the result that one should pick the dominant act in Newcomb's Puzzle. Conceptually, Egan's argument makes it seem like there are three things: (i) the

18 Indeed, some authors use something like Equation ($\star$) to justify rejecting the "both false" intuition in the Bizet-Verdi cases. For example, Stefánsson argues that "the interaction between our confidence[s]" in the counterfactuals like (1)-(2) justifies a truth-conditional semantics that accepts CEM as an axiom (Stefánsson 2018, Sec. 3). For more work on CEM and probability judgments, see Moss (2013b); Eagle (2010); Williams (2012); Mandelkern (2019), and the influential proposals in Skyrms (1980) and Skyrms (1981).

subjective probability of a state, given an act; (ii) the subjective probability that if the act *were* performed, the state *would* result; and (iii) the subjective probability one would have in that same counterfactual, if one learned (only) that the act was actually performed. But it has transpired that (i) and (iii) cannot be distinguished. A causalist-friendly response to the Collapse Lemma, then, is to leverage it to argue that the appearance of there being three things, rather than two, is simply mistaken. The argument from future credence in counterfactuals was just a disguised version of the same reasoning Causalists rightly learned to reject in Newcomb problems. Moreover, the causal decision theorist can provide a complete model theory compatible with this view of counterfactuals.

But looked at another way, the Collapse Lemma is clearly a bizarre result for CDT, too. For the CDTer must confront, not just Egan's particular counterexamples,[19] but also his *argument*, which coherently leverages *both* causal notions and concepts related to learning. Given the lemma, these would appear to be on a collision course. The CDTer cannot rely on the Collapse Lemma to deprive the argument of force, since what the proof may *really* indicate is that imposing Consequence 1 on the semantics of counterfactuals issues in a flawed formulation of CDT. Note that dialectically, Egan himself has no reason to endorse Consequence 1. If it fails, and the reduction does not hold, the argument from Murder Lesion can be weakened *from* an argument in favor of $\mathscr{U}_{\text{Egan}}$ (and hence, given the Collapse, in favor of $\mathscr{V}$) to a mere argument *against* $\mathscr{U}$, on grounds that many two-boxers will be tempted to accept. This position—not the endorsement of Eganized CDT, but merely an argument against the classic version of CDT—is indeed Egan's considered view, though for different reasons than the one advanced here.[20]

In the next section, I look at Gibbard and Harper's own treatment of the Collapse Lemma. Readers interested in the connection between the puzzle I have framed for CDT in this section and Lewis's approach to the semantics of counterfactuals, which rejects CEM (Lewis, 1973, 1981), are referred to Fusco (2017); there, I prove that the simplest CDT-friendly way of jettisoning Consequence 1 in fact does little to alter the basic dialectic sketched above.

The G&H discussion of the Collapse

Gibbard and Harper derive the basic version of the Collapse Lemma in passing, en route to another point (op. cit., pg. 130). But they return to it later in the paper, with greater emphasis, in a paragraph I reproduce below (first underlined passage). They pair it, as Egan does, with an implicit endorsement of Conditionalization (second underlined passage):

19 There are at least two with *Murder Lesion*'s structure in the original paper. The other widely discussed one is called "Psychopath Button" (Egan op. cit., pg 97).

20 op. cit., pg. 111.

> When a person decides what to do, he has in effect learned what he will do, and so he has new information. He will adjust his probability ascriptions accordingly. These adjustments may affect the $\mathcal{U}$-utility of the various acts open to him.
>
> Indeed, once a person decides to perform an act a, the $\mathcal{U}$-utility of a will be equal to its $\mathcal{V}$-utility. Or at least this holds if Consequence 1 . . . is a logical truth. For we saw in the proof of Assertion 1 that if Consequence 1 is a logical truth, then for any pair of propositions P and Q, Prob(P → Q | P) = Prob(Q | P). Now let $\mathcal{U}_a$ be the $\mathcal{U}$-utility of an act *a* as reckoned by the agent after he has decided for sure to do *a*, and let Prob give the agent's probability ascriptions before he has decided what to do. Let Prob_a give the agent's probability ascriptions after he has decided for sure to do *a*. Then for any proposition P, Prob_a (P) = Prob (P | *a*). Thus $\mathcal{U}_a(a)$ . . . $\mathcal{V}(a)$ . . . the $\mathcal{V}$-utility of an act . . . is what its $\mathcal{U}$-utility would be if the agent knew he was going to perform it.

This passage is from pre-Egan times, however, and the result is not viewed by Gibbard and Harper as unsettling. They continue:

> It does not follow that once a person knows what he will do, $\mathcal{V}$-maximization and $\mathcal{U}$-maximization give the same prescriptions. For although for any act *a*, $\mathcal{U}_a(a) = \mathcal{V}(a)$, *it is not in general true that for alternatives b to a,* $\mathcal{U}_a(b) = \mathcal{V}(b)$ . . . the distinction between $\mathcal{U}$-maximization and $\mathcal{V}$-maximization remains. (op. cit., pg. 157)

Note here that in Gibbard and Harpers's notation, $\mathcal{U}_c(d)$ (for arbitrary acts *c* and *d*) is $\sum_i Cr(s_i \mid c)Val(s_i \wedge d)$; hence in their notation, $\mathcal{U}_a(a)$ what I have called $\mathcal{U}_{\text{Egan}}(a)$.

While I am a partisan of CDT, I am not sure whether, or how, the underlined observation can defend the view from Egan's conceptual argument. Both the quoted passages above begin with a person who has come to *know*, via making up her own mind, that she will do some $a \in \mathcal{A}$. But how did this person decide and thereby *come* to know that she was going to do *a*? At stake is the identity of *a*—whether it is the $\mathcal{V}$-maximizing act, or the $\mathcal{U}$-maximizing act, of the decision problem the agent faces. (For simplicity, assume the problem is an NC Problem, so that the act must maximize *one* quantity but not both.)

With this in mind, we can re-phrase the observation that $\mathcal{U}_a(a) = \mathcal{V}(a)$ as an Egan-friendly hindsight check. If *a* is the $\mathcal{V}$-maximizing act (call this "a^E"), then a CDT-like procedure will verify it in hindsight, since the observation that $\mathcal{U}_{a^E}(a^E) = \mathcal{V}(a^E)$ can be glossed as the observation that a^E maximizes $\mathcal{U}$ on the condition that it is decided on. This feature plausibly generates a foresight condition: the agent is in a position to know *prospectively* that *if* she were to choose a^E, it would pass the verification check—meeting with the approval of her "future epistemic self", as Jeffrey might say. If the agent instead chooses the $\mathcal{U}$-maximizing act (call it "a^C"), her act will, by the same token, *fail* the hindsight check. Returning to Egan's example, the person may find herself quite confident that *if she were to shoot, she would hit*. But if shooting itself provides evidence that this counterfactual is really false, and she knows

this as she deliberates about whether to shoot, she can anticipate that her high confidence in the counterfactual will be decimated by the evidential impact of taking the shot, reducing $\mathcal{U}(a^C)$ to the already low $\mathcal{V}(a^C)$ (viz., to $\mathcal{U}_{a^C}(a^C)$.)[21]

Hence the interpretive question, in looking at these Gibbard and Harper passages, is this: how does it help to emphasize, as Gibbard and Harper do in the third underlined passage, that "it is not in general true that for alternatives *b* to *a*, $\mathcal{U}_a(b) = \mathcal{V}(b)$"? To have a name for both what is granted and what is not, we can set out:

Fact 3. **(The Two Faces of Collapse)**. Although, in any NC Problem, for any act *a*:

(i) $\mathcal{U}_{\text{Egan}}(a) = \mathcal{U}_a(a) = \mathcal{V}(a)$,
the Causal Expected Utility of a if the agent conditions on a is equal to the prior Evidential Expected Utility of a;

(ii) it is not the case that for $a' \neq a$: $\mathcal{U}_a(a') = \mathcal{V}(a')$.
It is not the case that for other acts a', the Causal Expected Utility of a' if the agent conditionalizes on a is equal to the prior Evidential Expected Utility of a.

In the context of *Murder Lesion*, for example, (ii) is the point that while

$\mathcal{U}_{\text{shoot}}(\text{shoot}) = \mathcal{V}(\text{shoot})$
and
$\mathcal{U}_{\text{not shoot}}(\neg\text{shoot}) = \mathcal{V}(\neg\text{shoot})$
are true, the following inequality also typically holds:
$\mathcal{U}_{\text{shoot}}(\neg\text{shoot}) \neq \mathcal{V}(\neg\text{shoot})$

But it is not clear why this last fact is of interest.

I'll consider three alternative takes on the underlying dialectic, ending up with one that I favor. As advertised, the third consideration will avert to the peculiarity of predicaments involving (knowingly) choosing one's evidence.

---

21 The hindsight check I frame here has much in common with the spirit of *ratifiability* of acts (Jeffrey op. cit., pgs 15-16; see also Egan's discussion, pg. 107 ff.). However, it does different dialectical work from the original use of the notion, as can be seen by considering Newcomb's Problem again. Two-boxing is the only ratifiable act in Newcomb's Problem, even though one-boxing maximizes evidential utility from the point of view of Jeffrey-style EDT, because whether the agent comes to condition on 1*B* or on 2*B* (given Collapse, whether the agent becomes nearly certain that $[(1B \rightarrow P1) \wedge (2B \rightarrow P1)]$ or becomes nearly certain that $[(2B \rightarrow P2) \wedge (2B \rightarrow P2)]$), it maximizes evidential expected utility to do 2*B*. That dialectic does not apply just as stated to *Murder Lesion*, because it is not the case that whether Mary conditions on shooting *or* she conditions on not shooting, it maximizes her expected utility to shoot. Rather, *Murder Lesion*, like Gibbard and Harper's "Death in Damascus" case, is an example of nontrivial decision dependence (Hare and Hedden, 2015).

Take One: The Immediate Post-Act Perspective is Practically Unimportant

A first, simple thought is that even if there is no argument against "pre-conditioning" on one's own acts, this perspective is ultimately unimportant. This is because it is immediately "trumped" by an agent's learning something obviously much more important: viz., learning which $s_k \in \mathscr{S}$ is actual, and thus completely fixing her total payoff for the decision problem.

Most of the cases we have looked at suggest this. Recall again Mary, the agent in *Murder Lesion*. Egan's story emphasizes that as soon as Mary takes the shot, she will instantly have evidence that her shot is unlikely to hit Alfred. If Mary conditionalizes on her act *as* the bullet flies, she will lose hope in a good outcome. But this is obviously a fleeting moment: she is about to *see* whether the bullet hits Alfred. The same dynamic is at work in the classic version of Newcomb's Problem. The way that it is typically told, the choice of either two-boxing or one-boxing leaves little time for epistemic readjustment: the fact of the matter as to whether the big box contains a million dollars, or not, is instantly revealed when the choice is made.[22]

These intermediate moments—*between* the time when an act is chosen and the time when all is revealed—thus occupy an odd position; they loom large in the dialectic that motivates Eganized CDT over classical CDT, but in the context of the cases we've been asked to consider, they seem too ephemeral to be significant loci of epistemic or practical concern. When Mary acts, she is invested in the fate of what we might call her *posterior* future self—that is, the fate of the person who either lives out her days under tyranny, liberates the nation from Alfred's grip, or is jailed by his cronies. She is *not* directly invested in the features, epistemic or otherwise, of her *proximal* future self—the one who witnesses some temporary fluctuation in attitude towards the utility of her presently available acts.

Alas, this attempt to deflate Eganized causal decision theory—by *practically*, if not exactly *epistemically*, undermining the immediate post-act perspective—is unsuccessful. For the fleetingness of the post-act, pre-revelation period is just an artifact of particular cases. By stretching this period out, allowing plenty of time for the registration of one's regrets, the critical period can be rendered epistemically significant. In a variation on *Murder Lesion*, for example, we can imagine that Mary's option is to lob a javelin at Alfred from an enormous distance. This choice would allow her plenty of time to reflect on her chances of success before "all is revealed".

Moreover, and more potentially embarrassing, a lengthened critical period can be rendered *practically* important, by creating a context in which the agent can *act* on her regrets. If an agent like Mary will predictably regret her choice after she has acted, an enterprising third party can swoop in and offer her—for a fee, of course—a *further* act which will partially offset the consequences of her previous choice in the state she now believes to be most

22 But see Seidenfeld *op. cit.*, pgs. 204-205.

likely. Mary's behavior overall will reflect the self-defeating character of someone who fails to account for what she expected to learn upon acting.

This operationalization represents the logical next step in the evolution of Egan's counterexamples. I know of four similar vignettes in the literature, two due to Ahmed (Ahmed 2014, Ch. 7.4.3; Ahmed, 2017), one due to Meacham (Meacham, 2010, pg. 64-65) and one due to myself (Fusco, 2018, Sec. 3). I provide a simple, *Murder Lesion*-friendly version here, referring the reader to this literature for a more thorough discussion of the way the dialectic interacts with the literature on sequential choice:[23]

> *Murder Lesion, Snailmail Edition*. Mary is deliberating about whether to try to assassinate Alfred by mailing him a bomb. Mary has good evidence that a certain kind of brain lesion, which she may or may not have, causes murderous tendencies but also causes would-be assassins to have significant dyslexia in the writing down of the addresses of their intended victims. This dyslexia is bad enough that if Mary has it, her package is unlikely to reach Alfred, and likely to be delivered to an innocent stranger living somewhere else instead. Mary is currently fairly confident that she does not have mailing-address dyslexia, and not very confident that she will put a bomb in the mail.

We construct a decision matrix that duplicates counterfactual state-descriptions and the payoff relations in the original *Murder Lesion*:

	Mail → address incorrect	**Mail → address correct**
Mail	–1000	+1000
¬ Mail	0	0

Suppose that, as a classic CDTer, Mary goes ahead and mails the package. She is now quite sure she has the lesion, and thus quite sure her act will have a terrible outcome. I represent this by omitting the ¬ Mail option and bolding in the state Mary now assigns the most probability mass to:

	Mail → address incorrect	**Mail → address correct**
Mail	**–1000**	+1000

An entrepreneur now offers Mary a deal: she can pay $100 for him to intercept her package. Assuming that interception brings about the same results as never having mailed the

23 See esp. Ahmed (2014, pg. 204) and Ahmed (2017, Sec. 3).

package at all, taking the deal maximizes causal expected utility from the perspective Mary now occupies:[24]

	Mail → address incorrect	**Mail → address correct**
Mail, no Deal	–1000	+1000
Mail, Deal	0–100	0–100

Looked at as a whole, though—especially if Mary knew, in advance, that someone would offer her the deal if she mailed the package—this course of action is bizarre. Mary has paid $100 to secure an outcome she could much more easily have guaranteed for free: a life under monotonous tyranny (value: $0).

I conclude that emphasizing the unimportance of the immediate post-act perspective is not a fruitful way to respond to the Collapse's challenge to CDT.

Take Two: The Factivity of a

A second possibility for understanding the inequality in Fact 3 is that Gibbard and Harper are drawing attention to the factivity of knowledge. Because one cannot know what is false—the thought goes—a deliberating *rational* agent could not genuinely have access *both* to the epistemic position of someone who *has learned* she will do a^E and the epistemic position of someone who *has learned* that she will do a^C. After all, one of these acts is irrational. Hence one quantity, either $\mathcal{U}_{a^E}(a^E)$ or $\mathcal{U}_{a^C}(a^C)$, is not really a rationally accessible causal expected utility: for at least one act $a \in \{a^E, a^C\}$, $\mathcal{U}_{\text{Egan}}(a)$ $(= \mathcal{U}_a(a))$ corresponds to the causal expected utility a rational agent would assign to a if she learned something she cannot possibly learn—namely, that she is going to do it. Returning to Jeffrey's maxim to chose for the person you expect to be when you have chosen, the argument could be put like this: one of the two epistemic perspectives afforded by the decision problem $Cr^t(\cdot \mid a_E)$ or $Cr^t(\cdot \mid a_C)$ is simply *not* a perspective a rational agent has *any* chance of occupying, once she has chosen.

While this way of cashing out the passage is coherent, it is unsatisfactory as it stands, for two reasons. First, it is a quite general fact about decision-making that agents use conditional probabilities which reflect views on what happens conditional on propositions they regard themselves as having no chance of learning. Suppose that I am faced with an option a that brings some risk of death, and I am spitefully contemplating whether my acquaintances

24 Calculation: suppose for concreteness that initially, Mary's confidence $Cr^t(M \rightarrow \neg C) = .1$ and $Cr^t(M \rightarrow C) = .9$. Then $\mathcal{U}(M) = -1000(.1) + 1000(.9) = 800$. However, M is good evidence for $M \rightarrow \neg C$: $Cr^t(M \rightarrow \neg C \mid M) =$ (by Collapse) $= Cr^t(\neg C \mid M) = .75$. It follows (since $Cr^t(\cdot)$ is a probability function) that $Cr^t(C \mid M) = .25$. Hence the second offer ("D") will be calculated at:

$\mathcal{U}(D) = .75(-100) + .25(-100) = -100$
$\mathcal{U}(\neg D) = .75(-1000) + .25(1000) = -500$
Hence $\mathcal{U}(D) > \mathcal{U}(\neg D)$.

will be remorseful in the event that I die. If this makes a difference to my utilities, then the expected utility of *a* will depend in part of *Cr*(remorse | die). This is true even if I regard it as impossible for me learn that I've died.[25]

Second, it isn't clear in general how this dialectical maneuver generalizes to the waxing and waning of credence. In a credal, rather than a full-belief, context, a defender of Egan CDT can accept the letter of factivity while avoiding much of its spirit. On at least *many* views of the latitude we have in deliberation, each epistemic position—knowing that one is going to do a^C, and knowing that one is going to do a^E—can be *approximated*, even if it cannot be fully accepted, by a rational agent in the throes of deliberation.[26] A rational agent who Jeffrey Conditionalizes on a surging resolve to perform a^C, for example, can come arbitrarily *close* to the epistemic position she will occupy if she does a^C.

Indeed, we could simply re-define Egan causal expected utility by appeal to this kind of approximation. The current definition pegs $\mathcal{U}_{\text{Egan}}$ to the $\mathcal{U}$ an agent will assign to *a* once she has conditionalized on *a*, which, on a reasonable construal of Conditionalization, is justified only if she comes to *know a*. But Egan could instead have defined it like this:

$$\mathcal{U}_{\text{Egan}}(a) = \lim_{Cr^t(a)\to 1} \mathcal{U}(a)$$

$$= \lim_{Cr^t(a)\to 1} \sum_i Cr(a \to s_i) Val(a \wedge s_i)$$

On this definition, $\mathcal{U}_{\text{Egan}}(a)$ takes the limit of the classic causal expected utility of *a* as the agent Jeffrey Conditionalizes on increasing confidence that she will perform *a*. In all of the example cases that are widely discussed in the literature,[27] this definition yields the same, classical CDT-unfriendly verdicts as the original version of Eganized CDT.

Take Three: Eganized Credences are Fake News

I proposed above that one way of reading the Gibbard and Harper response to the Collapse is as an argument that one of $Cr_{a^E}(\cdot)$ and $Cr_{a^C}(\cdot)$ represents an illegitimate epistemic perspective for a rational agent. One take on interpreting "legitimacy" had to do with the what can be learned by agents who are rational, and thus cannot make irrational choices. But that interpretation is unpersuasive.

---

25 This point can also be made with reference to exercises like Jeffrey's *Death Before Dishonor* (Jeffrey, 1983, pg. 89) and, in the semantics literature, by Richmond Thomason's "cheating spouse" examples, discussed by van Fraassen (1980).

26 On one family of such views, one can "try on" the epistemic perspective of someone who performs an act *a* by *supposing* that one will do *a* (Joyce, 2007, Sec. 3; Velleman, 1989). This justifies an agent in provisionally increasing his confidence in *a*.

27 Including *Murder Lesion*, Egan's *Psychopath Button* (op. cit., pg 97), Gibbard and Harper's *Death in Damascus*, Richter's asymmetric Death in Damascus variant (Richter, 1984, pg. 396).

Another way of interpreting epistemic legitimacy—the interpretation I will argue for in the rest of this paper—has to do with evidential quality. This interpretation grants that both a_E and a_C are, in the relevant sense, accessible to the decisionmaker; however, it emphasizes that one of these acts is misleading in respect of which proposition in $\{s, \neg s\}$ is true. The important "causalist" observation is that while, at the moment of decision, the agent can *choose*, by acting, whether her total future evidence will support s or $\neg s$, she cannot choose which of s or $\neg s$ is actually true.

I've thus far assumed the following about an agent, like Mary, facing an NC problem or *Murder Lesion*-like problem (where Gibbard and Harper's rule recommends a^C and EDT/Eganized CDT recommends a^E):

(a) The agent can, in the relevant sense, perform both available acts. Hence both are potentially learnable for her: she can learn a^C (if she does a^C) and she can learn a^E (if she does a^E).
(b) She will conditionalize on her chosen act.
(c) Sequential decision problems which exploit the post-act perspective no matter what she does are possible if she does a^C. (Example: the longform version of *Murder Lesion, Snailmail Edition*.)

However, I have not granted the normative upgrade of (b):

(d) The agent *ought*, epistemically, to conditionalize on her chosen act.

Indeed, I suggest that (d) is not generally true. So long as we assume (b), then, a CDTer has an argument to the effect that an agent who conditionalizes on a^C ends up in a no better (and potentially worse) epistemic position than she occupied before acting.

Before going on to sketch the argument, it is worth being explicit about how, if successful, it affects the dialectic. We seek an account of why causal expected utilities are better calculated according to classical CDT, rather than Eganized CDT. Egan's argument depends on the idea that in decision situations, we *ought* to defer to our future credences. But *if* the CDTer has reason to think her future degrees of belief in $\{s, \neg s\}$ are no better than her current ones, she can reject this call to deference.

To make the case, it will be useful to adopt a time-slice perspective (Hedden, 2015), which conceives of a single persisting agent as an aggregation of different agents at different times. Following Moss (2012), one can take a further step into that perspective by conceiving of what are ordinarily glossed as diachronic epistemic norms, like Conditionalization, as norms of communication—more particularly, of *testimony and knowledge-transmission*—between the different time-slices that constitute the agent.

Within this structuring metaphor, here are a few platitudes. In general, given that a messenger is statistically reliable with regard to signals $\{e_1, \ldots, e_n\}$ indicating $\{h_1, \ldots, h_n\}$,

I should take the messenger's signaling that e_i as evidence that h_i is true. However, this moral must be applied with caution in cases where *I am* the messenger. When I am contemplating different signals $\{e_1, \ldots, e_n\}$ I could send into the ether, I should *not* generally hold that *my* signaling e_i is evidence for *my future self* that h_i is true. This caveat holds even if I have a past track record of being highly reliable on h-related matters.

I take it that the reason for this is no great mystery. My past track record of being a reliable messenger is underwritten by adherence to alethic norms: I generally tried to send only the signal $e_j \in \{e_1, \ldots, e_n\}$ which I antecedently held to be likely on *my own* evidence. In a diachronic context, this means I generally tried to send e_i only when my prior probability for e_i on my total evidence was high. Given my current choice of signals, if I send a signal in e_j which total evidence does not support, I have knowingly flaunted the mechanism which was responsible for my past reliability; I thereby gain no posterior reason to take my signal as novel evidence that h_j is true.

These platitudes can be fruitfully applied in the context a classic NC problem, in which a^C is a signal that is a statistically reliable indicator of $\neg s$. The agent faces a choice between a^C, which immediately secures her a "sweetener" of Δ, and a^E, which foregoes it. The agent can either send her future self good news, or send her future self bad news—"goodness" and "badness" here being understood with respect to whether her act statistically indicates that s is true or false. The sweetener is available just in case the agent sends herself bad news. In only one case, though, will her act constitute a signal to her future self that is non-*misleading* according to her current total evidence: and that is just in case she performs the act which accords statistically with the state-hypothesis (s or $\neg s$) that her *prior* supports.

	s	$\neg s$
a^E	Θ	0
a^C	$\Theta + \Delta$	Δ

The agent located at t, therefore, faces a tradeoff between prudential and (what we might call) testimonial goods. It is a *testimonial* good to refrain from sending misleading evidence to one's future self. (After all, this evidence may be called upon later in future utility-maximization problems.[28]) But it is a *prudential* good to pick up sweeteners while one can. Without a more detailed description of the tradeoff—one which, for example, connects future epistemic states to future utility-maximization problems at particular, specified stakes, or states that the agent directly values reliable testimony for its own sake—there is no clear answer to the question of what all-things-considered rationality requires in such a

28 For a classic example of a tradeoff between aiming to maximize expected utility in one's current decision problem and aiming to better one's epistemic position in view of anticipated *future* decision problems, see the "exploration"-"exploitation" tradeoff in Multi-armed Bandit Problems (Robbins, 1952; Berry and Fristedt, 1985).

case.[29] All that can be said is that opting for the sweetener is what achieves the *immediate* goal of maximizing utility. And while this *may* be permissible as far as all-things-considered rationality is concerned, it gives the agent no reason to regard her future credences in act counterfactuals—those influenced by suspect testimony—as better than her current ones. Without that presumption, Egan's call to deference to her future credences is blocked.[30]

Conclusion

To sum up: we canvassed the collapse lemma itself, and how it is derived from the Ratio Formula and from commitments regarding the semantics of the counterfactual which are endorsed by Gibbard and Harper's classic account of CDT. I also provided an opinionated account of the lemma's relationship to Egan's counterexample to CDT, as well as Gibbard and Harper's response to the lemma in their 1976 paper.

In closing, it is fruitful to briefly compare the Collapse to Lewis's "Bombshell"—that is, his triviality results for the indicative conditional. In "Probabilities of Conditionals and Conditional Probabilities" (1976), Lewis proved the bizarre result that if $Cr(B \mid A) = Cr(A > B)$, then $Cr(A > B) = Cr(B)$. Putting the two equalities together and framing the result diachronically, this means conditioning on A does nothing to one's posterior credence in (arbitrary) B. But this is obviously absurd: if learning A at t does anything at all, it changes posterior credences. The weak link in this road to paradox is apparently the commitment about the semantics of the indicative conditional.

For comparison, The Gibbard-Harper collapse result states that, in an NC problem, conditioning $\mathrm{Cr}(\cdot)$ on an act a leads to the distinction between causation and correlation's being obliterated in hindsight. Framed diachronically, this means that when agents look backwards, they are insensitive to the difference between knowing they caused s and knowing

---

29 A similar moral has been emphasized by careful commentators on diachronic Dutch books and other arguments for Conditionalization:

> The claim is not that dynamic coherence and reflection are sufficient for all-things-considered rationality. The claim is that dynamic incoherence and violations of reflection are indicators of epistemic irrationality . . . it is perfectly rational (in the all-things-considered sense) to prefer a situation in which one is slightly epistemically irrational to a situation in which one is perfectly epistemically rational but has to pay all sorts of nonepistemic costs. (Huttegger, 2013, pg. 423; emphasis in original)

30 Relevant here is Nissan-Rozen (2017), who argues that, in NC problems, the agent's high (even degree-1) credence $Cr(\neg s \mid a^c)$ is *Gettiered* and hence fails to be probabilistic knowledge in the sense of Moss (2013a). Furthermore, Nissan-Rozen claims that *the agent is in a position to know this* about the relevant high conditional credence (*op. cit.*, pg 4813). Whether Nissan-Rozen's view aligns with the diachronic suggestion floated here depends, however, on the question of how agents ought to update in such cases (for example, in the degree-1 case, whether an agent should use Modus Ponens on a conditional she believes, when she *also* knows that her conviction in that conditional is Gettiered.) For more about updating for Causalists, see also Cantwell (2010).

they merely sent a signal that indicates *s* without causing it. I think this is as absurd as Triviality: if the causation-correlation distinction does anything at all, it does something which must be capable of being appreciated in hindsight as well as in foresight. Once again, the weak link seems to be a commitment about the semantics of a type of natural language conditional—this time, the counterfactual conditional, rather than the indicative one. But there are many ways for a causal decision theorist to tackle this puzzle, and I have only gestured at one. Confronting this outstanding issue in a model theory of credence is a frontier for formal developments of CDT.

References

Ahmed, Arif (2013). "Causal Decision Theory: A Counterexample." Philosophical Review, 122: pp. 289–306.

——— (2014). Evidence, Decision and Causality. Cambridge University Press.

——— (2017). "Exploiting Causal Decision Theory." Draft, University of Cambridge.

Berry, D. A., and B. Fristedt (1985). *Bandit Problems: Sequential Allocation of Experiments*. London: Chapman and Hall.

Cantwell, John (2010). "On an Alleged Counter-Example to Causal Decision Theory." *Synthese*, 173: 127–52.

Eagle, Anthony (2010). "'Might' counterfactuals." Retrieved through the author's website.

Egan, Andy (2007). "Some Counterexamples to Causal Decision Theory." *Philosophical Review* 116, no. 1: 93–114.

Fusco, Melissa (2017). "An Inconvenient Proof: The Gibbard-Harper Collapse Lemma for Counterfactual Decision Theory." In *Proceedings of the 21st Amsterdam Colloquium*, edited by A. Cremers, T. van Gessen, and F. Roelofsen, 265–75.

——— (2018). "Epistemic Time-Bias in Newcomb's Problem." *Newcomb's Problem*, edited by A. Ahmed. Cambridge: Cambridge University Press.

Gibbard, Allan, and William Harper (1978). "Counterfactuals and Two Kinds of Expected Utility." In *Foundations and Applications of Decision Theory, Vol. 1*, edited by C. A. Hooker, J. J. Leach, and E. F. McClennen, Dordrecht: D. Reidel.

Hare, Caspar, and Brian Hedden(2015). "Self-Reinforcing and Self-Frustrating Decisions." Noûs, 50, no. 3: 604–28.

Hedden, Brian (2015). "Time-Slice Rationality." Mind 124, no. 494: 449–91.23.

Huttegger, Simon (2013). "In Defense of Reflection." *Philosophy of Science* 80 no. 3: 413–33.

Jeffrey, Richard (1983). *The Logic of Decision*. Chicago: University of Chicago Press.

Joyce, James (1999). *The Foundations of Causal Decision Theory*. Cambridge: Cambridge University Press.

——— (2007). "Are Newcomb Problems Really Decisions?" *Synthese* 156: 537–62.

Kment, Boris (2019). "Decision, Causality, and Pre-Determination." Draft, Princeton University.

Lewis, David (1973). *Counterfactuals*. Oxford: Blackwell.

——— (1981). "Causal Decision Theory." *Australasian Journal of Philosophy* 59, no. 1: 5–30.

Mandelkern, Matthew (2019). "Talking about Worlds." *Philosophical Perspectives*.

Meacham, Christopher (2010). "Binding and Its consequences." Philosophical Studies 149: 49–71.

Moss, Sarah (2012). "Updating as Communication." *Philosophy and Phenomenological Research* 85, no. 2: 225–8.

——— (2013a). "Epistemology Formalized." *Philosophical Review* 122, no. 1: 1–43.

——— (2013b). "Subjunctive Credences and Semantic Humility." *Philosophy and Phenomenological Research* 87, no. 2: 251–78.

——— (2018). *Probabilistic Knowledge*. Oxford: Oxford University Press.

Nissan-Rozen, Ittay (2017). "Newcomb meets Gettier." *Synthese* 194: 4479– 814.

Nozick, Robert (1969). "Newcomb's Problem and Two Principles of Choice." In *Essays in Honor of Carl G Hempel*, edited by N. Rescher. Dordrecht: Springer.

Quine, W. V. O. (1950). *Methods of Logic*. New York: Holt, Reinhart and Winston.

Richter, Reed (1984). "Rationality Revisited." *Australasian Journal of Philosophy* 62, no. 4: 392–403.

Robbins, H. E. (1952). "Some Aspects of the Sequential Design of Experiments." *Bulletin of the American Mathematical Society* 527–35.

Savage, Leonard (1972). *The Foundations of Statistics*. New York: Dover.

Skyrms, Brian (1980). Causal Necessity: A Pragmatic Investigation of the Necessity of Laws. New Haven, CT: Yale University Press.

——— (1981). "The Prior Propensity Account of Subjunctive Conditionals." In *Ifs: Conditionals, Belief, Decision, Chance, and Time*, edited by W. Harper, R. Stalnaker, and G. Pearce. Dordrecht: D. Reidel.

Spohn, Wolfgang (1986). "The Representation of Popper Measures." *Topoi*, 5.

Stalnaker, Robert (1968). "A Theory of Conditionals." In *Ifs: Conditionals, Belief, Decision, Chance, and Time*. Dordrecht: D. Reidel.

Stefánsson, H. Orri (2018). "Counterfactual Skepticism and Multidimensional Semantics." *Erkenntnis* 83, no. 5: 875–98.

van Fraassen, Bas (1980). "Review of Rational Belief Systems by Brian Ellis." *Canadian Journal of Philosophy* 10: 497–511.

Velleman, David (1989). "Epistemic Freedom." *Pacific Philosophical Quarterly* 70: 73–97.

Williams, J.R.G. (2012). "Counterfactual Triviality: A Lewis-Impossibility Argument for Counterfactuals." *Philosophy and Phenomenological Research* LXXXV, no. 3: 648–70.

3

WHO'S AFRAID OF NORMATIVE EXTERNALISM?

Zoë Johnson King

1. Moral Uncertainty

This paper is about what someone should do when she is not only unsure what first-order moral theory is true but also unsure about whether this moral uncertainty is itself morally relevant.

I myself am not sure what first-order moral theory is true. Of course, I have some sense of which things matter morally. But I don't take myself to have figured out precisely how many and which things matter morally, the exact relative degrees to which each of these things matters morally, and the metaphysical relationships (causal or constitutive) that obtain between them. The true first-order moral theory, fully spelled out, will settle all these matters. So, there are a great many possible moral theories that are consistent with my rough sense of which things matter morally. I am unsure which of these theories is true.

This paper addresses the question of how someone like me—who is not sure what first-order moral theory is true—should act. The question has been a hotbed of philosophical activity in recent years. One central insight that emerges from the recent literature is that we can use the tools of decision theory to discuss and generate possible answers to this question. We can do this because we can construe my moral uncertainty as a kind of uncertainty about what the world is like: it is uncertainty about what the world is like *morally*. With this construal in hand, we can model my uncertainty using the familiar table that we all know and love from undergraduate courses on decision theory. Simplified drastically, it looks something like this:

	T1 (0.6)	T2 (0.2)	T3 (0.2)	Expected Value
A1	10	2	4	7.2
A2	4	8	6	5.2
A3	2	6	50	12.4

Here the columns represent each moral theory in which the agent has some non-zero credence. (I have some non-zero credence in far more than three moral theories, but, for ease of exposition, I'll just discuss an imaginary possible agent who divides her credence exhaustively between exactly three theories.) The rows represent acts that she might perform. And each box contains a number representing the moral value of performing the corresponding act according to the corresponding theory. For example, the table shows that theory 1 says that performing act 1 would have 10 units of moral value.

The phrase "units of moral value" might raise a skeptical eyebrow. It is difficult to explain what this can mean in a way that remains neutral between first-order moral theories. And it is even more difficult to devise a way to render different moral theories' assessments of acts commensurable such that they can all be placed on a single scale—which is necessary for the numbers in the table to even make sense. These are major problems that have been discussed at length in the literature on moral uncertainty.[1] I do not think that these problems can be easily solved. But I will set them aside for the purposes of this paper, since I am interested in an entirely different problem. For present purposes, I will proceed as if we have found a clear way to make intertheoretic comparisons of moral value.

Lots of different possible decision rules dictate how someone should act given the kind of moral uncertainty depicted in the table above. These decision rules for moral uncertainty correspond to rules in traditional decision theory saying how an agent should act given uncertainty about mundane empirical matters, like whether it will rain or whether an egg that she is considering adding to her omelet is rotten.[2] The most widely discussed such rule is this:

> Maximize Expected Objective Value (MAXEOV): Morally uncertain agents should *maximize expected objective value*—i.e., perform the act with highest expected objective moral value.

MAXEOV is explicitly inspired by the traditional decision theorist's refrain that an uncertain agent ought to maximize expected value. According to MAXEOV, a *morally* uncertain agent ought to maximize expected *moral* value. This is accomplished by calculating the expected moral value of each act—a weighted sum of the amounts of moral value that it has according

1 Some important contributions to this literature are Gracely (1996), Lockhart (2000), Ross (2006), Sepielli (2009, 2013a, 2013b), Hedden (2016), and Hicks (forthcoming). N.B. It might be possible to solve the problem of intertheoretic value comparisons by "consequentializing" all moral theories (on which see, e.g., Dreier 2011; Hurley 2013; Portmore 2007). But I suspect that even this would not enable us to place the theories' evaluations on a single scale.

2 For the omelet example, and a classic statement of traditional decision theory, see Savage (1954).

to each moral theory in which the agent has some non-zero credence, weighted by the agent's credence in each of the corresponding theories—and then performing the act with the highest total expected moral value. In the case represented in the table above, this rule says that the uncertain agent should perform A3. Defenses of MAXEOV appear in Lockhart (2000), Sepielli (2009), and Enoch (2014), among others.

Myriad other possible decision rules for moral uncertainty would mimic other principles from traditional decision theory. For example, there are possible decision rules stating that morally uncertain agents should use maximin, or maximax, or minimax regret, or that they should maximize risk-weighted expected moral value in the style of Lara Buchak (2013), and so forth. Thankfully, these rules do not yet all have advocates in the literature on moral uncertainty.

Another widely discussed rule is this:

> My Favorite Theory (MFT): A morally uncertain agent should just consult the moral theory in which she has the highest credence, and should perform the act that is ranked best by this theory, ignoring all the others.

MFT holds that a morally uncertain agent should perform the act recommended by the moral theory in which she has the highest credence. In the case represented in the table above, this rule says that the agent should perform A1. This approach is attractive because it completely avoids having to compare the degrees of moral value that acts have according to different moral theories—a clear advantage of MFT over MAXEOV, and over all the other decision-theoretic rules just mentioned. Nonetheless, MFT is not very popular, because it generates some highly counterintuitive results. Most notably, if an agent has a plurality of credence in a theory that ranks A1 the highest, but all of the other theories in which she has some positive credence hold that A1 would be a serious moral disaster, MFT cheerfully recommends A1. This holds even if there is a "safe" alternative act, A2, which all theories agree would be very good. That seems like a bad result.[3] Still, MFT has received some sophisticated recent defenses, most notably from Johann Gustafsson and Olle Torpman (2014).

One last decision rule will play a starring role in this paper. Here it is:

> Do The Right Thing (DTRT): Morally uncertain agents should do whatever is in fact the right thing to do, regardless of their credences.

---

3 One could render this result less implausible by adding a restriction to MFT that requires uncertain agents to follow the dictates of their favorite theory only if their credence in this theory is above 0.5—this would avoid recommending actions that the agent thinks are more likely than not to be a serious moral disaster. But this version of MFT still entails that someone should perform A1 if she has credence 0.5000001 in a theory that holds that A1 is slightly better than a safe alternative, and credence 0.4999999 in a theory that holds that A1 would be a serious moral disaster. This verdict still seems unacceptably morally risky to many people.

DTRT is very different from the other decision rules. According to DTRT, an agent's moral uncertainty is morally irrelevant. What a morally uncertain agent should do is the same as what someone who is certain of the true first-order moral theory should do, which is also the same as what someone who is certain of a false moral theory should do: she should do whatever is in fact required by the theory that is in fact the true first-order moral theory. DTRT has been defended by Brian Weatherson (2013) and Elizabeth Harman (2015). It follows from the approach termed "normative externalism" and discussed by Weatherson in his (2019), according to which the true moral norms apply to agents irrespective of our epistemic states with respect to them, and thus are in an important sense "external" to us. On this view, an agent faced with the kind of moral uncertainty depicted in the table above should ignore the table entirely and do whatever is in fact morally right—however risky or unreasonable this may seem to her. Moreover, on this view, debates between "internalist" views like MAXEOV and MFT about how agents' credences in moral theories make a difference to what she should do are totally wrongheaded. They are like debates about what blood type R2D2 is. R2D2 is a robot, so he doesn't have any blood. Thus, an inquiry into R2D2's blood type would be a total waste of time. Similarly, according to DTRT, agents' credences in moral theories make no difference to what they should do. Thus, an inquiry into what kind of difference they make is a total waste of time.

I am inclined to think that something like MAXEOV must be correct. But, for a long time, I was troubled by the possibility that normative externalism might be true. This is partly because, as just discussed, if this position is true then the debates about moral uncertainty that seem to me interesting and important are in fact a total waste of time. But that is not all. I was also troubled by the fact that, while I am inclined to think that something like MAXEOV must be correct, I certainly do not assign credence 1 to its correctness. The other decision rules just mentioned all have arguments in their favor that are good enough for me to assign *some* positive credence to each of them being correct. So, in addition to being uncertain about what first-order moral theory is true, I am in a state of *higher-order uncertainty*: uncertainty about what a morally uncertain agent (such as myself) should do. As an internalist, I am inclined to think that I should in some way take account of each of the higher-order theories in which I have some positive credence when I decide what to do. However, it turns out to be quite difficult to know how to take account of the possibility that normative externalism is true—and that I am wasting my time, because the correct decision rule for morally uncertain agents is just DTRT—within a MAXEOV-like framework. That is what this paper is about.

2. Higher-Order Uncertainty

As just discussed, I am not sure which theory about how people should act when they are morally uncertain is true. But here is one thing of which I am fairly certain: I am fairly certain that the term "should" in the phrase "how people should act when they are morally

uncertain" is a *moral* "should." I find it hard to see what other "should" it could be. It does not, for instance, seem to be prudential, epistemic, or aesthetic, given that the arguments for and against MAXEOV and its rivals trade on moral intuitions rather than intuitions pertaining to any of these other normative domains. Nor does it seem plausible that the "should" in question is an all-things-considered "should," akin to the Gibbardian primitive "ought" of practical deliberation (2003, especially 152-158). This primitive "ought" is the one in the final "ought I to φ?" question that an agent asks herself before acting; it is the one such that it is an incoherent combination of attitudes to judge that one ought to φ and yet not intend to φ. On this point, here is Gibbard (2003, 153):

> To think something the thing to do is to plan to do it. To think, for instance, that the thing now to do is to defy the bully who torments me is to plan to defy him. And planning right now to defy him right now, to do it at this very moment, amounts to setting out to do it. My theory thus yields internalism in a strong form[4]: if I think that something is now the thing to do, then I do it. My hypothesis about ordinary *ought* judgments is that they are judgments of what to do, of what is the thing to do. I don't, then, think that I ought right now to defy the bully unless I do defy him. If I fail to defy him, then as a matter of the very concept of *ought*, I don't believe I ought to.

This does not seem to be the case for the "should" in claims of the form "the true principle governing how people should act when they are morally uncertain entails that, in light of my moral uncertainty, I should φ." One need not plan right now to φ in order to count as accepting a claim of this form. On the contrary, it seems perfectly intelligible to make such a claim and then nevertheless ask, "But shall I φ?" (And it seems perfectly intelligible to answer "no, I shan't," if there are weighty prudential or epistemic or aesthetic considerations counting against φ-ing.) The fact that such questions always remain open tells us that it is at least not a *conceptual* truth that the verdicts of principles governing morally uncertain agents always settle the practical question of what to do. But the primitive "ought" does settle this practical question, and it does so, as Gibbard says, "as a matter of the very concept of *ought*" (2003). So, the "should" in principles governing morally uncertain agents is not akin to the Gibbardian primitive "ought." The "should" here is no primitive "should." Given that the arguments for and against MAXEOV and its rivals trade on moral intuitions, then, I take the "should" in such principles to be a *moral* "should."

---

4 N.B. This is not the sense of "internalism" that contrasts with the term "externalism" as used by Weatherson. Gibbard is a *motivational* internalist: he holds that one does not count as making a normative judgment (in the quotation above, a judgment about what I ought to do) unless one is motivated to act accordingly. Weatherson is a *normative* externalist: he holds that the most important norms governing thought and action are external in the sense that they apply to us regardless of our epistemic states with respect to them. Despite the names, these views are about totally different things.

This means that the decision rules surveyed above are all partially self-referential. They apply to morally uncertain agents, and they are themselves (putative) moral principles. So, each one applies, in part, to uncertainty about whether it is indeed the correct moral principle governing morally uncertain agents.

I am therefore in a state of *higher-order moral uncertainty*. Not only am I uncertain about which first-order moral theory is true, but I am also uncertain about what I (morally) should do in light of this fact.

We can draw up a decision-theoretic table representing the higher-order moral uncertainty of people like me. It would look something like this:

	MAXEOV (0.95)	**MFT (0.025)**	**DTRT (0.025)**	**Expected Value**
A1	7.2	10	?	?
A2	5.2	4	?	?
A3	12.4	2	?	?

Here, again, the columns represent each (higher-order) moral theory in which the agent has some credence, the rows represent her available acts, and each box contains a number representing the moral value of performing the corresponding act according to the corresponding theory. (Again, for ease of exposition, I will discuss an imaginary agent who divides her credence exhaustively between exactly three higher-order theories, though I myself have some non-zero credence in more theories than just these.)

Some of the columns in this table are easier to fill out than others. It is easy to fill out the MAXEOV column; the moral value of each act according to MAXEOV is just the expected objective value of the act. This is easy to calculate using one's first-order decision table. It is also easy to fill out MFT's column; the value of each act according to MFT is just the value that it has according to the first-order moral theory in which one has the highest credence, which is easy to see from one's first-order decision table. But it is impossible for a morally uncertain agent (such as myself) to fill out the DTRT column of her higher-order decision table. I can describe the values that this column should contain in opaque terms: it should contain the moral value of each act according to the *true* first-order moral theory. But I do not know what the true first-order moral theory is—therein lies my moral uncertainty. Since I do not know what the true first-order moral theory is, I do not know how much moral value each of my available acts has according to the true theory. So, I am not in a position to fill out the DTRT column of my higher-order decision table with actual numbers.[5]

5 This paragraph overstates the difference between MAXEOV and DTRT a bit. Just as I do not know what the true first-order moral theory is, I do not know *precisely* what credence I have in various first-order moral theories, since my credences are not completely luminous to me. So, technically, the numbers in the MAXEOV column should be narrow ranges of the sort that I discuss when describing the upper-and-lower-bounds strategy later in this section. Nonetheless, in the vast majority of cases, the ranges will be sufficiently narrow to generate a clear ordering over my available acts. And, if I do know precisely what my first-order

These observations troubled me for a long time. I was troubled because I thought that they highlighted a problem for MAXEOV. Like all theories about how morally uncertain agents should act, MAXEOV is partly self-referential—it applies, in part, to uncertainty about whether it is indeed the correct principle governing morally uncertain agents. But MAXEOV simply does not work when applied to the higher-order decision tables of agents, like me, who have some positive credence in normative externalism. I cannot fill out the DTRT column of my higher-order decision table, as I am unsure what first-order moral theory is true. But this means that I cannot apply a maximizing algorithm. I cannot calculate weighted sums of the values in each row of my decision table if one box in each row contains a question mark instead of a number. The algorithm crashes; it needs precise numbers in each box as input. My credence in DTRT thus has a devastating impact on the higher-order application of MAXEOV. Indeed, the problem is not just that I am no longer able to identify a choice set among my available acts. It's worse than that; I am unable to rank the acts at all, nor even to assess the higher-order expected objective value of any one of them.

This does not mean that MAXEOV cannot handle *any* uncertainty as to whether it is the correct principle governing morally uncertain agents. On the contrary, it is possible to calculate the higher-order expected objective value of one's available acts if the higher-order theories in which one has some credence are all internalist. If all higher-order theories in which I have some positive credence take the value of my available acts to depend in some way on the contents of my first-order decision table, then I am okay. For example, an agent who divided her credence entirely between MAXEOV and MFT would have no trouble applying MAXEOV to her higher-order decision table, since these theories both assign precise numbers representing the value of the agent's available acts in light of her first-order uncertainty. It is only when someone has some positive credence in an externalist principle, like DTRT, that she cannot apply MAXEOV to her higher-order uncertainty.

Nonetheless, I found this troubling. I think it is plausible that the true theory about how morally uncertain agents should act should apply to both first-order and higher-order uncertainty. This would offer an attractively unified picture of normative reality. Certainly, if one theory were true for first-order moral uncertainty and a totally different theory were true for higher-order uncertainty, that would be a striking fact that calls out for explanation.[6] And I cannot think of a satisfying explanation for why MAXEOV should be the true theory

credences are, then I can complete the MAXEOV column precisely. By contrast, in *all* cases, uncertain agents will be unable to fill out the DTRT column of their decision table with actual numbers, and—as we will see below—the ranges will frequently be too broad to generate a clear ordering over my available acts. (The solution that I propose in §3 could be used to accommodate uncertainty about the values of acts according to MAXEOV and MFT, as well as DTRT.) Thanks to Billy Dunaway for pointing all of this out to me.

6 Here I deviate from Sepielli (2013b), who suggests that each agent simply resolve her uncertainty at level N by appeal to whichever theory she is most confident of at level $N+1$, if there is one. I am not so laissez-faire.

for first-order uncertainty but not for higher-order uncertainty—certainly, "the math doesn't work" does not seem like a satisfying explanation. But I also think it is plausible that the true theory about how morally uncertain agents should act should be able to withstand agents' credence in alternative theories. After all, the point of these theories is to guide the uncertain. It would therefore be ironic if such a theory were unable to identify a choice set, nor to rank the available acts, nor even to assess the value of any individual act, for all uncertain agents who have some positive credence in a particular alternative theory. And this seems to be the case here.

One possibility for the friend of MAXEOV is to dismiss DTRT out of hand. Perhaps we should leave agents' credence in DTRT out of their higher-order decision tables, or revise our maximizing algorithm to ignore it (renormalizing the agents' credence in other theories accordingly). It has been suggested to me that friends of MAXEOV may defend this move by way of an analogy between DTRT and moral nihilism. Some philosophers hold that, in a first-order decision table, it is permissible to ignore the possibility that moral nihilism is true and that nothing has any moral value whatsoever. After all (so the argument goes), decision theorists typically ignore states of the world in which all acts result in the same outcome, since these outcomes "cancel out" when we assess the expected utility of each act. But nihilism is the possibility that nothing has any moral value. So, it is a possibility according to which all acts result in the same outcome, namely no moral value. Thus, it can be safely ignored.[7] The defender of MAXEOV might try suggesting that DTRT is relevantly analogous to moral nihilism. Nihilists answer the question, "Which first-order moral theory is true?," by saying, "None of them." And, similarly, normative externalists answer the question, "Which decision-theoretic principle governs morally uncertain agents?," by saying, "None of them." So perhaps, like first-order moral nihilism, normative externalism can be safely ignored.[8]

I think that this is the wrong way to think about DTRT. The possibility that Normative Externalism is true is not the possibility that all acts have the same moral value. Rather, it is the possibility that the first-order moral uncertainty of agents like me is morally irrelevant, and that what we should do is determined by the true moral theory alone, rather than by our credences in moral theories. So, as long as the true moral theory holds that different acts have different degrees of moral value, DTRT holds this too. This is therefore not a principle according to which all acts are equally good. It is rather a principle such that agents who are first-order uncertain cannot tell which acts it evaluates as better than others, or by how much.

I think it would be strange for different theories to be true at different levels. So, if someone takes this to be the case, I hope that she has a satisfying explanation of why the levels are different.

7 This suggestion comes from Ross (2006) and is criticized in MacAskill (2013). The idea that a state of the world in which all acts yield the same outcome can be ignored is the idea that motivates the Allais paradox; see Allais (1953).

8 Thanks to Andrew Sepielli for raising this possibility to me in conversation.

Relatedly, it is a mistake to see DTRT as a decision rule according to which there is no answer to the question of what someone should do in light of her moral uncertainty. On the contrary, DTRT issues verdicts as to what someone should do in light of her moral uncertainty—but they are the same verdicts as those that the rule issues for agents who have credence 1 or credence 0 in the true first-order moral theory. To put the same point another way: on this view, as on MAXEOV or MFT, there is a function from an agent's epistemic state and the moral facts to a choice set. But the striking thing about DTRT is that this function's output stays the same as the agent's credences in moral propositions change, so long the moral facts (and morally relevant nonmoral facts) are held fixed.[9] The trouble with DTRT is not that it fails to evaluate or compare the available acts of first-order uncertain agents; it is just that first-order uncertain agents cannot tell what these evaluations and comparisons are.

So, simply dismissing normative externalism out of hand seems unmotivated. Another option is for me to make an educated guess as to what goes in the DTRT column of my higher-order decision table. I can use my first-order decision table to think about what the values in the "DTRT" column of the higher-order decision table might, for all I know, turn out to be. Granted, I cannot say confidently what the values in this column are, since I am unsure what first-order moral theory is true. But I do have a partition of possible first-order theories between which I divide my credence. And each of these theories assigns a precise degree of objective moral value to each of my available acts. (Recall that we are assuming, for present purposes, that there is a way to make intertheoretic comparisons of value such that a first-order decision table can intelligibly be completed.) So, I can place *upper and lower bounds* on the degree of moral value that each of my available acts might turn out to have. The upper bound is the highest degree of objective value assigned to the act by a moral theory in which I have some positive credence, and the lower bound is the lowest such degree of value. Although I am not sure precisely what degree of moral value the act has, I am sure that it falls somewhere within this range.[10]

Given the first-order decision table presented at the beginning of §1, this strategy will yield a higher-order decision table that looks like this:

9 Elizabeth Harman (2015) makes this point especially clearly. Defending a version of normative externalism that she calls "Actualism," Harman repeatedly emphasizes that she offers Actualism as a theory about what uncertain agents *subjectively* ought to do (58). As she puts it, "Actualism is a proposed answer to the very same question the [defenders of MAXEOV] are interested in, namely: how should a person act, taking into account her beliefs and credences (including her moral beliefs and credences), given that one sometimes must act while experiencing moral uncertainty?" (70). Since this is Harman's own characterization of her view, I think it is a fair characterization of the view.

10 Thanks to Brian Weatherson for raising this possibility to me in conversation.

	MAXEOV (0.95)	MFT (0.025)	DTRT (0.025)	Expected Value
A1	7.2	10	[2–10]	[7.14–7.34]
A2	5.2	4	[4–8]	[5.14–5.24]
A3	12.4	2	[2–50]	[11.88–13.08]

This is progress. With a range within which the value of each act according to DTRT must fall, it is possible to identify a range within which the act's higher-order expected objective value according to MAXEOV must fall. This is accomplished by calculating two weighted sums, one that takes the value of the act for DTRT to be the lowest value in the range and one that takes it to be the highest in the range. Since the agent is sure that the value of the act according to DTRT falls somewhere within the range, she can establish that a higher-order application of MAXEOV would yield a degree of higher-order expected objective value somewhere within the bounds specified in the final column. Moreover, in some cases—like the one above—this provides enough information to rank some of the acts, since some acts' ranges of possible degrees of expected objective value are wholly above or below some of the others. In the table above, the ranges are sufficient to totally order the acts (A3 > A1 > A2), and to identify a singleton choice set ({A3}).

However, this ordering is *only* possible when one act's range of possible degrees of expected objective value is wholly above another's. And this will not always be the case. Indeed, it can very easily fail to hold. For example, if the agent's first-order decision table is exactly as presented at the beginning of §1 except that the moral value of act 3 according to theory 3 is 25 rather than 50, then her higher-order decision table will be as follows:

	MAXEOV (0.95)	MFT (0.025)	DTRT (0.025)	Expected Value
A1	7.2	10	[2–10]	[7.14–7.34]
A2	5.2	4	[4–8]	[5.14–5.24]
A3	7.4	2	[2–25]	[7.13–7.705]

Now it is no longer possible to rank acts A1 and A3, since A1's range of potential expected objective value is wholly contained within A3's. There is no answer to the question, "Which is greater, *between 7.14 and 7.34* or *between 7.13 and 7.705*?"—the question is ill-formed. So, the uncertain agent cannot order these actions, nor can she identify a choice set from among them. Of course, the defender of MAXEOV could propose a further rule for what to do in such cases: the rule might be maximax, or maximin, or a rule telling us to choose the act whose range of potential expected objective value has the highest midpoint, or whatever. But the very fact that such a wide range of rules for this situation are available makes the choice of any particular one seem arbitrary and unmotivated. So, in the absence of a positive argument for any particular rule, I suggest that we try a different tack.

3. A Solution

In this section, I will present what I think is the best strategy for defenders of MAXEOV to take account of the credence that agents like me have in DTRT.

The strategy begins with the same observation as the upper-and-lower-bounds strategy above: an uncertain agent like me has a first-order decision table displaying a partition of possible first-order theories between which she divides her credence, each of which assigns a precise degree of objective value to each available act. As we have just seen, this means that we can identify upper and lower bounds to what might, for all the agent knows, be the true moral value of each act—and thus its value according to DTRT. My preferred strategy begins with the observation that we can do even better. For the agent also assigns a precise credence to each of these first-order theories' being true. She is still unsure which of them is true, and thus which assigns the values to acts that DTRT endorses. But, for each theory, she has some precise credence that this theory specifies the values of her acts according to DTRT. After all, DTRT is the decision rule of normative externalism, which instructs uncertain agents to do whatever is in fact required of them by the true first-order moral theory. So, the agent's credence that T1 assigns the values to acts that DTRT endorses is just her credence that T1 is the true first-order moral theory. The same holds for all other first-order moral theories in which she has some non-zero credence; her credence that DTRT evaluates and compares acts in accordance with the dictates of this theory should just be her credence that this theory is true.

With the foregoing in mind, here is my strategy. The uncertain agent should think of the possibility that normative externalism is true, not as one possibility, but as many possibilities—as many as there are first-order theories to which she assigns some positive credence. There is no way for normative externalism to be true without some particular first-order theory being true. And the agent has at her disposal a partition of first-order moral theories that she thinks might be true. So, she should divide the credence assigned to DTRT in her higher-order decision table between these possibilities. That is to say: rather than including one single "DTRT" column in the higher-order table, she should include as many columns as there are first-order moral theories to which she assigns some positive credence, with these columns representing the combined possibilities that DTRT is the correct rule for uncertain agents *and* T1 is the true first-order moral theory, that DTRT is the correct rule for uncertain agents *and* T2 is the true first-order moral theory, and so on. Her credence in each combined possibility can be easily calculated; it is the product of her credence in DTRT and her credence in the first-order theory.[11]

11 This assumes that DTRT is probabilistically independent of each first-order theory in which the agent has some credence. I take this to be a reasonable assumption, since the existing arguments for and against normative externalism are all supposed to hold independently of which first-order theory is true. Nonetheless, I confess to being unsure as to whether it really is probable that DTRT is probabilistically independent of all first-order theories. For instance, maybe DTRT is more plausible if objective consequentialism is the

The result will be a higher-order decision table that looks something like this:

	MAXEOV (0.95)	MFT (0.025)	DTRT and T1 (0.015)	DTRT and T2 (0.005)	DTRT and T3 (0.005)	Expected Value
A1	7.2	10	10	2	4	7.27
A2	5.2	4	4	8	6	5.17
A3	12.4	2	2	6	50	11.88

Call this strategy "the repartitioning strategy."

As is clear from this table, the repartitioning strategy has some merits. In contrast to the upper-and-lower-bounds strategy, the repartitioning strategy will always enable the uncertain agent to complete her higher-order decision table with precise numbers in every box. This ensures that it is always possible to apply a maximizing algorithm to the higher-order decision table, notwithstanding the agent's credence in DTRT. Moreover, the agent's credence in DTRT is not simply being set aside or ignored, as in the strategy that treats normative externalism like moral nihilism. On the repartitioning strategy, the agent's credence in DTRT is fully taken into account by being repartitioned into her credences in the combined possibilities that serve as inputs to its maximizing algorithm. Moreover, the repartitioning strategy enables uncertain agents to calculate the higher-order expected objective value of each available act in a manner that always yields a total ordering (in the example above, A3 > A1 > A2) and a choice set (in this case, {A3}), so long as their lower-order decision tables are sufficiently detailed to yield such things. The repartitioning strategy thus solves pretty much all of MAXEOV's problems. So, I think that this is the best strategy for defenders of MAXEOV to take account of the credence that agents like me have in DTRT.

However, the repartitioning strategy has a surprising implication. We can begin to grasp this by considering what would happen if the agent whose uncertainty is depicted in the table above were to become somewhat less confident in MAXEOV and somewhat more confident in DTRT. Suppose, for instance, that she shifts from credence 0.95 in MAXEOV and credence 0.025 in DTRT to credence 0.775 in MAXEOV and credence 0.2 in DTRT (retaining credence 0.025 in MFT). Her higher-order decision table would change as follows:

	MAXEOV (0.775)	MFT (0.025)	DTRT and T1 (0.12)	DTRT and T2 (0.04)	DTRT and T3 (0.04)	Expected Value
A1	7.2	10	10	2	4	7.27
A2	5.2	4	4	8	6	5.17
A3	12.4	2	2	6	50	11.88

true first-order theory than if subjective consequentialism is the true first-order theory, just because this would be a less weird combination of lower-order and higher-order facts about the normative relevance of an agent's credences.

Now suppose that she shifts to credence 0.525 in MAXEOV and credence 0.45 in DTRT. Her higher-order decision table would then look like this:

	MAXEOV (0.525)	MFT (0.025)	DTRT and T1 (0.27)	DTRT and T2 (0.09)	DTRT and T3 (0.09)	Expected Value
A1	7.2	10	10	2	4	7.27
A2	5.2	4	4	8	6	5.17
A3	12.4	2	2	6	50	11.88

Finally, suppose that she assigns credence 0 to MAXEOV and credence 0.975 to DTRT. Her higher-order decision table would then look like this:

	MFT (0.025)	DTRT and T1 (0.585)	DTRT and T2 (0.195)	DTRT and T3 (0.195)	Expected Value
A1	10	10	2	4	7.27
A2	4	4	8	6	5.17
A3	2	2	6	50	11.88

The general point is as follows. On the repartitioning strategy, an uncertain agent can shift any amount of credence back and forth between MAXEOV and DTRT without this making any difference to her choice set, to the ordering of her available acts, or even to the precise higher-order expected objective value of any act. This holds in full generality; no matter which first-order theories the agent has some credence in, and no matter which other higher-order theories (besides MAXEOV and DTRT) she has some credence in, it will remain the case that the agent can shift an arbitrary amount of credence from MAXEOV to DTRT and back again without anything in the "Total" column of her higher-order decision table changing at all.

This is not just some wild coincidence. It happens because the part of the higher-order calculation whereby MAXEOV takes account of my credence in DTRT, on the repartitioning strategy, is much like a simple first-order calculation of expected objective value. I take each first-order theory in which I have some credence, take the objective value of the act according to that theory, and then multiply this number by the product of my credence in the first-order theory and my credence in DTRT. I do this for each first-order theory in which I have some credence. I then sum the results. This part of my calculation of the higher-order expected objective value of the act is mathematically equivalent to simply taking its first-order expected objective value and multiplying it by my credence in DTRT. To verify that this is so, compare these equations:

(i) $(a * m * q) + (b * n * q) + (c * o * q)$
(ii) $((a * m) + (b * n) + (c * o)) * q$

Here the variables *a*, *b*, and *c* stand for my credences in T1, T2, and T3, while *m*, *n*, and *o* stand for the value of an act according to each theory (respectively), and *q* stands for my credence in DTRT. The first equation is the part of my higher-order calculation that takes account of my credence in DTRT, on the repartitioning strategy. The second is a first-order calculation of expected objective value, multiplied by my credence in DTRT. These equations are straightforwardly equivalent. So, on the repartitioning strategy, the part of my higher-order calculation that takes account of my credence in DTRT is equivalent to simply taking the act's first-order expected objective value and multiplying it by my credence in DTRT. Now, consider the part of my higher-order calculation that takes account of my credence in MAXEOV. This part involves taking the act's first-order expected objective value—the value of the act according to MAXEOV—and multiplying it by my credence in MAXEOV. It is no wonder, then, that the uncertain agent employing a repartitioning strategy can shift any amount of credence back and forth between MAXEOV and DTRT without the higher-order expected objective value of any act changing at all. Whatever credence she assigns to MAXEOV and to DTRT, she will then effectively just multiply each of these credences by the first-order expected objective value of the act and sum the results. Shifting credence back and forth between MAXEOV and DTRT will not change this sum.

This is a curious result. The repartitioning strategy ensures that, in all cases, my credence in DTRT may as well be credence in MAXEOV. So it turns out that, after all, I *can* safely ignore the possibility that normative externalism is true—just as the analogy with nihilism suggested that I do. But I am no longer ignoring this possibility based on a misunderstanding of normative externalism, as the analogy with nihilism did. On the contrary, I am now ignoring this possibility on principled grounds. I am ignoring it because the most promising way to accommodate it in my calculations of expected objective value implies that my credence in this theory may as well just be credence in MAXEOV. So, I can acknowledge that normative externalism might be true, without this acknowledgment ever making any difference to what I should do, nor to the ordering over my available acts, nor even to the higher-order expected objective value of any act. In short, on this strategy, my acknowledging that normative externalism might be true doesn't change anything.

This might make the repartitioning strategy seem unfair to normative externalism. Why accommodate this theoretical possibility in a way that ensures that it never makes a difference to what I should do? But that is too quick. The points above all apply only to my shifting credence between DTRT *and MAXEOV*. If there are other higher-order theories in which I have some credence—such as MFT—then shifting some credence from *these theories* to DTRT does change the higher-order expected objective value of my available acts. And this can make a difference to what I should do, depending on how much credence shifts and on what my first-order decision table looks like. So, my credence in DTRT does matter, sort of.

I say "sort of" because shifting a certain amount of credence from another theory to DTRT will always have the same impact on an act's higher-order expected objective value as shifting the same amount of credence from this other theory to MAXEOV. It remains the

case that my credence in DTRT may as well be credence in MAXEOV. But my credence in DTRT does matter, sort of, because it matters how much credence I assign to DTRT *rather than MFT*, and rather than any other higher-order theory, although it does not matter how much credence I assign to DTRT rather than MAXEOV.

There is another way in which the repartitioning strategy genuinely is unfair to normative externalism. It is unfair because normative externalism holds that uncertain agents should perform the act required by what is in fact the true first-order moral theory, even if they have *no* credence in this theory. The DTRT decision rule evaluates my available acts according to the true first-order moral theory, whether or not this theory appears in one of the columns of my first-order decision table. So, if I were to fully acknowledge the possibility that normative externalism is true, I would have to take account of the possibility that the true moral theory is one to which I assign credence zero. This is not a possibility that the repartitioning strategy is able to accommodate.

But that does not seem such a terrible result. After all, we are considering the possibility that the true moral theory is one to which I currently assign credence zero. By stipulation, my credence in this possibility must be zero. And possibilities to which I assign credence zero are not usually ones that I am required to take into consideration when deciding what to do. For example, the possibility that my typing this sentence will summon an evil square circle that destroys the universe does not lead me to hesitate in typing this sentence, notwithstanding the fact that it would be very bad if the universe was destroyed, as I assign this possibility credence zero. Within a decision-theoretic approach to thinking about moral uncertainty, there is no way to accommodate hypotheses in which I have no credence. But this is simply part and parcel of the decision-theoretic approach to thinking about how uncertain agents should act. So, while the repartitioning strategy cannot capture everything involved in the possibility that normative externalism is true, it seems to be the best that MAXEOV can do.

This last point should be appealing to those attracted to a Gibbardian view of decision-making as a matter of deciding what to do in light of one's plan-laden beliefs. (To be clear, I think the strategy I have described is the best strategy for defenders of MAXEOV regardless of how they feel about Gibbard. But the fact that it fits well with Gibbard's view is, to my mind, a nice bonus.) On a Gibbardian view, the possibility that normative externalism is true is simply not the sort of possibility that we can plan for. We plan what to do in possible subjective states that we might be in, where these are states that result from conditionalizing a set of initial credences on a set of total evidence. So, we do not plan for certain sets of objective facts to hold—such as the fact that a certain first-order moral theory is true—but for us to acquire evidence indicating that there is a certain probability that they hold. Here is Gibbard (2003, 57):

> A plan must be one the agent can carry out with the information at her disposal. "Buy low, sell high," for example, is not a plan, if one has no way of telling whether prices have reached their

> peaks or their troughs. An *occasion*, as I have characterized it, contains much that the agent has no way of knowing, but one's plans must respond to features of the occasion available to the agent. Alternatives must be subjectively characterized, so that the same alternatives are available on subjectively equivalent occasions. And a plan must permit the same alternatives on subjectively equivalent occasions.

If Gibbard is right that a plan must permit the same alternatives on subjectively equivalent occasions, then a morally uncertain agent cannot plan to perform different acts depending on what first-order moral theory is in fact true, irrespective of her subjective states. And, if a plan must respond to features of the occasion available to the agent, then plans cannot respond to the objective moral facts. The best we can do as planners is to respond to our credences in various hypotheses as to what the objective moral facts might be, having conditionalized appropriately on our total evidence. Thus, the repartitioning strategy—or something very much like it—is the best that Gibbardian planners can do to take account of the possibility that normative externalism is true. We cannot be expected to respond to the facts (if they are facts) that externalism and some first-order moral theory are true. We can respond only to our credences in the combined hypotheses that they are true. Again, then, the repartitioning strategy seems to be the best that we can do.

If my argument in this paper is correct, it means good news for expected-objective-value-maximizers: we simply do not have to worry about the possibility that normative externalism is true. We can respond to arguments for externalism by listening patiently and then going about our days as if the arguments had never happened. And we can do this safe in the knowledge that, to the extent that the arguments we have heard have moved some of our credence in the direction of DTRT, the best way to incorporate this into our decision-making renders it mathematically equivalent to simply becoming even more confident of the truth of MAXEOV. Thus, maximizers can focus on their real enemies: other internalist decision rules proposing rival ways for morally uncertain agents to take account of their credences in first-order moral theories. Proponents of MAXEOV can pretty much just ignore DTRT, safe in the knowledge that, by doing so, they are acknowledging its possible truth as best they can.

4. Conclusion

Who's afraid of Normative Externalism?

Not me!

References

Allais, Maurice (1953). "Le comportement de l'homme rationnel devant le risque: critique des postulats et axiomes de l'école Américaine." *Econometrica* 21: 503–46.

Buchak, Lara (2013). *Risk and Rationality*. Oxford: Oxford University Press.

Dreier, James (2011). "In Defense of Consequentializing." In *Oxford Studies in Normative Ethics, Volume 1*, edited by Mark Timmons, 97-118. Oxford: Oxford University Press.

Enoch, David (2014). "A Defense of Moral Deference." *Journal of Philosophy* 111: 229–58.

Gibbard, Allan (2003). *Thinking How to Live*. Cambridge, Mass.: Harvard University Press.

Gracely, E. J. (1996). "On the Noncomparability of Judgments Made by Different Ethical Theories." *Metaphilosophy* 27, no. 3: 327–32.

Gustafsson, Johann, and Olle Torpman (2014). "In Defence of My Favorite Theory." *Pacific Philosophical Quarterly* 95: 159–74.

Harman, Elizabeth (2015). "The Irrelevance of Moral Uncertainty." In *Oxford Studies in Metaethics, Volume 10*, edited by Russ Shafer-Landau, 53–79. Oxford: Oxford University Press.

Hedden, Brian (2016) "Does MITE Make Right? Decision-Making under Normative Uncertainty." In *Oxford Studies in Metaethics, Volume 11*, edited by Russ Shafer-Landau, 102–28. Oxford: Oxford University Press.

Hicks, Amelia (forthcoming). "Moral Uncertainty and Value Comparison." In *Oxford Studies in Metaethics, Volume 13*, edited by Russ Shafer-Landau, 161–83. Oxford: Oxford University Press.

Hurley, Paul (2013). "Consequentializing and Deontologizing: Clogging the Consequentialist Vacuum." In *Oxford Studies in Normative Ethics, Volume 3*, edited by Mark Timmons, 123–53. Oxford: Oxford University Press.

Lockhart, Ted. (2000). *Moral Uncertainty and Its Consequences*. Oxford: Oxford University Press.

MacAskill, William (2013). "The Infectiousness of Nihilism." *Ethics* 123: 508–20.

Portmore, Douglas W. (2007). "Consequentializing Moral Theories." *Pacific Philosophical Quarterly* 88: 39–73.

Ross, Jacob. (2006). "Rejecting Ethical Deflationism." *Ethics* 116: 742–68.

Savage, Leonard (1954). *The Foundations of Statistics*. New York: John Wiley and Sons.

Sepielli, Andrew (2009). "What to Do When You Don't Know What to Do." In *Oxford Studies in Metaethics, Volume 4*, edited by Russ Shafer-Landau, 5–28. Oxford: Oxford University Press.

——— (2013a). "Moral Uncertainty and the Principle of Equity among Moral Theories." *Philosophy and Phenomenological Research* 86: 580–89.

——— (2013b). "What to Do When You Don't Know What to Do When You Don't Know What to Do . . ." *Noûs* 48: 521–44.

Weatherson, Brian (2013). "Running Risks Morally." *Philosophical Studies* 167: 1–23.

——— (2019). *Normative Externalism*. Oxford: Oxford University Press.

4

WHAT EPISTEMIC REASONS ARE FOR:

Against the Belief–Sandwich Distinction

Daniel J. Singer and Sara Aronowitz[1]

The standard view says that epistemic normativity is normativity of belief. If you're an evidentialist, for example, you'll think that all epistemic reasons are reasons to believe what your evidence supports. Here we present a line of argument that pushes back against this standard view. If the argument is right, there are epistemic reasons for things other than belief. The argument starts with evidentialist commitments and proceeds by a series of cases, each containing a reason. As the cases progress, the reasons change from counting in favor of things like having a belief to things like performing ordinary actions. We argue that each of those reasons is epistemic. If the argument succeeds, we should think there are epistemic reasons to consider hypotheses, conduct thought and physical experiments, extend one's evidence, and perform mundane tasks like eating a sandwich, just as there are epistemic reasons to believe what one's evidence supports.

1 Both authors contributed equally to this work. We'd like to thank all of those who helped us work through these ideas, including Daniel Singer's Spring 2019 graduate seminar on epistemic normativity at the University of Pennsylvania, Sandy Goldberg, David Plunkett, Xingming Hu, Amy Sepinwall, Josh Peterson, and Billy Dunaway. Also thanks to Richard McGehee and Silvio Levy for kindly providing the sphere eversion graphic, one of many wonderful images from their work with the University of Minnesota's *Geometry Center*.

1. "Belief, though, can't aim literally; It's we who aim"

At the beginning of "Rational Credence and the Value of Truth," Gibbard (2007, 143) says "Belief aims at truth—or so it is said, . . . Belief, though, can't aim literally; it's we who aim." Gibbard then argues that to explain that feature of epistemic rationality, we must see practical guidance value as playing an important role in what makes beliefs epistemically rational. Here we extend Gibbard's idea that what's epistemically required of us can be understood in terms of what's involved in an agent aiming for truth.

According to the standard view, epistemic normativity only governs belief. Put in terms of reasons, the claim is that all epistemic reasons are reasons to believe:

E↔B
Something is an epistemic reason if and only if it is a reason (of the right kind)[2] to believe.

E↔B is incompatible with there being epistemic reasons to consider hypotheses, draw helpful diagrams, request books related to one's research from the library, or look at birds in the sky to learn about birds.[3] But, insomuch as we see epistemic normativity as the kind of normativity that governs the project of promoting truth in our beliefs, we should think there can be epistemic reasons to do those kinds of things, since each of them can be an important part of believing the truth. Here we'll show that there is an intuitively compelling line of argument for thinking exactly that. We'll present a series of cases, each of which contains a reason. As the series progresses, what the reasons count in favor of becomes progressively more like doing an action than having a belief. Despite that, we'll argue that we should see each reason as an epistemic reason. If our argument succeeds, it will follow that just as we can have epistemic reasons to believe what our evidence supports, we can have epistemic reasons to eat a sandwich.

There is already a large literature on E↔B, but the focus of that literature is primarily on whether there are practical reasons for belief, that is, whether the right to left direction of E↔B is true (see, e.g., Reisner 2009; Markovits 2014; Rinard 2015; Leary 2017). We focus

2 One version of this view takes practical reasons to be the "wrong kind of reason" to believe and, in turn, treats epistemic reasons as the only "right kind of reason" to believe (e.g., Hieronymi 2005, 2013), whereas another version takes putative practical reasons to believe to be actually practical reasons to bring belief about (e.g., Kelly 2002, 171; Parfit 2001; Shah 2006, 498).

3 The standard view is a background assumption in a lot of work in epistemology from the last century. The view is argued for (or at least explicitly assumed) by Berker (2018) who argues that "epistemic and practical normativity are individuated by their objects of assessment: the normative assessment of belief (and other doxastic attitudes) is epistemic normativity, and the normative assessment of action is practical normativity" and by Turri (2009) and Raz (2009) who treat the standard view as an uncontroversial assumption. For more examples, see Sylvan (2016), who reviews theories of epistemic reasons where the assumption is that these are reasons for doxastic attitudes (e.g., one view of epistemic reasons is "the fact that p is a reason to have doxastic attitude D towards q").

on the opposite direction, that is, whether there are epistemic reasons for things other than belief. Our claims will be independent of whether there might be practical reasons to believe.

We take it that our argument would uninterestingly beg the question if it started by assuming a truth-loving consequentialist epistemology like the one endorsed by Goldman (1999) or Singer (2018). If there are epistemic reasons to do anything that results in maximizing the number of true beliefs we have, it would immediately follow that there are epistemic reasons to consider certain hypotheses, do certain calculations, set up scientific and educational institutions in certain ways, and drink coffees before certain soporific talks (assuming all of those things promote having more true beliefs).[4]

Instead, we'll take as our starting point a traditional picture of epistemic normativity that might lead someone to accept E↔B. According to *evidentialism*, epistemic normativity is about having the beliefs that are supported by one's evidence. On this view, all and only pieces of evidence are epistemic reasons. And because what evidence does is stand in support relations to beliefs, proponents of this view might conclude that all epistemic reasons are reasons to believe. Often, evidentialists, like Conee and Feldman (1985, 2000, 2004), make this conclusion explicit. We'll assume that our reader will grant the evidentialist starting intuition, that is, epistemic reasons spring from evidence and that we have epistemic reasons to believe what our evidence supports—though evidence may be understood in an internalist or externalist manner. What our argument will show is that someone who accepts those claims, along with a few other plausible claims about how reasons work, should think there are epistemic reasons for things other than belief.

Below we'll put the argument in terms of reasons, but we hope that it's clear that the force of the argument is neutral on the metanormative question of whether reasons, rationality, "ought"s, or something else is normatively fundamental. What we hope to show, put in more ecumenical terms, is epistemic normativity is not normativity of belief. Epistemic norms apply to mental and physical "doings" just as they do to beliefs.

Finally, we won't consider objections to our arguments below that appeal to worries about whether we can have reasons for things we cannot control. The cases we consider below are designed to be at most as controversial in this respect as paradigmatic cases of belief formation. The literature on doxastic volunteerism is well-trodden with arguments on all sides of the issue. Our starting point assumes that there are some epistemic reasons (reasons to believe what one's evidence supports), and the reader may choose their preferred way to handle objections to that on the basis of whether we control our beliefs.

With those clarifications in mind, we turn to the first step in the argument.

---

4 For the same reasons, we won't start with a similar picture offered by Kornblith (1983) where epistemically valuable actions and epistemically valid reasoning are seen as reflecting two different normative perspectives. We'll be concerned with a single notion of epistemic normativity that governs both belief and action.

2. Epistemic Reasons to Not Believe

The first step in the argument should be uncontroversial. Evidentialism says that we have epistemic reasons to *believe* what our evidence supports. A ham-fisted reading of that entails that all epistemic reasons must be reasons for beliefs, not other doxastic attitudes. But it seems like we can also have reasons to suspend belief:

> Marathon with Broken Watch
>
> You're running a marathon. Based on using your watch during training, you thought that it would take a little over 4 hours to finish. But, you just learned that the watch was broken the whole time. It incorrectly reported your training times, and you're not sure about the magnitude or direction of the errors.

In this case, once you learn the watch was broken, your evidence no longer supports believing anything about whether you'll finish in four hours, since the fact that the watch was broken defeats any reason you otherwise had to believe anything else. So it looks like you ought to suspend belief about whether you'll finish in four hours. Assuming one ought to have a doxastic state only if one has a reason to have it, there must be epistemic reasons to suspend belief.[5]

Of course, it's easy for our opponent to accept this by modifying their view to say that we have epistemic reasons to have whatever doxastic states are supported by our evidence, where suspension of belief is seen as a kind of doxastic state. The next case puts pressure on this refined view.

3. Epistemic Reasons to Consider

Consider this case:

> Competing Scientific Theories
>
> You're a chemist, and you have a long-standing dispute over a theoretical postulate with your colleague Tirrell. You both agree that the results of a titration experiment will shine light on your dispute. After repeating the titration multiple times, you discover that, in fact, the volume of the titrant is 50 mL, which is a result neither of you expected. You and Tirrell are both puzzled by the outcome. You're discussing the result with Tirrell in the department lounge

5 What should the evidentialist think this reason is? It's natural to think that the fact that the watch is broken is itself that reason, but there are many other options. The reason might instead be a higher-order reason arising from the lack of first-order reasons to believe, or it might be a standing reason that one always has to suspend when one doesn't have sufficient reason to have a certain belief. What's important though is that there must be some epistemic reason to *not* believe or to suspend belief.

when your colleague Jorrie, who was not previously a party to the conversation, interjects with what appears to be an elegant and plausible explanation of the result.

In this case, we'll argue that you have an epistemic reason to consider Jorrie's proposed explanation.[6] Of course, the claim isn't that you're obliged to accept Jorrie's proposal, since the proposal might be implausible on other grounds. It might even be the case that *all (epistemic) things considered*, you shouldn't consider the proposal; perhaps Jorrie is widely regarded as a crank who ought not be listened to. Nonetheless, the fact that Jorrie proposed a simple, elegant, and plausible explanation of the otherwise puzzling result is *a reason* to consider it, we'll argue.

The defender of the standard view will likely think that any reason to consider is practical and not epistemic. One reason to think it's practical is that, in scientific institutions, financial and professional incentives are often structured so as to promote epistemically good outcomes. So it's no surprise that you'd have a practical reason that's doing intuitively "epistemic" work.

We don't think the reason to consider Jorrie's hypothesis can plausibly be construed as merely practical though. To see why, consider cases in which the scientists' practical and epistemic reasons come apart. Here is an extension of Competing Scientific Theories:

Impractical Competing Scientific Theories

Unlike Tirrell, you do not do research in the area of the anomalous result, and except for your conversations with Tirrell, the explanation of that result will never have any practical import for you. Moreover, you're giving an important talk tomorrow on your own research. If you consider Jorrie's explanation of the anomalous result, you'll begin to doubt your ability to do science well and you'll perform badly at the talk.

In this case, your practical reasons overwhelmingly recommend ignoring Jorrie's proposed explanation. But, suppose that just before you leave for your talk tomorrow, we ask you what might explain the anomalous result. If you don't have a good reason to ignore it (like by having considered it and rejected it), we can criticize you for disregarding Jorrie's proposal. For example, if you said, "I literally have no idea about what might explain it," we can reasonably expect that you must have rejected Jorrie's explanation for good cause. What is the basis of the negative evaluation we'd give if you simply ignored Jorrie? It can't be a failure to be practically rational (because you're being practically rational). It also can't be an interpersonal failure, like a failure to be nice to one's colleague, since we can additionally stipulate that

6 This case and the cases below begin by stipulating that the agent does have a doxastic attitude about the proposition in question. We do this to avoid objections from evidentialists, like Feldman, who countenance reasons to believe what one's evidence supports only if the agent has or is going to have a doxastic state about the proposition.

no one will find out that you didn't consider Jorrie's hypothesis or that the stakes are high enough to warrant violating such a norm (like if the question were posed by authorities in a murder investigation). Excluding those, the most natural conception of the criticism is that it's epistemic. What the criticism exposes is that if you fail to consider Jorrie's proposal, you're making an epistemic mistake. That mistake is a failure to comply with your epistemic reason to consider Jorrie's proposal.

Another way to argue that the criticism is epistemic and not practical is by looking at the types of agent-impartial features it depends on. Notice that any agent who has the same beliefs and evidence as you is subject to the same criticism when they fail to consider Jorrie's explanation without having a good reason not to. The criticism depends only on the agent sharing those epistemic features with you. Importantly, it doesn't depend on the agent's other attitudes, such as their wants or needs.[7] Since the practical reasons of agents are grounded in that second class of things, the criticism here can't be practical. Instead, the natural conclusion here is that this is an epistemic reason to consider.

Finally, we can see that conceiving of the criticism as epistemic follows naturally from our Gibbardian starting picture of epistemic normativity. Because it is we who aim for truth, not our beliefs, Gibbard (2007, 146) says that a "minimal test" for whether a method of forming beliefs is epistemically rational is whether "if one forms beliefs that way, it will be as if one were, by one's own lights, forming beliefs voluntarily with the aim of believing truths and not falsehoods." Supposedly, you believe that the puzzling titration result stands in need of explanation. If so, then whatever method of forming beliefs that resulted in your ignoring or rejecting Jorrie's explanation without good reason couldn't be one that by your lights was aiming at truth in your beliefs, since Jorrie's explanation was plausible, elegant, and the only available explanation of the result you had. So if epistemic rationality requires that you form beliefs in ways that would be as if you were forming beliefs with the aim of believing truths, then it looks like epistemic rationality requires you to have considered Jorrie's hypothesis (and so, in reasons-speak, you must had had an epistemic reason to consider it).

Could the evidentialist grant that reasons to consider are epistemic but also think that they can be reduced to reasons for (or against) having certain doxastic states? Might she think, for example, that the reason to consider Jorrie's hypothesis can be understood in terms of reasons to believe Jorrie's hypothesis or have at least a minimal credence in it? We don't think that simple reduction works because reasons to consider come prior to reasons to believe. For example, suppose that, despite Jorrie's hypothesis *seeming* elegant and plausible, if you considered it, you'd immediately see that the proposal is actually insane. If so, then it

7 Likewise, one version of practical normativity has it that practical reasons come from some form of decision theory. In that case, what you have a reason to do normally depends on your utility function. Decision theory arguably gives rise to some epistemic norms—such as the maxim to never turn down free evidence—but once we add in a practical cost to an act of evidence gathering, its rationality will depend on assessing that cost against the agent's utilities.

seems like you have a reason to consider her proposal (due to its apparent fecundity), but you have no reason to believe it (due to its insanity that would immediately become apparent to you in considering it). This shows against a straightforward reduction of reasons to consider to reasons to believe.

Might reasons to consider reduce to more sophisticated reasons to believe, like reasons to believe that the hypothesis is plausible or reasons to believe that you should take it to be a possibility? We do not think so. A reason to consider P is a reason to treat P as a serious option for belief, but treating P as a serious option for belief doesn't amount to having any particular beliefs, even including the belief that P is a serious option for belief. Treating P as a serious option for belief is a doxastic *process*. You go through the process in order to facilitate either believing or rejecting the thing under consideration. The process of considering can include activities like testing whether the hypothesis really makes sense of the data or questioning whether it's supported or refuted by other things you know. Believing is not naturally construed as a process like this. So epistemic reasons to consider don't amount to reasons to believe. Reasons to consider are for a kind of mental *doing* that is distinct from having or not having certain beliefs. Because of that, we should think that there are epistemic reasons to consider, in addition to epistemic reasons to believe.

4. Epistemic Reasons to Infer and Calculate

We also have epistemic reasons to infer and do complex calculations, we'll argue. Let's start with a simple inference:

> Lungfish
>
> Sasha is considering whether any fish have lungs, and he's pretty sure none do. He recalls from his childhood biology class that lungfish have lungs, and Sasha knows that lungfish are fish. Sasha has not yet put these two facts together, but he knows that if he were to think about it, he would perform an inference that left him believing more of what his evidence supports.

In this case, the evidentialist should think that Sasha has an epistemic reason to do the inference. Why's that? It's because performing the inference just is how Sasha should come to believe what his evidence supports. To see this, first notice that if Sasha came to believe that some fish have lungs but only because he believed that whales were fish, there's a sense in which he still wouldn't have the doxastic states supported by his evidence. Sasha has evidence for more than just that some fish have lungs. Sasha's evidence also supports basing his belief that some fish have lungs on his two other beliefs, that lungfish are fish and lungfish have lungs. Inferring from those two beliefs to the one belief just is the activity of forming the one belief based on his evidence. So, since the evidentialist grants that Sasha has epistemic reasons to believe what his evidence supports, the evidentialist should grant that Sasha has an epistemic reason to do the inference.

More generally, the evidentialist should accept the following principle:

Reasons from Evidence (Weak)
If ***S*** knows that ***ϕ***ing constitutes ***S*** coming to believe strictly more of what her evidence supports, then ***S*** has an epistemic reason to ***ϕ***.

Since the evidentialist grants that we have epistemic reasons to believe what our evidence supports, the evidentialist should similarly grant that we have epistemic reasons to do the things that constitute coming to believe what our evidence supports. So, in cases like Sasha's, the evidentialist should think that Sasha has an epistemic reason to perform the activity of according his beliefs with his evidence, since that just is what it means for Sasha to do what the evidentialist says he has reason to do. In more complex cases though, the agent might need to take a series of steps to achieve that outcome. Consider this case:

Frog Calculation
Fred doesn't know much about frogs, but he's trying to figure out which garden plots frogs find most hospitable. He's performing an experiment to test how frogs thrive in lily pads compared to mud. Before the experiment, he believed that frogs would do equally well in both conditions, but the initial data appear to show against that initial belief. Fred knows that calculating the effect size would give him a better sense of how much these differences matter. Fred is unable to calculate the effect size in his head though. Fortunately, he's near a computer and can calculate the effect size using it.

Here, when Fred gets the surprising initial data, he is epistemically required to suspend judgment, since that data appear to show against his initial belief. After Fred suspends, he knows that if he performs the effect size calculation, he will come to believe more of what his evidence supports, and as such, he'll be in a better epistemic position by the evidentialist's lights. Because Fred knows that calculating the effect size will bring him to an epistemically better outcome, Fred has an epistemic reason to do the calculation, we'll argue. If asked why he is doing the calculation, Fred could, after all, appeal to the fact that it is a way to figure out what his evidence supports.

Notice that our claim about Fred's reasons doesn't follow from the weak version of Reasons from Evidence given above, since using a nearby computer doesn't *constitute* Fred coming to believe more of what his evidence supports. Using the computer is simply a step in how Fred will form his belief. So to defend our claim about Fred, we'll need a stronger principle:

Reasons from Evidence (First Pass)
If ***S*** knows that ***ϕ***ing would help bring about ***S*** believing strictly more of what her evidence supports, then ***S*** has an epistemic reason to ***ϕ***.

Why might an evidentialist accept this principle? Evidentialists think that we epistemically ought to believe what our evidence supports. In cases where the Reasons from Evidence principle applies, the agent knows that they don't believe everything that's supported by their evidence and know that ϕing would have them believing strictly more of what their evidence supports.[8] The agent epistemically ought to achieve this end and the agent knows they can get strictly closer to it by some means; that must at least give the agent some reason to take those means.

This principle is close to, but doesn't follow from, popular necessity-based transmission principles, which say things like, "If one has a reason to ϕ and ψing is a *necessary means* to ϕing, then one has a reason to ψ" (e.g., Darwall 1983, 16; Kolodny 2007, 251; Schroeder 2009, 245). But, as Kolodny (n.d.) shows, requiring that the means are necessary for reason transmission is too strong. Kolodny argues that the means need only "probabilize" the ends: "If there is reason for one to E, and there is positive probability, conditional on one's M-ing, that one's M-ing, or some part of one's M-ing, helps to bring it about that one Es . . . , then that is a reason for one to M . . . ," Kolodny concludes.[9] Reasons from Evidence does follow from Kolodny's transmission principle, by evidentialist lights. Since the evidentialist says *S* has epistemic reasons to believe what her evidence supports, and *S* knowing that ϕing would help bring about an end requires that ϕing at least 'probabilizes' that end, that general transmission principle entails that *S* has a reason to ϕ.

The evidentialist could also get to the Reasons from Evidence principle in another way. Many evidentialists will accept a picture of epistemology according to which epistemic value consists in agents believing what their evidence supports. If they accept such a picture, then the evidentialist could see our principle as flowing from a more general view about the nature of reasons according to which the existence of reasons is grounded in value production, like the one defended by Maguire (2016). On that kind of view, reasons just are the kind of thing that points an agent to promoting things of value.[10] On that picture, agents will have epistemic reasons to promote the epistemic value of believing what their evidence supports, so Reasons from Evidence will follow.

---

8 Note that an agent knowing that they don't believe everything supported by their evidence doesn't require the agent to believe *of some particular proposition* that their evidence supports it and they don't believe it. By requiring that the agent believe *strictly* more of what their evidence supports, the principle avoids trade-off worries where the agent might have to consider giving up epistemically good beliefs to get many other good beliefs.

9 In the part of the quote excluded, Kolodny outlines (complicated) provisos to the transmission principle, including that the bringing about must not be superfluous and that the end in play be not itself a means to some other end. All of the applications of this principle we use avoid those issues though, so we simplified the presentation for ease of exposition.

10 Note that Maguire's view is committed to a stronger principle than ours though since it links increased value to the existence of reasons. Our principle only generates reasons to ϕ when ϕ-ing would *knowingly* make a better outcome than not.

Two worries remain. The evidentialist might grant that the agent has a reason but insist that it's a practical, rather than epistemic, reason. This is a natural move for the defender of the standard view as well, since as Frog Calculation brings out, Reasons from Evidence will generate reasons for actions, not just beliefs.

On the face of it, it doesn't look like the reason generated would be practical rather than epistemic, though. For one, if the reason were practical, the agent would have to have the particular desires needed to ground that practical reason. Since being subject to the evidentialist requirement to believe what is supported by one's evidence doesn't put any substantive constraints on an agent's desires, the evidentialist would need a story about why agents in particular evidential situations must have the desires needed to ground a practical reason generated by Reasons from Evidence. (Feldman (2000, 676), for example, seems to deny that any such story could be had, since he takes the evidentialist requirement to apply to us solely in virtue of our role as believers, and as such, independently of our desires.) Moreover, if the reason were practical, that would suggest that we could incur practical obligations solely in virtue of having certain kinds of evidence. Of course, gaining new evidence can change our practical obligations *in conjunction with our beliefs, desires, and moral considerations*. Learning that there's a humanitarian crisis might oblige us to donate to certain charities, for example. But if the reasons generated by Reasons from Evidence were practical, there would be practical reasons that are grounded in evidence itself, not one's desires or moral considerations. That would be a surprising result.

In fact, Feldman (2000, 676) thinks that epistemic oughts apply to us in virtue of us occupying a certain role as believers. If that's right, then we should think that the reasons one has in virtue of occupying that role, including the ones that follow from Reasons from Evidence, are also epistemic. It would be a mistake to assume that all reasons generated by one's role as a believer are reasons for belief, just as it would be a mistake to assume that all reasons generated from one's role as a baker are reasons to bake. Being a baker additionally gives one reasons to buy certain ingredients, be well informed about public flavor preferences, and cultivate one's dough-kneading skills. Analogously, what we take the argument to be showing is that being a believer can generate reasons to calculate using a computer, not just reasons to believe.

A second worry one might have about Reasons from Evidence is that, as stated, the principle is unclear about what 'ϕing would help bring about *S* believing strictly more of what her evidence supports' means. In most cases, like Frog Calculation, taking steps to bring it about that one believes more of what one's evidence supports doesn't change the evidence one has very much. Using a computer to analyze data might give you new evidence about what the computer looks like, for example, but in normal circumstances, using a computer to analyze data doesn't change what the data itself support. Nonetheless, we can imagine cases where taking the means to believing what one's evidence supports significantly changes what one's evidence supports. In those cases, it's unclear when the principle applies. Is the antecedent satisfied when one believes more of what one's evidence supports after they ϕ?

Or is it the pre-ϕing evidence that's relevant? And if that's the case, what about cases where after ϕing, the agent's evidence no longer supports what the agent comes to believe?

To resolve these ambiguities, we'll modify the principle in two ways: First, we'll stipulate that the agent has a reason to ϕ only when ϕing would bring it about that the agent believes more of what her evidence supported before ϕing. Second, we'll exclude cases where ϕing importantly changes what one's evidence supports by using a more restricted version of Reasons from Evidence:

> Reasons from Evidence
>
> If at $\boldsymbol{t_0}$, ***S*** knows that $\boldsymbol{\phi}$ing would help bring about ***S*** believing strictly more of what her evidence supports about some propositions at $\boldsymbol{t_0}$ without changing what ***S***'s evidence supports about those propositions, then at $\boldsymbol{t_0}$, ***S*** has an epistemic reason to $\boldsymbol{\phi}$.[11]

Returning to the case of Fred and whether he has an epistemic reason to use the computer, we can see now that he does by a simple application of Reasons from Evidence.

In discussions of this case, our interlocutors often accept that Fred has some reason, but they strongly resist the conclusion that the reason is epistemic. Our interlocutors cite only the fact that it counts in favor of performing a physical action (manipulating the computer) rather than performing a mental "action" in their defense. As we argued above though, we think there are good reasons to take the reasons produced by Reasons from Evidence to be epistemic. But, to assess further whether the physicality of the object of the reason is relevant to the type of reason it is, let's compare Frog Calculation to an internalized version of the same case:

> Mental Frog Calculation
>
> As in the previous case, Mahdi, who shares Fred's initial belief, is also performing an experiment to test how frogs thrive in lily pads compared to mud. Mahdi gets the same experimental data that Fred did. However, Mahdi is great at mental calculation, so instead of calculating the effect size on a computer, he can easily calculate it mentally.

In this case, the objector should admit that Mahdi has an epistemic reason to do the calculation, since Mahdi can comply with his reason with only his mentally available resources. If the objector concedes that though, we take the main battle of this section to be won.

In conceding that Mahdi has such an epistemic reason, the objector admits that there can be epistemic reasons to calculate, which is most of what the argument aims to show here. That's because reasons to calculate, like reasons to consider, cannot be understood in terms of reasons to believe. One can see this by considering an agent who has a reason to calculate

---

11 Note that this principle will not give rise to wrong kinds of reasons cases since all the cases we discuss with it are cases of action, and wrong kinds of reasons problems don't occur for those kinds of cases (Hieronymi 2005).

but no reason to believe the output of the calculation. This might happen, for example, if the output of the calculation is undefined (though the agent doesn't suspect this in advance), perhaps due to some statistical anomaly.

Another way to see that reasons to calculate don't reduce to reasons to believe is to compare two agents who have the same beliefs and evidence over time, but one has the post-calculation beliefs on the basis of doing the calculation and the other has them on the basis of wishful thinking. We take it that the first person is epistemically better off because they respected their evidence in forming their belief in a way that the second did not, despite the two landing on the same beliefs. The most natural explanation is that the first acted in accord with his reasons to calculate and the second did not. Reasons to calculate, like reasons to consider, are reasons to go through a certain kind of process. That process doesn't amount to merely having certain beliefs. To calculate, one must at least also have those beliefs in a certain order and have the subsequent ones on the basis of prior ones. For real agents, calculating also involves going through other mental processes, which typically involve nonbelief-like actions, like carrying the "1" when numbers in the same "column" of an addition sum to more than 9. Doing a calculation then doesn't amount to just having certain beliefs.

What does that show about the possibility of epistemic reasons to calculate using a computer, rather than doing it mentally? At this point, we hope that insisting that epistemic reasons can only count for mental activity strikes the reader as ad hoc. Both Fred's and Mahdi's reasons are reasons to perform a particular statistical calculation. The difference between Mahdi and Fred is the method they use to perform the action. Without begging the question by assuming that epistemic reasons can only be for beliefs, what explanation might one give for why reasons to calculate can only be epistemic if the calculation is done mentally? We doubt there is any principled explanation to be had.

Crucially, to argue that there can be epistemic reasons for activities like calculating with a computer, one need not be committed to anything like Clark and Chalmers' (1998) extended mind hypothesis, which holds that the environment outside of one's body can be part of one's mind. Though accepting their view would make this argument easy, you can reject their view while accepting this argument. All this argument requires is that calculation can be done outside the body (an uncontroversial claim), and that a reason's status as epistemic does not depend on how it is complied with, at least in some cases involving external means.

So it looks like Fred and Mahdi share a reason to calculate. For Mahdi, it's easy to comply with it by calculating mentally. For Fred though, calculating mentally is not a real option, but he does have a computer. So Fred has an epistemic reason to calculate using the computer.

5. Epistemic Reasons for (Thought) Experiments

What the previous section purports to show is that the evidentialist should accept that there are epistemic reasons to take simple means to believe what one's evidence (already) supports, regardless of whether those means are internal to the agent's mental life. Here we'll extend

the argument to show there can also be epistemic reasons to perform more complex actions and actions that extend one's evidence.

Consider this case:

> Sphere Eversion
> Serena has a budding interest in topology, but she hasn't learned much about spheres. She thought that spheres could not be everted (turned inside out) without cutting, tearing, or creasing, but in conversation, her mathematician friends seem to presuppose that they can. Serena can't imagine how that would work, but she has a spherical toy on her desk and knows that manipulating the toy will help her get closer to figuring it out.

The argument below will conclude that Serena has an epistemic reason to manipulate the desk toy.

First notice that it's not obvious that a simple argument like the one used in the previous section will work in this case: After suspending her belief about the evertability of spheres, does Serena *know* that she would believe more of what her evidence supports if she manipulates the desk toy?[12] For instance, on a phenomenal view of evidence, where one's evidence consists in a set of seemings, Serena's initial evidence need not support that a sphere can be everted even though it is a necessary truth. If she doesn't know that she will come to believe more of what her evidence supports by playing with the desk toy, it doesn't follow from Reasons from Evidence that she has an epistemic reason to do so.

We'll start with an internal analogue of the case under consideration:

> Mental Sphere Eversion
> Alice has a budding interest in topology, but she hasn't learned much about spheres. She thought that spheres could not be everted without cutting, tearing, or creasing, but in conversation, her mathematician friends seem to presuppose that they can. Alice knows that she's very good at mental shape transformations, and she knows that manipulating a mental model of a sphere will help her get closer to figuring it out.

Like in the previous section, we'll give an argument that Alice has an epistemic reason to perform the thought experiment involving a mental manipulation of a sphere and then infer that Serena has an analogous epistemic reason to do that same process using her desk toy, since there is no epistemically relevant difference between the two cases.

---

12 To be as ecumenical as possible, our argument won't assume that our evidence always supports logically necessary propositions. Relaxing this assumption will mean that the relation "is evidence for" will not be transitive, a point we'll rely on in our argument. If, however, one's evidence does always support logically necessary propositions, this case will reduce to the cases in the previous section where the initial evidence does indeed support the conclusion at the outset.

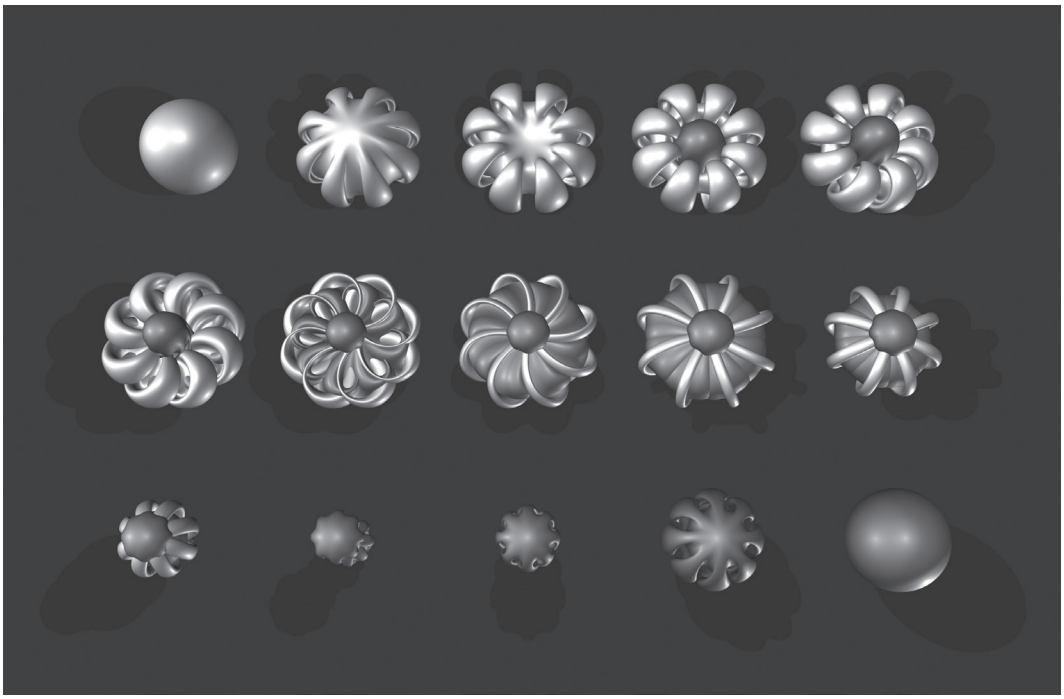


Figure 1 Steps to evert a sphere. Image reprinted with permission from Silvio Levy.

In fact, spheres are evertable, but it's hard for most people to imagine how that might work. (See figure 1 for one way to do it.) So, let's grant that the evertability of the sphere is not supported by Alice's evidence before she begins the thought experiment.[13] Because of the complexity of eversion, we can assume that if Alice were to do the thought experiment, she would have to go through many stages of imagining shapes, starting with a sphere and moving to the next shape by simple permissible transformations. Suppose that the stages of Alice's thought experiment are represented by figure 1.

Now suppose Alice has already started the thought process, and she's currently visualizing a shape at some stage of the transformation. As stipulated, Alice is good at mentally manipulating shapes. If she visualizes a shape, it is obvious to her what other similar shapes can be reached from the visualized shape by small transformations. So, when Alice is visualizing a certain shape, at that moment, her evidence supports thinking that the next shape can be reached by small transformations, even if Alice doesn't yet have a belief about that particular shape.[14]

So suppose (1) Alice has progressed to some stage S_n of the progression and is visualizing that shape, (2) she does not yet believe that S_{n+1} is possible to reach from a sphere without

13 If our opponent denies this, then the argument from the previous section applies directly here.

14 We encourage the reader to try this. Imagine an ellipse, for example. Isn't it obvious that the ellipse can become "dented" on one side without cutting, tearing, or creasing it? It's obvious because when you're visualizing it, your evidence supports thinking that it's possible.

cutting, tearing, or creasing, and (3) this pair of stages S_n, S_{n+1} is the last one in the sequence that meets conditions (1) and (2). (Such a pair must exist if she hasn't already figured out how to evert the sphere.) Because Alice is good at mentally manipulating shapes, we know that at S_n, Alice's evidence supports thinking that the shape at S_{n+1} can be reached. Moreover, Alice knows she's good at mental manipulations of shapes, and she knows that when she visualizes a transformation, she correctly believes whether it is possible. So, she knows at S_n that if she continues the visualization process, she will believe more of what her evidence supports at S_n. She can do this by moving from stage S_n to stage S_{n+1} of the visualization process. Thus, by Reasons from Evidence, she has an epistemic reason to do the mental transformation from S_n to S_{n+1}.[15]

What about the worry that visualizations in thought experiments don't come in discrete stages? Presumably, Alice visualizing one shape transforming into another will only induce the belief that the transformation is possible if Alice can visualize the transformation happening in a smooth way that doesn't admit of discontinuities or jumps between shapes. Since Alice has a strong skill in visualizing, we can assume she knows this as well. So Reasons from Evidence tells us that, as she goes through the process, Alice gets a series of epistemic reasons to continue going through the process of smoothly visualizing the shape changing, not just to visualize discrete stages of that transformation.

What this shows is that, once Alice starts the process, she has epistemic reasons to continue the process. Similar reasoning entails that Alice has an epistemic reason to begin the process: Consider Alice's state of mind before she starts visualizing. As stipulated, she knows a little about topology, and she originally thought sphere eversion was impossible. When her mathematician friends presupposed otherwise, Alice at least has an epistemic reason to suspend her prior belief. But since Alice knows some things about topology, she must know basic generalizations about how transformations can apply to shapes. For example, she knows the generalization that if a shape "*o*" can be reached from a sphere, then the shape "*O*"—shape "*o*" stretched vertically—can also be reached. She also knows that if "*u*" is a shape with a rounded edge that can be reached, then "*w*" can also be reached, since "*w*" is "*u*" with a new rounded indentation added to the bottom edge. Because she knows generalizations like these, Alice must know that there is some shape such that she doesn't believe it to be reachable from a sphere but her evidence supports thinking it is reachable. Why's that? It's because Alice knows that she doesn't already have a complete map of how spheres can be transformed. If she did, she would know already whether spheres are evertable. So if shape *B* is one she knows is reachable from a sphere, there must be some shape *B'* that can be reached by a simple transformation from *B*, even though Alice doesn't yet believe *B'* is

15 Since the argument doesn't assume that her initial evidence supports the proposition that the sphere is evertable, this reasoning also shows that the relation of evidential support is not transitive in the sense that (if one's evidence at t_1 supports believing P and if one believed P at t_2 then one's evidence would support believing Q), then one's evidence at t_1 supports believing Q.

reachable. Since she knows basic generalizations about simple transformations though, we can assume that the *B*-to-*B'* transformation is characterized by one of the generalizations she knows. So her evidence must support thinking that *B'* is reachable from a sphere, since it follows from the things she knows by simple logical entailment (namely, that *B* is reachable and the generalization that characterizes the *B*-to-*B'* transformation). So Alice must know that there are some shapes whose reachability is supported by her evidence but that she doesn't yet have beliefs about.

From that, it follows that she knows that if she visualizes a sphere and mentally applies the transformations in a thoughtful way, she will move from visualizing shapes that she already knew were possible to visualizing shapes that she didn't yet know were possible but whose possibility was already supported by her evidence. As such, Alice will know that she will believe more of what's supported by her present evidence if she begins the visualization, so by Reasons from Evidence, she has an epistemic reason to do so.

One might worry that this reasoning, if correct, would generate too many epistemic reasons. Our evidence almost always supports beliefs that we shouldn't waste our time forming. At any given time, we have evidence that bears on all kinds of things like whether the room we're in could safely hold more or less than twelve brown bears or whether Trump's hair could really just be three blades of oddly shaped grass. So, doesn't the kind of reasoning here entail that we have reason to do things to get ourselves to have the beliefs our evidence supports about those propositions? Doesn't it entail that we have epistemic reasons to imagine brown bears stacked in the room we're in and epistemic reasons do a thought experiment about Trump's hair?

No, the reasoning does not have that implication. Notice that in the kinds of cases the objector mentions, the agent isn't assumed to believe that imagining bears or thinking about Trump's hair would help her believe more of what her evidence supports. But Reasons from Evidence requires that the agent *know* that her action will bring about her believing more of what her evidence supports, so it doesn't apply to cases where the agent doesn't even believe it.

One could also respond to this worry by modifying the argument: Notice that, as we put it above, the evidentialist view is subject to a similar overgeneration worry. The evidentialist says that one ought to have the doxastic states supported by their evidence, which seems to entail that one ought to have beliefs about all kinds of things that intuitively we can simply ignore, again like how many brown bears could safely fit in this room. Feldman (2000, 679) considers this objection and modifies his view to avoid it by relativizing it to a particular proposition and making the normativity conditional on the agent having any doxastic state toward that proposition. So instead of making the blanket claim that everyone ought to have exactly the doxastic states supported by their evidence, the revised evidentialist view only says that if an agent has a doxastic state about some proposition, then she ought to have the doxastic state that her evidence supports having. Our argument can employ the same move to avoid the overgeneration worry. By relativizing Reasons from Evidence to a

particular proposition and making it conditional on the agent having a doxastic state about the proposition, it no longer follows from the principle that the agent has the kinds of epistemic reasons the objector was worried about, since the agent wasn't going to have a doxastic state about those things anyway.

Relativizing Reasons from Evidence to a particular proposition and making it conditional in that way doesn't threaten our conclusion. In each of the cases described, there is a particular proposition in play and the agent is stipulated to have a doxastic state about that proposition (suspension, in particular). Since the goal of the argument is conservative, in the sense that it aims to argue for the existence of these reasons, not show that they're abundant, the few (somewhat contrived) cases we give are sufficient.[16] So there is no overgeneration worry for the argument we present here, and it follows that there are epistemic reasons to begin thought experiments and continue doing them once one has started. Since there is no epistemically relevant difference between doing that and doing the experiment with a desk toy, the argument similarly concludes that Serena has an epistemic reason to manipulate the desk toy.

One interesting implication of this section compared to the previous section is that, in this section, the argument purports to show that there are reasons to do things that end up giving the agent new evidence. One might have thought that no new evidence is gained during the thought experiment in the way we describe it, since at each successive step, Alice only comes to believe what was supported by her evidence in the previous step. That's not right though. When Alice acts in accord with her epistemic reasons to visualize the next step, the new visualization itself gives her new evidence about what future steps are possible. So though she only believes what's supported by her prior evidence at each successive step, she gains new evidence by virtue of how she comes to form that belief, namely by visualizing its possibility. The case for this is even clearer when the reason is one to manipulate the desk toy. So, epistemic reasons can (perhaps surprisingly) be epistemic reasons to do things that will result in gaining new evidence.

6. Epistemic Reasons to Act

Something that is brought out by the cases above is that, if the argument is right, what epistemic reasons an agent has can depend on highly contingent and nondoxastic features of the agent. Alice has an epistemic reason to think through the thought experiment in virtue of her having a particular psychological makeup that allows her to easily manipulate mental models of spheres. And Fred has an epistemic reason to compute using the computer because of his close proximity to it. These reasons are quite particular to those agents, but

16 That said, in fact, we'd be happy to be saddled with the conclusion that there are epistemic reasons to do all kinds of things. Of course, in cases like the objector mentions, the reasons would be quite weak, so they wouldn't entail much about what the agent epistemically ought to do.

what we've seen is that accepting the existence of those reasons is naturally supported by accepting the more general claim that everyone has epistemic reasons to believe what their evidence supports.

In the final part of the argument, we'll extend that idea to show that there can be epistemic reasons for even the most mundane actions. Let's consider Alice again:

> Distracted Alice
>
> Alice has been imagining transformations of spheres for an hour in search of the elusive eversion technique. She finds herself getting hung up on repeating a particular sequence of transformations that she's already determined to be unhelpful. To stop herself from getting distracted by that particular sequence, she knows she could perform a mindfulness exercise that involves focusing on the failed step.

Notice that what happens when Alice gets distracted in this case is that she metaphorically takes a wrong turn in her thinking—her overall sequence of steps leads her astray from her goal. The defender of the standard view will surely see the deviation from her goal as a practical failing, so let's grant that. That said, she still has epistemic reason to do the mindfulness exercise, not because it will help her reach her goal, but because doing so will knowingly bring her to believe more of what her evidence supports.

One supported thing she'll come to believe is that the particular sequence of steps she's stuck on isn't useful. There's something else she'll believe too though. Recall that the previous section established that at each step of the thought experiment, she has reason to continue to the next step because in doing so, she knowingly comes to believe more of what her evidence supports. Then let S_l be the last stage in her thinking before the distracting sequence. Alice knows that, in theory, she could restart her thinking at S_l and visualize a different transformation of the shape and in doing so form new beliefs that are supported by her current evidence.[17] But, as stipulated, Alice is getting distracted whenever she gets to S_l, and she knows this. She also knows that doing the mindfulness exercise would allow her to move past the distractions at S_l and believe what her evidence supports about how the shape at S_l can be transformed. So again by Reasons from Evidence, she has an epistemic reason to do the mindfulness exercise.

The argument developed in the sections before this one purported to show that one can have epistemic reasons to do things like consider hypotheses, calculate effect sizes, and learn about spheres by manipulating a physical model. If you started reading this with a liberal conception of what epistemology is about, none of those results would have been

17 Recall that above we showed that if Alice doesn't already know that the sphere is evertable, there must be some stage of the sequence where this is true. The argument is also supposing that Alice remembers, and her evidence still supports, that a sphere can be transformed into the shape at S_l, which is what would make the new beliefs also supported by her evidence.

very surprising. All of those activities are paradigmatic elements of what's involved in both scientific research and general deliberation, both of which are squarely in a broad conception of the domain of epistemology. What's interesting about this stage of the argument is that, unlike the previous cases, it purports to show that we can also have epistemic reasons to do activities that will strike almost no one as epistemic. In Distracted Alice, that activity is the mindfulness exercise. The point generalizes much further though:

> Hungry Alice
>
> Alice got past the distracting sequence, but now she has been imagining transformations of spheres for over 4 hours. She's close to everting the sphere, but as she tries to go on, she finds that her low blood sugar is making her foggy-headed and unable to concentrate. Alice knows that she could eat a sandwich to raise her blood sugar, which will help her move on to the next stage of the thought experiment.

Here, Alice has an epistemic reason to eat the sandwich. No additional argument for this is necessary, since the argument for why Alice has an epistemic reason to do the mindfulness exercise, *mutatis mutandis*, applies equally well. Alice has a reason to eat the sandwich because, again, doing so will knowingly make her believe more of what's supported by her evidence.

This section just takes the reasoning of the previous two sections to their natural end. The argument shows that even evidentialists should think that there can be epistemic reasons to do all kinds of mundane activities that aren't typically thought of as in the domain of epistemic reasons. We discussed this here using the example of Alice needing mindfulness exercises and sandwiches to make it through a thought experiment. But, Mahdi could have faced the same distractions as Alice when he tried to mentally calculate the effect size in his frog experiment, and if he did, by this line of argument, he too would have had an epistemic reason to eat a sandwich. The same is obviously true in a broad range of other possible cases.

7. Conclusion

The argument laid out above purports to show that even those who start out only thinking that there are epistemic reasons to believe what is supported by one's evidence should allow that there are also epistemic reasons to suspend belief, consider hypotheses, do both thought and physical experiments, extend one's evidence, and perform mundane tasks like eating a sandwich. To create counterexamples to the standard view, our argument relied on intuitions and arguments about particular cases. But similar arguments can show that there are epistemic reasons to ask the right kinds of questions, buy certain statistical software packages, set up one's office in a certain way, hire a certain post-doc, get feedback from others on papers, apply for grants, and pay attention to certain things rather than others.

If this argument is right, it allows us to conceive of epistemic normativity as the kind of normativity that governs all aspects of knowledge creation and dissemination, not just what beliefs one ought to have in particular evidential situations. This is the picture of epistemic normativity that falls out of Gibbard's idea that it is we who aim for truth, not our beliefs.

That said, what we should learn from the argument is meager in the sense that it only shows that there are some epistemic reasons for these things, not that these reasons are strong, abundant, or sufficient for action. It is compatible with our argument that, all things considered, acting on these reasons is supererogatory, merely permissible, or even prohibited. In that way, if our argument is right, we should see the realm of epistemic reasons as similar to the realms of moral and practical reasons. Just as there can be practical and moral reasons to do all kinds of things that one ought not do all things considered, the same would be true of epistemic reasons.[18]

As we mentioned at the beginning, we don't see this argument as being about reasons *per se*. What the argument really shows is that epistemic normativity, regardless of whether it's understood in terms of epistemic reasons or epistemic rationality, governs not just our beliefs but all sorts of other things, including what we eat. In that way, if the argument is right, it goes against the long-standing dogma of meta-epistemic theorizing that epistemic normativity is normativity of belief. But, as mentioned above, the goal of this discussion was only to lay out the argument and show why we find it plausible, not to assess the full case for or against its conclusion.

References

Berker, Selim (2017). "A Combinatorial Argument against Practical Reasons for Belief." *Analytic Philosophy*, 59, no. 4, 427–70.

Clark, Andy, and David Chalmers (1998). "The Extended Mind." *Analysis* 58, no. 1: 7–19.

Conee, Earl, and Richard Feldman (2004). *Evidentialism*. Oxford: Oxford University Press.

Darwall, Stephen L (1983). *Impartial Reason*. Ithaca, NY: Cornell University Press.

Feldman, Richard (2000). "The Ethics of Belief." *Philosophy and Phenomenological Research* 60, no. 3: 667–95.

Feldman, Richard, and Earl Conee (1985). "Evidentialism." *Philosophical Studies* 48, no. 1: 15–34.

18 One possible conclusion from the arguments we've presented is that just as epistemic normativity extends to objects other than belief, so does deliberation and even evidential support. In sphere eversion and hungry alice, playing around with a desk toy and eating a sandwich contribute to figuring out an answer in much the same way as the internal mental activities of thought experiments and mental exercises. And further, these activities, whether mental or external, play a role in processing and understanding relations of evidential support. On this basis, one author suspects that deliberation, understood as the activity of figuring out what you should do and believe based on your evidence, might still be the sole locus of epistemic rationality even if the argument is right that epistemic reasons go far beyond belief.

Gibbard, Allan (2007). "Rational Credence and the Value of Truth." In *Oxford Studies in Epistemology*, edited by Tamar Szabo Gendler and John Hawthorne. Oxford: Oxford University Press.

Goldman, Alvin (1999). *Knowledge in a Social World*. Oxford: Oxford University Press.

Hieronymi, Pamela (2005). "The Wrong Kind of Reason." *Journal of Philosophy* 102, no. 9: 437–57.

——— (2013). "The Use of Reasons in Thought (and the Use of Earmarks in Arguments)." *Ethics* 124: 114–27.

Kelly, Thomas (2002). "The Rationality of Belief and Some Other Propositional Attitudes." *Philosophical Studies* 110, no. 2: 163–96.

Kolodny, Niko (2007). "How Does Coherence Matter?" *Proceedings of the Aristotelian Society* 107, no. 1pt3: 229–63.

——— (n.d.). "Instrumental Reasons." In *The Oxford Handbook of Reasons and Normativity*, edited by Daniel Star. Oxford: Oxford University Press.

Kornblith, Hilary. 1983. "Justified Belief and Epistemically Responsible Action." *The Philosophical Review* 92, no. 1: 33–48.

Leary, Stephanie (2017). "In Defense of Practical Reasons for Belief." *Australasian Journal of Philosophy* 95, no. 3: 529–42.

Maguire, Barry (2016). "The Value-Based Theory of Reasons." *Ergo* 3. https://doi.org/10.3998/ergo.12405314.0003.009.

Markovits, Julia (2014). *Moral Reason*. Oxford: Oxford University Press.

Parfit, Derek (2001). "Rationality and Reasons." In *Exploring Practical Philosophy, Essays in Honour of Ingmar Persson*, edited by D. Egonsson et al. Aldershot: Ashgate.

Raz, Joseph (2009). "Reasons: Practical and Adaptive." In *Reasons for Action*, edited by David Sobel and Steven Wall. Cambridge: Cambridge University Press.

Reisner, Andrew (2009). "The Possibility of Pragmatic Reasons for Belief and the Wrong Kind of Reasons Problem." *Philosophical Studies* 145, no. (2): 257–72.

Rinard, Susanna (2015). "Against the New Evidentialists." *Philosophical Issues* 25, no. 1: 208–23.

Schroeder, Mark (2009). "Means-End Coherence, Stringency, and Subjective Reasons." *Philosophical Studies* 143, no. 2: 223–48.

Shah, Nishi (2006). "A New Argument for Evidentialism." *Philosophical Quarterly* 56, no. 225: 481–98.

Singer, Daniel J. (2018). "How to Be an Epistemic Consequentialist." *Philosophical Quarterly*, 68 (272), 580–602.

Sylvan, Kurt (2016). "Epistemic Reasons I: Normativity." *Philosophy Compass* 11, no. 7: 364–76.

Turri, John (2009). "The Ontology of Epistemic Reasons." *Noûs* 43, no. 3: 490–512.

II

WARRANTED FEELINGS

5

ASSESSING FEELINGS

Simon Blackburn

I

In a recent paper about his views and mine, Allan Gibbard was kind enough to mention the first time we met, in Michigan in 1983. For me that meeting was the beginning of a friendship that I have cherished ever since, and from that time, thirty-five years ago, Gibbard has been a fixed point in my own philosophical life. From the very beginning I admired not only his work but the whole cast of his mind, and the many years since then have only seen that admiration grow. In dark days when books and papers were pouring off the presses excoriating expressivism, I always took comfort from being able to reflect that if a philosopher of Gibbard's stature continued to refine and defend it, then it couldn't be all that bad. I would have felt more oppressed if the criticisms had been directed at Blackburn's expressivism; when they were directed instead at Blackburn and Gibbard's expressivism, I slept more easily. So it is a great pleasure and a privilege for me to contribute to this volume, honoring someone I regard as a touchstone of philosophical acumen and imagination.

In the paper I mentioned Gibbard explores, with great delicacy and subtlety, one of the few issues on which he and I appear to diverge.[1] He says, of this divergence, "My impression is that Blackburn sees no need for this aspect of my views. He doesn't denounce it, as far as I know, though Blackburn knows how to denounce. I wonder whether, on this point, I am the rare fool whom Blackburn suffers gladly." So let me say at once that I certainly would not dream of denouncing Gibbard's view, and even more certainly do not regard him as any

1 Allan Gibbard, "Improving Sensibilities." In *Passions and Projections*, *Themes from the Philosophy of Simon Blackburn*, edited by Robert Johnson and Michael Smith. Oxford: Oxford University Press, 2015.

kind of fool. But as an aside, I want to enter a small plea. I really think I am quite slow to identify anyone as a fool, and even quite good at suffering the few I do so identify gladly. As when dealing with students, I think that I express annoyance only at writings that seem to me not so much foolish—in fact quite often not foolish at all—but instead false to some values I take to be central to the spirit of philosophical enquiry. An enquiry should be an unswervingly honest and cautious attempt to see if a suggested opinion stands up, whether it needs modification, or even rejection. I hope I only get hot under the collar about writings that are too polemical, prejudiced, one-sided, or glib to fit this description, and I cannot imagine Gibbard presenting any such thing, any more than I could imagine other giants, such as David Lewis, doing so.

The divergence that troubles Gibbard concerns a certain kind of psychological dissonance, and the ingredients we need in order to make sense of it. Gibbard's central example was the time in the 1960s when young men started wearing their hair long, and, he confesses, he felt disapproval, "but when I asked myself whether anything was really wrong with long hair for men, I answered myself no." In his later terminology, he felt that his own feelings were not warranted, and to say that a feeling is or is not warranted expresses a special state of mind, which involves what Gibbard calls "accepting a norm." The special nature he accords to this state is an aspect of his own views that, he confesses, he finds suspicious, but he doesn't know how to do without it.

The problem is not raised by any old case of expressing a norm, for after all Gibbard's feeling that it was wrong for youths to have long hair is itself a matter of a norm, and if he told a friend of this he would be expressing it. But this is not in itself the problem he is posing. The problem is set rather by cases in which someone has an attitude or a feeling but is at the same time afraid that the attitude or feeling is not warranted. They are, as it were, worried about the justice of their own reaction. The difference between us is not, of course, that I don't recognize the phenomenon, nor do I mind saying that the second of the mental states in question involves "accepting a norm." The difference is that Gibbard is afraid that it might take what we could call non-Humean ingredients to identify this state of mind, whereas I am more sanguine or more relaxed about it. I might mention in passing that there is a pleasing symmetry about Gibbard wondering whether I treat this too lightly, since I once wondered similarly whether he had helped himself too lightly to materials that enable us expressivists to pass over Frege's abyss, separating the propositional from the purely ejaculatory.[2] I now think that criticism was misjudged. But in each case we have a potential stumbling block in front of the project we share, of giving a naturalistic account of humanity's involvement with normativity.

If it were not for this ambition, then it would be quite easy to describe Gibbard's state, simply by using normative language. Gibbard felt queasy about seeing young men with long hair, but he did not think this state was justified, or right, or warranted or "fitting" and

2 Simon Blackburn, "Gibbard on Normative Logic." *Philosophical Issues* 4 (1993): 60–66.

he feared that there were no good reasons for his discomfort. The philosophers sometimes called "reasons-primitivists" such as A. C. Ewing, or in the recent scene Tim Scanlon or Derek Parfit has no difficulty in saying that, and in effect stopping there.[3] But on the face of it they are also involved with nonnatural, Moorean notions of fittingness, or of a primitive relationship denoted by one thing being a reason for another, and few naturalists are likely to be satisfied with these as useful stopping points. So the problem becomes one of understanding these locutions and the qualities of states of mind that they serve to express, on a natural basis.

There are, of course, different ways of regarding the requirement of naturalism. Writers such as John McDowell or Peter Strawson counsel a relaxed naturalism that has no particular difficulty including the normative elements in our language and our natures, whereas others find it necessary to do some work to understand these.[4] Gibbard and I belong to the second group. I do not regard this as a matter of finding the right definition of naturalism, as if it has a Lockean real essence we should set about finding. Instead I think it is a question of explanatory ambition: just how far back can we start in order to give the most insightful genealogy of our moral natures, or indeed our normative tendencies in general? I return to this below.

Gibbard recognizes that in *Ruling Passions* I addressed this problem, but he may well feel that I prowled around it rather than clearing it up satisfactorily.[5] I mentioned and rejected a simple proposal for a solution, which applied to the present case would have it that on the one hand Gibbard's younger self disliked long hair on young men, but on the other he desired not to have this dislike. This would be the account along lines discussed by Moore and Frankfurt, identifying our values not by means of what we desire but by what we desire to desire. The main problem I highlighted for this account is posed by the case of Satan, whose sense of the evil he follows and the good he has lost does not issue in desire to change at any level. I thought, however, that some modifications of this idea would put us on the right track. In order to count as being aware of his own evil, Satan, I supposed, would at least *regret* his state, or *rue* the day he fell from his previous state, or in some similar way manifest a split between what he actually desires and what he can "identify with," where his being unable to identify with his desire would issue in such states of mind as discomfort, irritation with himself, discontent, or even a wish he could have been different, although that wish need not issue in an actual desire, for wishes with counterfactual content typically do not. When our desires appear to ourselves not to be appropriate to their objects there is a disruption of the pleasant harmony between our desires, on the one hand, and the sense of their defensibility or propriety, on the other, that we like to enjoy.

---

3 A. C. Ewing, *The Definition of Good*. New York: Macmillan, 1947; Tim Scanlon, *Being Realistic about Reasons*. Oxford: Oxford University Press, 2014; Derek Parfit *On What Matters*, vols. 1–3. Oxford: Oxford University Press, 2011–17.

4 John McDowell, *Mind, Value, and Reality*. Cambridge: Harvard University Press, 1998; Peter Strawson, *Skepticism and Naturalism: Some Varieties*. New York: Columbia University Press, 1985.

5 Simon Blackburn, *Ruling Passions*. Oxford: Oxford University Press 1998, 58–68

One aspect of it that I think I hadn't fully recognized at the time is the "wrong kind of reason" problem, introduced by Wlodek Rabinowicz and Toni Ronnow-Rasmussen.[6] Here someone may have reasons for wanting to hold onto an attitude, but which are not reasons for supposing the attitude to actually fit its object, or equally she might have reasons for wanting to change an attitude, which are not reasons for supposing the attitude unfitting. There may be threats or rewards surrounding admiring someone, for instance, providing a motive for doing so independent of whether there is anything admirable about them. Or, there may be costs to feeling some way about something arising not from threats or bribes but in other ways. Gibbard gives the example of someone who feels outrage over bullying, but "perhaps feeling outrage over bullying doesn't improve the situation; the bullying will go on undiminished however I feel about it and my feelings of outrage about what I can't prevent just depress me." Here there is a reward in sight for becoming indifferent to bullying, namely freedom from depression, and hence this person might regret his sensitivity to the evils of the world but still think bullying is abhorrent and deserves outrage. So the challenge remains. Is it possible to delineate more accurately the kind of dissonance involved, using only Humean materials?

Part of the difficulty here may be knowing just what count as Humean materials. Suppose, for instance, the naturalist suggests that holding an attitude to be fitting is a matter of identifying yourself with it, and that in turn is a matter of feeling able to stand by it, or defend it, and at the very least feeling no shame or guilt about holding it. In this example, the subject may wish he didn't feel so strongly about bullying. But he would not feel ashamed of feeling so strongly about it. Nor would he feel less proud of another, his daughter say, on finding that she shares his abhorrence of bullying. Whereas if she says, "I used to feel like that, but I realize it did no good, and now I just don't care," then his daughter is not really presenting herself in an admirable light, and perhaps he becomes a little less proud of her. Would this be enough for a Humean to solve Gibbard's problem? Perhaps not: pride and shame are themselves moral emotions, so appealing to the special kinds of comfort or discomfort that they signal is, perhaps, ducking the challenge. One can imagine a reasons-primitivist insisting that they in turn signal an awareness of the right reasons in the case, with pride following upon supposing that we are aligned with them (or, in this case, that our daughter is), and shame, the feeling that in some respect we are in an indefensible position, following upon supposing that we are not after all aligned with right reason.

Perhaps so, but just as I have felt bolstered by finding Gibbard standing beside me, so now, I suggest, we might together feel bolstered by the formidable naturalist history standing behind the pair of us. The first element is Humean: we should remember that when Hume turns to moral psychology in Book II of the *Treatise* the very first sentiment that he treats is that of pride. A major support of his thinking about this is due to Adam Smith. Following on from Hume, Smith explores in some detail the mechanisms of internalization, whereby

6 Wlodek Rabinowicz and Toni Ronnow-Rasmussen, "The Strike of the Demon." *Ethics* 114 (2004): 391–423.

the actual or imagined attitudes of others become the voice of the impartial spectator, the 'man within the breast' whose verdicts on ourselves we have to listen to, sometimes uncomfortably enough.[7] These two give us inroads into the problem of dissonance between how we feel, on the one hand, and what we feel comfortable or even proud about feeling, on the other. Thus, I may feel bitter and resentful on being told that a rival has got some benefit that I had hoped would come to me. But if I am half aware that an impartial spectator would see nothing unjust about this outcome—for the rival deserved it just as much as I did—then I cannot expect others to feel indignation on my behalf (in Smith indignation is the third-person counterpart of private resentment). My resentment then becomes a social orphan, unsupported and unloved. It would be better to be without it. In this way, the imagined gaze of the impartial spectator is enough, sometimes, to give us pause. I return to this approach below, in connection with our interests in achieving a common point of view with others.

A further naturalist project in contemporary work is the pragmatist substitution of genealogy for analysis. The idea is that a naturalist agenda can be pursued not by traditional "reductions," seeking to analyze apparently awkward customers into naturalistically respectable components, but by advancing accounts of how creatures such as ourselves, starting with an intelligible endowment of mental states, and having typical human problems to solve, might be expected to end up talking, thinking, or feeling as we do. A paradigm of such an account is Hume's own treatment of the emergence of conventions, whereby self-interested creatures such as ourselves, anxious only to secure benefits and avoid threats and costs, might naturally evolve the combinations of restraints and expectations that follow upon the arrival of conventions—including those conventions governing such institutions as language, money, property, promises, law, and government.[8] These accounts creep up on the complexity of the normative way of thinking, in ways taking off from relatively simple interactions, such as tit-for-tat behaviors, that could be found amongst nonlinguistic or protolinguistic communities. Expressivism is of course itself an example of this kind of naturalism in action. Deflationist views about truth are of the same kind.

Again, we need to be careful about the naturalistic credentials, and such approaches face two kinds of hurdle. The first is that the starting point for the stories they sketch should be free of whatever problems made the story desirable in the first place. The second is that there should be no saltations or incredible jumps in the imagined evolution. Applied to morality this means that moral commitments must not be felt to belong to an especially spooky area of knowledge, in danger of requiring transactions with nonnatural properties and relations. There should be nothing similarly spooky in the psychological starting point,

7 Adam Smith, *The Theory of Moral Sentiments*, Part 1, Section 1, Chapter 4.

8 David Hume, *Treatise*, III, 2, 2; *Enquiry Concerning the Principles of Morals*, Appendix 3. Hume's is of course a "bunking" genealogy, fitted to produce increased confidence in what we have done and what we have become. It contrasts with the more familiar "debunking" genealogies associated with Nietzsche. But for a more nuanced view of Nietzsche's overall work, see Matthieu Queloz, "Nietzsche's Pragmatic Genealogy of Justice." *British Journal for the History of Philosophy*, 25, no. 4 (2017): 1–23.

and no jump from the natural to anything nonnatural at the end point. This is certainly the intent of Hume or Smith's forays into moral psychology.[9]

Returning to Gibbard's challenge, I think we should first notice that the *kind* of dissonance involved seems to go far beyond cases with a peculiarly moral flavor. For in fact, almost across the board, whenever we enter a verdict on the basis of a personal feeling or attitude, we seem capable of checking ourselves, either wondering whether we are wrong, so that our verdict is not the one justified by the case, or fearing that it is even *likely* to be wrong, so that we ought to some extent to discount our own reaction. Here are some examples.

X sees the cloth as brown, but is unsure whether it is actually brown, since the artificial light has a sodium tinge.
X hears the music as a jumble of notes, but is unsure whether it is a melody he is failing to catch.
X was not herself bored by the play, but fears that it might have been boring, because she knows her interest was sustained by the fact that her daughter was acting in it.
X does not himself enjoy the wine, but can tell that it may be excellent. He is a jaded connoisseur, who knows his wine, but has lost his appetite for it.
X finds Y somewhat creepy, but suspects he may be perfectly likeable, and just has an awkward manner.
X enjoys the paintings of Renoir, but has been led to worry whether they are too easy and sentimental.

We could go on indefinitely. In none of these cases is there a direct question of a moral judgment at stake. But there is a possibility, which X might cheerfully or ruefully accept, of herself being deficient, or the circumstances in which they have been exposed to the object of their feeling being "off" in some way, either making them reluctant to advance their own judgment as authoritative or significant or meaning that their opinion is to be discounted if they are not so reluctant. Either they themselves are suffering some disadvantage or perhaps the situation is not one in which we expect judgment to be reliable. As Hume remarked in his great essay on the topic,

> A man in a fever would not insist on his palate as able to decide concerning flavours; nor would one, affected with the jaundice, pretend to give a verdict with regard to colours. In each

9 "Normativity" arrives on the scene with positions being ones that attract social approval and disapproval, expressed through sanctions: "when a man says he promises anything he in effect expresses a resolution of performing it; and along with that, by making use of this form of words, subjects himself to the penalty of never being trusted again in case of failure." Hume, *Treatise* III, ii, vi, 522.

> creature, there is a sound and a defective state; and the former alone can be supposed to afford us a true standard of a taste and sentiment.[10]

Together with this there are better and worse situations for judgment. We do not or should not trust our own judgment if the light is bad, the theater is hot, the concert hall is noisy, and so on, and we will discount the verdicts of others when we learn of inappropriate circumstances in which they made their judgment. But furthermore we may not be sure that we were in a sound state—perhaps the day's events were preying on our minds, perhaps if we revisited the matter after a good night's sleep we would feel entirely different. So we worry about whether our judgment was trustworthy. Importantly as well, the good critic must clear his mind of particular personal prejudices. A critic who cannot do this

> never sufficiently enlarges his comprehension, or forgets his interest as a friend or enemy, as a rival or commentator. By this means, his sentiments are perverted; nor have the same beauties and blemishes the same influence upon him, as if he had imposed a proper violence on his imagination, and had forgotten himself for a moment. So far his taste evidently departs from the true standard; and of consequence loses all credit and authority.[11]

Why do we have these practices? Why do we hold our reactions of the moment subject to potential failure and revision? In questions of taste, a large part of the explanation is going to be that in conversation, while we may or may not be interested in how a particular individual responded to something, it is seldom that our interest is confined to that. When someone tells us something our normal concern is what *we ourselves* are to think about it and typically this is expected to be what *we together* are to think about it. Simply by asserting that something is so, they are, as Brandom puts it, in the game of giving and asking for reasons.[12] So if, to take the first example, someone tells us that the cloth is brown we may be set to act upon what they say, and this would set us up for disappointment in one project or another if in fact (which implies in normal daylight) the cloth is not brown, but, for instance, red. Things would not turn out as we had expected or hoped they would. The informant's excuse that it looked brown to him may go some way to exonerating him, but if he adds that of course the light was not at all standard then we may reasonably feel that he should have been more careful. In short, the public expression of a reaction does not only answer to the sincerity of the informant. It is telling people what they may expect to think or what they *are to* think about something, and this purpose will be ill-served by the fact that *I*, the

10 "Of the Standard of Taste." In *Essays Moral, Political and Literary*, edited by Eugene F. Miller, 233–4. Indianapolis, IN: The Liberty Fund, Part I, Essay 23.

11 Ibid., 23; Miller, p. 238.

12 R. Brandom, *Making It Explicit*. Cambridge: Harvard University Press, 1994.

informant, felt something, if in addition it seems likely that for internal or external reasons, *I* was in a poor position to judge.

This explanation of our practice is pragmatic in spirit. It identifies coming to judgments about things as a "common pursuit."[13] This should not however be misunderstood. It might sound as if success is achieved (we score a goal in the game of giving and asking for reasons) if our verdict is accepted, for instance, by enough of our peers. But our practice goes beyond trying to come to an agreement, or trying to come to solidarity with others, to use Rorty's phrase.[14] In a world permeated by the madness of crowds, solidarity with crowds is scarcely a goal. Our aim is to have a properly grounded view, or in other words to search for truth rather than agreement. Rorty parses this as the hope of being justified not just to a present audience but also to an imagined future audience. But if so, the future audience needs to merit its status as a superior or more authoritative arbiter (it is no skin off my nose if I cannot justify myself to future persons who are ignorant or incapable). It is, of course, not easy to know how to think about superiority, authority, or truth, in aesthetic matters, since variation of taste and judgment is such a pervasive fact of life, and skepticism about the pretensions of the connoisseur is not uncommon. But even in the face of such divergence, there is the hope for opinion that can be "assimilated, validated, corroborated and verified," to use William James's phrase.[15]

Following Hume we can approach this by reflecting on the merits that make someone worth listening to on his subject. After many illustrations Hume hits on the necessary qualities for the critic whose verdicts can be assimilated, corroborated, validated, and verified:

> Strong sense, united to delicate sentiment, improved by practice, perfected by comparison, and cleared of all prejudice, can alone entitle critics to this valuable character; and the joint verdict of such, wherever they are to be found, is the true standard of taste and beauty.[16]

The rest of us, lacking these credentials, may have our feelings and indeed express them. But we stand to be corrected. If our sentiments are not delicate then we miss things others find to matter; if we are coming to some kind of work with no practice or have no suitable comparison class in our repertoire then we literally don't know what we are talking about; and if we are apparently biased by other prejudice then we forfeit authority. If we insist on asking *why* this character alone entitles the critic to being someone worth listening to, then pragmatism comes to the rescue (for it would be tossing in the sponge just to say that these virtues are indicators of aesthetic truth). The good critic can show us things we had

13 The phrase comes from T. S. Eliot, who defined literary criticism as "the common pursuit of true judgment." The phrase was later used as the title of a book by the formidable literary critic F. R. Leavis.

14 Richard Rorty, "Solidarity or Objectivity" in his collection *Objectivity, Relativism and Truth*. Cambridge: Cambridge University Press, 1990.

15 William James, *Pragmatism*. New York: Longmans, Green & Co., 1907, 197.

16 Hume, ibid., 23; Miller, p. 241.

otherwise missed, enable us to place works in their traditions, to come to understand what is more satisfying and permanently satisfying, and thereby increase our enjoyment. Virtue, here as everywhere else in Hume, denotes a quality of mind whereby a person is 'useful or agreeable to himself or others' and when a critic meets Hume's criteria, this is what we find her to be.

The genealogy of our own capacity for worrying about our own reactions is now apparent. If people act upon my having told them that something was brown, or boring, or disgusting or enjoyable and not only fail to find them so but also fail to find anything commonly supposed to put such a reaction in place, then I will be put in a metaphorical dock. I will lose status; people may start to whisper behind my back. I am deemed to be unreliable, and that is a serious criticism. I hope I am not risking it, and may revolve in my mind the chances that I am.

It is, I think, worth remarking that just as Peirce and James approach the problem of truth in general by asking about the "particular go of it," or in other words the human practices of enquiry and the satisfaction of doubt, so Hume approaches aesthetic truth by reflecting on the discriminations and practices that, in this area, constitute enquiry. We may start by being suspicious of any abstract conception of "the truth" in matters of taste, or even in ethics. But we all happily deploy the difference between those who know what they are talking about and those who do not, those who are practiced and those who are not, and those from whom we can learn and those from whom we cannot.[17]

There are also cases in which we care little about achieving a common point of view. Kant thought that this was typically so with pleasures of the palate, where there is only sensation and no real imaginative involvement.[18] I can express my delight at chocolate ice cream without caring whether you have the same taste (although even here there are shadows of a desire for solidarity. If you do not share at least some of my tastes, then we are not going to get along very well). If I go round an art gallery with my wife, the afternoon will be enjoyable enough if we find ourselves having similar reactions to paintings, and we may not care very much whether one of Hume's sensible critics would feel and think differently. There are also limits to our worries about superior audiences to whom we may be unable to justify ourselves. We can, I suppose, play with the idea of a superior musical culture that finds the work of Mozart simplistic or jejune, but it need not affect our enjoyment, nor our sense that our enjoyment is well enough grounded. The thought that our standards are *our* standards does not undermine the practices of criticism.

If all this is so with aesthetic feelings and attitudes, it is even more so when we return to the moral case, for here feelings and attitudes matter more and have more direct expressions

17 This summarizes a longer treatment in Simon Blackburn, *On Truth*. New York: Oxford University Press, 2018.

18 Immanuel Kant, *The Critique of Judgment*, trans. Werner Pluhar and Mary Gregor. Indianapolis, IN: Hackett Publishing, 1987, Book I, part 3, 47.

in action. It need not matter all that much to me if some third party goes on admiring Renoir when I find him sentimental. But it will worry me more when our plans conflict, or when the practice of the other strikes me as boorish or depraved, indecent or dishonest. Hume talked of achieving a "common point of view," so, for instance, we can abstract from our own involvement in a state of affairs and disinterestedly contemplate the features of whatever we are judging. This enables us to take up attitudes to people in history, or even fiction, where our own interests are absent. This search for a common point of view is the essential ingredient in a Humean attack on Gibbard's problem, for it is this that separates simple likes, dislikes, and preferences from the more reflective and disinterested states of mind that underlie public approval and disapproval.

> When a man denominates another his *enemy*, his *rival*, his *antagonist*, his *adversary*, he is understood to speak the language of self-love, and to express sentiments, peculiar to himself, and arising from his particular circumstances and situation. But when he bestows on any man the epithets of *vicious* or *odious* or *depraved*, he then speaks another language, and expresses sentiments, in which, he expects, all his audience are to concur with him.[19]

It was an inability to make the same transition that was a crippling disqualification in a prejudiced critic, above.[20]

It is quite possible to imagine people lacking the capacity, or more likely having it only to a limited degree. In fact, Hume himself thought that some people were primitive in this respect. Like the wanton, they would be unable to stand back and ask whether they need to discount their personal prejudices and defects in the interest of a "common pursuit." Indeed he thought it was quite difficult for any of us to manage this when we are faced with serious emotional blocks in front of doing so. Allowing that our enemy has a musical voice may be one step too many for some people, and parents are notoriously slow to admit their children's deficiencies.

I believe that once we have the idea of a common point of view, or the common pursuit of true judgment, the "right reasons" problem largely solves itself. The right reasons for a verdict are just those features of a subject matter that can be advanced in a common pursuit: the ones that do not appeal only to individual peculiarities of prejudice or motivation or that only appear in certain circumstances. We can advance a feature such as her cheerfulness for calling someone an admirable mother, because we expect all unprejudiced audiences to find such a trait useful and agreeable to the subject herself, to her children, and to others. We cannot, or should not, find the fact that a tyrant bestows wealth on those who admire her as

---

19 Hume, *Enquiry Concerning the Principles of Morals*, edited by Selby Bigge. Oxford: Oxford University Press, 9.6, 272–73.

20 There has been much discussion of exactly how Hume thought of the common or general point of view, and how it relates to Adam Smith, and indeed to subsequent utilitarianism. A seminal paper is Geoff Sayre-McCord, "On Why Hume's 'General Point of View' Isn't Ideal—and Shouldn't Be." *Social Philosophy and Policy* 11, no. 1 (Winter 1994): 202–28.

a mother, and punishments on those who do not, as a reason for supposing her admirable, since it has not located any useful or agreeable trait that she possesses. And in any case the tyrant's bribe or threat is only likely to result in lip service. If a tyrant tries to bribe or coerce us into admiring a person who is quite the reverse of admirable, we may manage to behave as if we do, but as with Pascal's wager it would take a whole sea change for us to find ourselves doing so. Similar remarks apply across the whole field of gerundively laced descriptions: there are specific features that need to be cited if we are to defend a verdict that something is exciting, boring, delicate, enjoyable, creepy, and so on.

Do naturalists need to dig deeper than this? In the case of morals, much more than that of taste, it seems fairly straightforward to sketch a story. Animals evolve social natures when they need to coordinate, in hunting or managing a territory or just practicing skills as they grow. Wherever there is coordination, there is the possibility of defection, or public nuisance, and wherever this is so there is space for aggression against such nuisance. A famous example of this is the "canid bow," which is the signal whereby a dog signifies that it is willing to play; among packs of wild dogs, such as the Western coyote, a dog that gives such a signal, but then takes advantage of its victim being off guard to attack it, is shunned and in effect expelled from the pack. We have here the first signs of convention, of the will of the collective to enforce accordance with it, and of the internalization of the criticisms we know that defection and nuisance will bring down on us. Gibbard himself has written brilliantly about the social function of guilt in the face of anticipated anger.

So if I say that some male should be ashamed because they have let their hair grow, then outside of some special context (such as it being likely to be caught in the machinery they operate), I am criticizing them unjustly. I am probably supposing that increased vigilance will discover hidden vices or flaws in them, but very likely I would be wrong about that. I will feel uncomfortable if I fear that I risk being in that position, just as I might feel uncomfortable and ashamed at having committed many another social gaffe, or even having risked doing so, for example by having made a significantly off-color remark that fortunately went unheard.

I hope that these considerations take away the threat of there being inexplicable jumps in the journey from reactions such as desire or aversion toward judgments of good or bad, right or wrong. But the other naturalistic concern is the starting point of the genealogy. Have we smuggled naturalistically suspect materials into the very starting point of the account?

I think the only way to turn this worry into a criticism would be to complain that all psychological ascriptions involve us in "the space of reasons," and hence in "normative governance." So even a starting point that makes use of our foresight, or prudence, or desires or beliefs of any kind at all is starting, as it were, too far up. This is why in his fine and unjustly neglected book *Linguistic Behaviour* Jonathan Bennett took us back to primitive goals and registrations as features of quite simple animal life that nevertheless underpin more complicated cognitive states, such as desires and beliefs.[21] Here I have no space to do more than

21 Jonathan Bennett, *Linguistic Behaviour*. Cambridge: Cambridge University Press, 1976, Chapter 2.

commend Bennett's approach, as one that brings intentionality in general into the sphere of the natural.

To finish with I shall venture a remark about reasons and the naturalistic attempt to do better than "reasons-primitivism." The basic phenomenon is our habit of saying that A is a reason for B. I hold that we say this to express an attitude (Gibbard might say, a plan) of approval for a cast of mind that taking in A is disposed to be guided toward B. The range of the relation might include almost any mental state: desires, attitudes, plans, or beliefs. The domain of the relation includes beliefs, but not only beliefs, since experiences and events can also initiate admirable movements of the mind toward a conclusion. Among the central examples of events that give people reasons for beliefs are observations. A person who has been and looked and seen the eggs in the fridge has reason for believing that there are eggs in the fridge; a person who for some unaccountable reason is deluded into thinking that he has been and looked has less reason, and in one good sense has no reason at all. In this sense reasons include causes and an unfortunate implication of Sellars's distinction between things that do, versus things that do not, belong to the space of reasons is the implication that these spaces are entirely disjoint. This makes nonsense of the very notion of observation. Making an observation is putting yourself in line for a causal impact from a state of affairs, and provided the causal impact is one that can be expected if, but only if, the state of affairs obtains, it is a very good way of acquiring a reason for believing in the state of affairs.[22]

Then the same things can be said about the attitude of accepting one thing as a reason for another as we said above about verdicts and judgments in general: finding yourself disposed to react to A by moving toward B is one thing; being prepared to commend this movement publicly is something else. When we think we have been run away with by our hopes and fears, or manipulated or seduced by the salesman, these come apart. We can be as comfortable, or uncomfortable, without own tendencies and responses in this respect as in all the others I have been describing.

So my position is that Gibbard was absolutely right to insist that there is something special about our capacity for normative governance, but had no need to fear that what was special about it also took it out of the range of a Humean psychology.

I am sure Gibbard has long got over his discomfort with long-haired youths, and I concur with him in thinking that he had no reason for this discomfort. But I also think that it was at least excusable to be infected by the censorious attitudes more commonly found in the faraway days when we were both young.

22 I expanded this point, directing it against coherentists such as Davidson, in "Pragmatism: All or Some?" In *Expressivism, Pragmatism and Representationalism*, edited by H. Price, 67–84. Cambridge: Cambridge University Press, 2013.

6

A GIBBARDIAN ACCOUNT OF (NARROW) MORAL CONCEPTS

Stephen Darwall

When philosophers talk about morality, moral reasons, and the like, they are typically employing narrower normative concepts than just any normative, or even ethical, ideas. Thus, Bernard Williams contrasts the fundamental "ethical" question, "How should one live?" with "distinctive issues of morality," such as "What is our duty?" (Williams 1985, 4).

Allan Gibbard makes a similar distinction, but between what he calls "wide" and "narrow" conceptions of "morality" (Gibbard 1990, 40–45). On a wide conception, "moral" is synonymous with "practically normative," like Williams's "ethical." It refers to any normative question about how to act; "morality is simply practical rationality in the fullest sense": what it "makes sense," or what one ought or has most reason to do (Gibbard 1990, 40). "On the narrow conception," however, "moral considerations are just some of the considerations that bear on what it makes sense to do" (i.e., on widely moral action).

Like Williams, Gibbard identifies narrow morality with specifically *deontic* matters of right and wrong (Gibbard 1990, 41). And he follows Mill in connecting these conceptually to the aptness of distinctive attitudes that Mill associates with "punishment." "We do not call anything wrong," Mill says, "unless we mean to imply that a person ought to be punished in some way or other for doing it; if not by law, by the opinion of his fellow-creatures; if not by opinion, by the reproaches of his own conscience" (Mill 1998, Ch. V). Gibbard's Millian view identifies wrongness, and thereby narrow morality, through its connections to the attitudes of "guilt," "resentment," and "blame" (Gibbard 1990, 41–42). "What a person does is morally wrong," he initially proposes, "if and only if it is rational for him to feel guilty for doing it, and for others to resent [or blame] him for doing it" (Gibbard 1990, 42). (I add "or blame"

because resentment is only appropriate from victims or vicariously from those who identify with them; blame, by contrast, is an attitude that is apt for any or all "others," including the agent himself (guilt or self-blame) (Darwall 2013a, 32–37).)[1]

Gibbard recognizes that this ties wrongness to blameworthiness too tightly. An action may be wrong though not blameworthy if the agent has an excuse. So, he says that "standards for wrongness . . . are standards such that an agent is *prima facie* blameworthy if he does not use them to rule out acts that violate them" (Gibbard 1990, 45). I try to get at the same thing when I claim that it is a conceptual truth that if something is morally obligatory (all things considered), then it is an act of a kind it would be blameworthy to omit *without excuse* (e.g., Darwall 2016, 266).

Gibbard notes that the strategy of tying narrow (deontic) moral concepts to blameworthiness and of understanding blameworthiness in terms of a fundamental normative concept of what there is reason, it is fitting, or one ought, to blame was initially proposed by A. C. Ewing (Ewing 1939, 14). It is part of a more general program of understanding all normative concepts in terms of one fundamental normative concept—ought, fittingness, or what is justified or warranted—combined with distinctive attitudes to which the normative concept is conceptually related, in this case, moral blame.

The basic idea is that normative judgments are all "fraught with ought," in Sellars's famous phrase (Gibbard 2003, 21; Sellars 1962, 44). They all contain the same fundamental normative concept, to refer to which different philosophers may use different terms. Sidgwick says that ethical judgments all contain "the fundamental notion represented by the word ought" (Sidgwick 1967, 25). Ewing tends to prefer the term "fitting," and Gibbard generally uses "rational," "makes sense," or "warrant." "Reasons firsters," like Parfit and Scanlon, express what I take to be the same idea in terms of *normative reasons*, that is, reasons in favor of or *to* have some attitude or other (Parfit 2011; Scanlon 1998, 2014).[2]

Different normative concepts are then composed by combining this basic normative notion with the attitude to which the more specific normative concept is conceptually tied. Thus, the desirable is what there is reason to desire, the estimable, what one ought to esteem, dignity, what is fittingly respected (in the recognition sense), and so on.[3]

---

1 We could also add the prospective attitude that Strawson calls "sense of obligation" and that Howard Nye calls "prospective guilt-tinged aversion" (Nye 2009; Strawson 1968, 86). See also Pufendorf on "antecedent obligation" (Pufendorf 1934, 41).

2 In *Principia Ethica*, G. E. Moore famously argued that *good* is the basic ethical concept (Moore 1993). For arguments that Sidgwick rather than Moore was right on this fundamental point, see Frankena (1942) and Darwall (2003).

3 On the difference between recognition and appraisal respect, see Darwall (1977). More carefully, normative reasons must be restricted to "reasons of the right kind." Like Parfit, however, Gibbard tends to deny that what other philosophers call "reasons of the wrong kind" for attitudes actually are genuine normative reasons for those attitudes. He denies, for example, that pragmatic considerations can be reasons for belief (which must concern evidence), holding that these can only be reasons to desire to have some belief.

Of course, normative concepts are not restricted to the ethical or practically normative. We can also ask what we have reason (or ought) to believe, that is, what is *credible*? If we follow Gibbard, we can achieve a unification of all normative concepts by understanding them all in terms of a fundamental normative concept, whichever terms we might choose to use to refer to it, along with the different attitudes that different normative concepts distinctively involve. For example, we can understand evaluative concepts as involving the basic normative concept and evaluative attitudes, thereby distinguishing them from other normative concepts, such as deontic or epistemic concepts, by their distinctive attitudes. And we can distinguish *within* evaluative concepts by distinguishing different evaluative attitudes, desire and esteem, for example, that are distinctively involved in the concepts of the desirable and the estimable, respectively.

I follow Gibbard also in holding that deontic (narrow) moral ideas are conceptually connected to blameworthiness. This is one of two planks that underlie my claim in *The Second-Person Standpoint* and later work that deontic moral ideas are all second-personal notions (Darwall 2006, 2013a, 2013b). The other is the Strawsonian idea that reactive attitudes like moral blame are held from an implicitly interpersonal (I say "second-personal") perspective of implicit relating *to* their objects. A distinctive feature of blame as an attitude is that it implicitly addresses a putatively legitimate demand *to* its object (second personally) for them to hold themselves accountable for violating a legitimate demand (Strawson 1968, 85). In so doing, it presupposes the authority to make the demand and demands acknowledgment of this authority as well. Blame, like every reactive attitude, comes with an RSVP.

Blameworthiness conjoins two notions, the idea of blame and the fundamental normative idea, whichever term we choose to express it. By connecting deontic (narrow) moral ideas to blameworthiness, the approach Gibbard and I favor therefore ties them to the notion of what would be justifiably blamed. If we can use 'normative reason' to refer to the fundamental normative notion, then this approach is compatible with a "reasons first" account (Parfit 2011; Scanlon 1998, 2014). What is wrong is what there is normative reason to blame, lacking excuse.

As will become clearer later, however, deontic moral ideas like obligation, duty, right, and wrong, although applying to actions, cannot themselves be understood in terms of *reasons for action*, not even moral reasons. That does not mean that they can be understood independently of the idea of a normative reason though, if that is the fundamental normative idea. To the contrary, Gibbard's and my analysis of deontic (narrow) moral ideas makes ineliminable use of the fundamental notion of normative reason, only not normative reason *for action* but reasons for the holding-accountable attitude of moral blame. This is where the distinctive normative work of narrow morality is done. In determining what our moral duties and hence wrongful conduct consist in we are in effect determining what kinds of acts we have normative reason to blame if these acts are done without excuse.

Deontic moral concepts are not, however, the only kind of moral concepts there are. We appraise (completed) actions, motives, and character as morally good or bad. We speak of moral agents or persons as having a special kind of moral value or dignity. And beings of

other kinds are also said to have a kind of "moral status" in being "morally considerable," as giving us moral reason to act toward them in certain ways. The idea of moral reason seems also not itself to be itself a deontic (narrow) moral notion. It seems conceptually, if not indeed actually, possible, for an action to be recommended by moral reasons, to be morally *choiceworthy*, without being morally required.

How, then, are these ideas to be understood in relation to narrow morality? They do not seem comfortably to fit Gibbard's category since they do not seem to have a conceptual connection to blameworthiness, at least, not in any obvious way. In what follows, however, I shall be pursuing the possibility that all moral ideas *do* have a connection to blameworthiness, however unobvious it may seem to be.

Moral Reasons

Start first with the idea of a moral reason and note, to begin, that there is no single such idea. We can distinguish, at least, moral reasons for *action*, for *blame* (blameworthiness), and for moral *esteem or disesteem* (morally good and bad character, motive, and action). Take, then, the idea of a moral reason for acting. How are we to understand this idea?

It is helpful to start by showing how the narrow ideas of moral obligation, duty, right, and wrong cannot themselves be understood in terms of moral reasons *for action*. Since we often express someone's being morally obligated to do something by saying that this is what they "morally ought" to do, and since, as I said above, the basic normative idea can be expressed indifferently by "normative reason" or by "ought," this can seem a puzzling idea. If what there is normative reason for someone to do is what they ought to do, then why isn't what someone morally ought to do what they have most moral reason to do? There is no problem granting a sense of "moral ought," the *practical moral reasons sense*, as we might call it, in which this is true. Gibbard and I are committed to holding, however, that there is a different "narrow" or *fully deontic* sense of the moral ought that differs from such a moral reasons sense.[4]

To see this, consider the following "open question" (or as Gibbard puts it, "what is at issue?") argument (Gibbard 2003, 23–29; Moore 1993, 72). Imagine a case in which an agent is able to bring about a great good for others but at *very* great personal cost. And consider a disagreement about such a case between an orthodox act consequentialist and someone who accepts agent-relative "prerogatives" or permissions of the kind Scheffler argues for in *The Rejection of Consequentialism* (Scheffler 1982). Let us suppose that they agree that it would be morally best, that is, most recommended by practical moral reasons or most morally choiceworthy in this case for the agent to bring about the greater impartial good despite the significant personal cost. They disagree, however, about whether this would be

4 Perhaps it would be cleaner to reserve "morally ought" for the moral reasons sense and use "morally *must*" for the fully deontic sense. Nothing of substance hangs on this semantic choice, however. I am indebted here to discussion with Paul McNamara.

morally obligatory or required. Since it would bring about the greatest impartial good, the act consequentialist holds it would indeed be morally required. The Schefflerian denies this. They hold that the personal cost is sufficiently high that an agent-relative prerogative or permission makes it permissible for the agent to forego promoting the greater impartial good at that level of agential cost (see also Harman 2016).

Without prejudice as to which would be correct, it seems obvious that such a disagreement is conceptually possible. Since, however, they agree that bringing about the greater impartial good would be most morally choiceworthy, in disagreeing about whether this act is morally required, they both must implicitly hold that there is a conceptual difference between moral choiceworthiness, being best supported by moral reasons for action, on the one hand, and the deontic (narrow) idea of moral obligation, on the other. What they would be disagreeing about, on Gibbard's (and my) analysis, is whether, given the agent cost, it would be blameworthy for someone to refrain from the morally best act were they to lack excuse. Here the agent cost would be functioning not as an excuse but as a *justification*. It would show not that her action was wrong but not culpable. It would show that her action was not actually wrong, not a violation of an all things considered moral obligation.

We can make the same point by saying that when philosophers disagree about whether there can be such a thing as moral supererogation, actions above and beyond the call of moral duty, theirs is a normative rather than a conceptual disagreement. It seems obvious that someone who holds (with common sense!) that acts can be supererogatory is neither conceptually confused nor contradicting themselves.

Again, although deontic (narrow) moral concepts cannot be understood in terms of moral reasons for action, that does not mean that they cannot be understood in terms of normative reasons *for some attitude or other*. I take Gibbard to agree that they can be, only that the normative reasons are, in the first instance anyway, reasons for holding-accountable attitudes like moral blame rather than reasons for action. It is a conceptual truth that violations of moral obligations entail reasons that justify resentment (from the victim's perspective), guilt (from the agent's perspective), and blame (from anyone's perspective).

When I say that blame is warranted from the standpoint of any moral agent (or the moral community), I am referring to the *attitude* of blame and not to any speech (or other public) act that might be taken to express that attitude. Whether someone has standing to blame in these latter senses can depend on normative issues of standing that can be defeated by, for example, hypocrisy, entrapment, and so on. What matters is whether blame is a *fitting* response, not whether someone has justification for expressing it, or even, indeed for feeling it that relates not to whether the action was indeed culpable but to some reason for foregoing blame of some other kind (say, that it is inconsistent with something like *agape* love).

I will take it that normative reasons for blame, which can support conceptually related deontic (narrow) moral judgments of right and wrong, are a kind of moral reason. Someone might wonder whether there can be reasons for blame that are not moral, whether the range of the culpable extends beyond the realm of the moral. I cannot see, however, that there is

anything more than a semantic question here of how to use "moral." If a genuine issue of culpability is involved, then we must be in the range of genuinely normative, *de jure* obligation, and, on the view Gibbard and I favor, that just is the realm of narrow deontic morality.

Now, however, we face a puzzle. I have said that deontic (narrow) moral notions are normative for holding-accountable attitudes like blame, and not for action, in the first instance. But what deontic judgments *apply to* and demand is *actions*. So how can it be that they concern, not reasons for action, but reasons for blame? The answer is that moral obligations *do* conceptually implicate reasons for acting, both moral reasons and normative reasons period, but that they do so precisely because of their connection to warranted blame and presuppositions of blame itself (Darwall 2016).

To see this, ask yourself whether you can coherently have the attitude of blame toward someone (either yourself or someone else) and simultaneously accept that they had sufficient normative reason for acting as they did. I take it you cannot. I am not saying that it is impossible for someone to be in these simultaneous states of mind, but that the states of mind are themselves in conflict; their combination is incoherent in something like the way that inconsistent beliefs are. In blaming someone, you are implicitly making a putatively legitimate demand of them and implying that there cannot be sufficient reason for them to violate it. Any reason that would justify violation would constitute a *justification*, which would show that the action was not indeed something it would be blameworthy to do without excuse; it would defeat the claim that what one did was wrong, something one was obligated, all things considered, not to do.[5] Blame can be warranted only on the assumption that no such justification exists. Since moral obligation entails that an action is of a kind there would be reason to blame the agent for, absent excuse, it must also follow that if an action is obligatory, all things considered, then there is conclusive normative reason to do it.

But what, then, are moral reasons for action? If narrow moral concepts are deontic notions that conceptually implicate accountability and blameworthiness, and the concept of a moral reason differs from these in the ways we have just noted, how can moral reasons for action to be understood in relation to narrow morality, as Gibbard understands it?

Here we might take a step back and consider the question of what makes a reason for acting a moral reason. What hangs on this question?[6] We can of course stipulatively define "moral reason" in any way we like, say, by other-regarding content or in some other way, but nothing of normative interest can follow from that. And some ways of defining moral reasons, from example, that moral reasons are reasons *from the moral point of view*, may actually put the normative force of moral reasons for acting at risk (see, e.g., Baier 1957; Frankena 1977, 212).

5 This does not mean that whether something is morally wrong depends on whether there is, independently, normative reason, all things considered, not to do it. The fact that it is wrong can itself be a (further) reason (Darwall 2013a, 52–72).

6 What follows draws from Darwall (2017).

Suppose we say, for example, that the moral point of view is that of equal concern and respect, and therefore that moral reasons for acting are practical reasons from that point of view. Philippa Foot famously pointed out, however, that it is consistent with reasons having force *from a point of view* that they have no genuine (perspective-independent) normative weight at all (Foot 1972). In fact, Foot makes this point about moral reasons in particular. Although she grants that moral oughts are "categorical imperatives" in the sense of applying to and having weight from the point of view of morality irrespective the agent's ends and interests, the same is true, she says, of etiquette. But she notes that if it is consistent with the fact that, say, there is reason *from the point of view of etiquette* to answer in the third person an invitation addressed in the third person, that there might *actually*, and not just from this point of view, be *no normative reason* whatsoever to do this, *and* that the same might hold of morality. Identifying moral reasons by their relation to the moral point of view is consistent with their having no (perspective-independent) weight at all.

This problem seems especially pressing with respect to moral reasons, since, unlike reasons of etiquette, moral reasons have generally been thought to be especially weighty. I argue, however, that an approach to narrow morality like Gibbard's can provide an account of moral reasons on which whether a reason is a moral reason is a question with genuine normative upshot (Darwall 2017). The idea would be that although there is a conceptual difference, as we have noted, between an action's being morally obligatory and its being recommended by moral reasons, what makes something a moral reason for acting is that it is a *pro-tanto moral obligation-making consideration*, a fact that *tends* to make an action morally obligatory, all things considered.

Think about the kind of case we considered before in connection with a disagreement between an act consequentialist and someone accepting a Schefflerian agent-relative prerogative or permission. This was a case in which an agent could realize significant impartial good, but at a personal cost sufficient for an agent-relative prerogative to make it permissible for them not to do so. Let us then view this case as the Schefflerian would analyze it. The impartial good remains a pro tanto obligation-making consideration, one that still tends to make taking the action morally obligatory, even though, given the personal cost, it does not make producing it morally obligatory, all things considered. On the theory I am proposing, then, the fact that one could produce the impartial good remains a moral reason in favor of taking action.

Moreover, as a *moral* reason, on the current proposal, it outweighs the personal cost so that what the agent has most *moral reason* to do, all things considered, is to produce the impartial good even at the personal cost. In a case of this kind, as I am imagining it, the personal cost defeats the obligation to absorb it without creating a duty not to do so. The personal cost functions not so much as a moral reason to avoid the personal cost as it does a consideration that justifies doing so in the fully deontic sense of showing that avoiding the cost is permissible and not morally wrong.

This leaves the *pro tanto* obligation to produce impartial good (assuming there is one) unaffected. It remains the case that the agent is morally obligated to produce it, *other things*

being equal, though not, all things considered. On the theory I am proposing, therefore, it does not affect the weight of the moral reason to produce it. Crucially, since the personal cost does not override the impartial good as a pro tanto obligation-making consideration, it does not outweigh the impartial good as a *moral reason for acting*.

It might be objected, however, that being a pro tanto obligation-making consideration cannot explain why some fact—in this case, impartial good—makes something a moral reason to act when the obligation is defeated. After all, producing the impartial good ends up not being morally obligatory, all things considered. How can the fact that the act *would* be obligatory, all things considered, were it not for the personal cost, give one a moral reason to produce the impartial good if the personal cost is sufficient to defeat the moral obligation, all things considered?

But facts can defeat moral obligations in different ways. Some do so without affecting other facts' tendency to make an action obligatory, and so, on the proposed theory, without affecting the moral reasons for acting those facts create. Other facts, however, can show that a fact that otherwise might have had such a tendency does not actually have it. They defeat not just the all-things-considered moral obligation but also the pro tanto one and therefore, on the proposed theory, any moral reason that might otherwise have been created.

For example, suppose that I promise to bring you a book, but that you later release me from my promise because you no longer need or want it. This not only defeats any all things considered obligation to give you the book as promised but also any pro tanto obligation that might otherwise be in play owing to the promise. This is different from the kind of case we have been considering where the pro tanto obligation-making force of considerations of impartial good remains. And this explains the clear sense in which the impartial good remains as a moral reason to produce it, even at a personal cost that can defeat any all-things-considered obligation to do so, and why, in the case just described, the fact that I promised to give you the book no longer remains as any moral reason for me to do so.

This theory explains why, in the case involving impartial good and personal cost, it is morally *better*, more morally choiceworthy, to produce the impartial good at such a personal cost, although it would not be morally obligatory to produce it (or wrong to decline to do so). It explains why such an action would be morally supererogatory.[7]

Despite appearances, therefore, a Gibbardian approach to narrow morality, though having to deny that fundamentally deontic moral ideas can be understood in terms of moral reasons, can nonetheless account for moral reasons for action as a narrow moral idea, since moral reasons can be understood as pro tanto moral obligation-making considerations.

There is the further question of whether the fact that an action is morally obligatory or morally wrong can *itself* be a reason, and if so, a moral reason, to comply with the obligation or avoid the wrongful action. If the only moral ought were the ought of moral reasons, and if

7 Of course, this latter is a substantive normative moral claim. I am using it here just to illustrate a feature of the conceptual proposal that moral reasons be understood as moral obligation-making considerations.

there were no independent fully deontic notion of moral obligation, then the answer would seem to have to be "no." Just as a "buck-passing" view of value, on which to be valuable just is for there to be reason to value something, denies that the fact that something is valuable is *itself* a reason for valuing that thing, so also would the fact that one morally ought to do something not itself be a reason to do it (Scanlon 1998, 95–100). That one moral ought to do something would *consist* in there being moral reason to do it, so it could not itself provide a (further) practical moral reason.

It is crucial to a Gibbardian conception of narrow morality, however, that the fully deontic idea of moral obligation differs from the moral ought of moral reasons by virtue of its conceptual connection to accountability and blameworthiness. And because that is so, I have argued elsewhere, the fact of moral obligation *can* and *does* provide a further reason for acting as one is obligated beyond the reasons that would make omitting it something that would be blameworthy lacking excuse (Darwall 2010). A Gibbardian should be a buck-passer about blameworthiness; the fact that some action would be blameworthy is not itself a reason to blame someone for performing it, since it consists in all and only the reasons there are for such blame. But that does not dictate buck-passing about moral obligation and wrong *with respect to reasons for acting*. To the contrary, it actually supports the thought that the fact that an action is morally obligatory is itself a reason for acting.[8]

So, the concept of a moral reason can be plausibly understood as a narrow moral concept. Gibbardians will insist that narrow moral concepts like moral obligation and wrong must be understood, not in terms of reasons for action, even practical moral reasons, but through their conceptual connection to justified holding-accountable attitudes like moral blame. But that does not mean that the concept of a practical moral reason is not a narrow moral concept, since moral reasons can themselves be understood as moral obligation-making considerations.

Kinds of Moral Good

Once we have placed the concept of a moral reason for acting within a Gibbardian narrow moral framework, we can define the concept of the morally choiceworthy, what it is morally good, or best, *to do*, in terms of it as what there is most moral reason to do. Putting aside issues about moral dilemmas, suppose that there is an action in the circumstances that is morally obligatory, all things considered. I have already argued that we never have sufficient normative reason not to act as we are morally obligated, all things considered. So it should be uncontroversial that we never have sufficient moral reason not to do so. All

8 But is it a *moral* reason for acting? If moral reasons are *pro tanto* moral obligation-making considerations, it would seem not, since the fact that an action is morally obligatory is not itself moral obligation making. To handle this case, we can simply add that moral reasons are either moral obligation-making considerations or the fact that an act is morally obligatory itself.

things considered, moral obligations therefore bring conclusive moral reason and so moral choiceworthiness in their train.

But an action can be most recommended by moral reasons without being morally required; at least, it can if there is such a thing as supererogation. And in this case, the action is, though not a moral duty, nonetheless, most morally choiceworthy, or morally best to do.

Being a morally good thing *to do* is only one thing "morally good" can mean, however. What is morally choiceworthy is an action *type*, a kind of action to choose in the circumstances. However, action *tokens*, that is, completed (and hence, motivated) actions, can be appraised as morally good or virtuous also.[9] What seems to be basic to this kind of moral evaluation is motivation. As Hume put it, "All virtuous actions derive their merit only from virtuous motives" (1978, 478). We take morally good motivation as an expression of an agent's virtuous or morally good *character*, he says, when these virtuous motives come from "durable principles of the mind, which extend over the whole conduct" (Hume 1978, 575).

How can moral goodness of the kind realized in motive and character be accounted for on a Gibbardian conception of narrow morality? However difficult this may seem, it is actually surprisingly easy once we have the idea of a moral reason for action as a narrow moral concept. Consider, first, the question of what motives are morally good. This would appear to be no different from the question, what are the reasons, action on which is morally good? And what could these be other than *moral reasons for acting*?

An agent's motive can either be the consideration they themselves take as a reason for acting, *their* reason for acting, or it can mean the psychological state of their so taking and acting on it. Once we have delimited the set of moral reasons for acting (again, as pro tanto moral obligation-making considerations), we would appear to have all we need to answer the question of which motives are morally good in the sense that an agent's being moved by them (in *either* of these senses) makes the action token they thereby complete morally good.

We can derive a notion of morally bad motivation and action similarly. Morally bad motives involve an agent being motivated to do something by a consideration that constitutes a moral reason not so to act.[10] A moral reason not to do something is a consideration that makes it pro tanto obligatory not to do that thing, hence wrong to do it. Being moved by such reasons is morally bad, therefore.

Note that we can take this view whatever position we take on the vexed question of whether morally good motivation is *de dicto* or *de re*.[11] This is a question that has divided ethical philosophers since at least the eighteenth century. Francis Hutcheson is an excellent example of a philosopher who holds that morally good action is motivated by morally good motives *de re*—as he sees it, by forms of benevolence—without any thought of their moral goodness

9 W. D. Ross uses "act" and 'action' to mark this distinction (Ross 2002, 7).

10 I am indebted here to discussion with Allan Gibbard.

11 Michael Smith introduces the issue in these terms in *The Moral Problem* (Smith 1994, 73–76).

(Hutcheson 2004).[12] Examples on the other side are Butler and Kant. Butler holds that morally good action must always be governed by conscience, and Kant famously maintains that actions have "moral worth" only if they are done "from duty" (Butler 2017; Kant 1996). Whether these come to the same thing or not, they are both clearly instances of *de dicto* moral motivation.

Whether we take either view or some combination, we will end up with an account of morally good motivation that can be linked to our account of moral reasons for acting. On a Hutchesonian view, morally good motivation involves being moved by the considerations that are moral reasons independently of their being moral reasons. And on a Butler/Kant view, morally good motivation must involve an explicitly moral thought, whether conscious or unconscious, going beyond these—that these *are* moral reasons (for Kant, quite explicitly that they tend to make action morally obligatory).[13]

It is fairly easy, therefore, to provide a Gibbardian account of moral goodness in the sense of morally good motivation once we have on hand the proposed "narrow" account of moral reasons for acting. And we can easily extend this to provide an account of morally good or virtuous *character* and *agents* in the sense in which morally good actions (tokens) are supposed to reflect, or reflect on, these. Morally good and bad character traits will then be dispositions to be moved by the requisite moral reasons, and morally good agents, those who have these traits.

However, an issue now arises. Judging motives, traits, and agents to be morally good is presumably conceptually related to some form of the attitude of *esteem*. There seems to be no difference between saying these things are morally good and saying that they are morally estimable. On a Gibbardian account, what is estimable is what there is reason (of the right kind) or it is fitting or warranted to esteem. This follows from the general Gibbardian program of understanding normative notions in terms of the fundamental normative notion of normative reason or warrant and the distinctive attitude that is conceptually implicated in the normative notion in question.

Francis Hutcheson thought there is a distinctive attitude of *moral* esteem, which he called "approbation." He begins his *Inquiry Concerning Moral Good and Evil* as follows: "The word 'moral goodness,' in this treatise, denotes our idea of some quality apprehended in actions, which procures approbation" (Hutcheson 2004, 85). Against Hutcheson, Hume famously contended that any distinction between "moral virtues" and estimable features of other sorts is merely "verbal" (Hume 1978, 3.3.4.1). There is no fundamental *attitudinal* difference between what Hutcheson calls "approbation" (moral esteem) and esteem in general (which Hume often also calls "approbation"). Someone may be estimable for their fine trombone

12 For discussion, see Darwall (1995, 231–33).

13 For contemporary "reasons-responsiveness" (*de re*) accounts of morally good motives, see, for example, Arpaly (2004) and Markovits (2010). For a recent defense of the idea that morally motivation can include the *de dicto* desire to do what is right, see, for example, Carbonell (2013).

playing or for their fair and benevolent character, and calling the former a "natural ability" and the latter a moral virtue marks only a "verbal" distinction. So Hume argued.

Suppose we say that something is morally estimable when the reasons for esteeming them are moral reasons, in a way that reasons for esteeming a trombone virtuoso are not. That seems fine, but it lands us back with our question, what makes a reason for esteeming something or someone a moral reason for esteem, much like our earlier question, what makes a reason for acting a moral reason for acting. But if we have answered the latter question, will we not also have our answer to the former? Moral esteem is for how someone functions *as a moral agent*.[14] Morally estimable motives just are those we think an agent (morally) should be moved by. Morally estimable character traits are virtues on account of the deliberative light they shed, that is, what the person having them is disposed to construe as reasons to act.[15] And so on.

Once we have an account of moral reasons for action, then an account of morally estimable motives, traits, character, and agents would seem to follow in its train. It might be objected that there are morally good emotions, perceptions, and ways of seeing things that are not agential in the way that motives are. It is arguably morally bad to be amused, even involuntarily, by a racist joke, even if that amusement would be unlikely to have any motivational or behavioral upshot.

However tenuous the relation morally bad involuntary states have to motivation and action though, that does not mean that what makes them morally bad is independent of moral reasons for or against action. A racist joke that denigrates and belittles, for example, depicts its object as appropriately treated in ways there are conclusive moral reasons not to treat people. Even if the response is not itself agential, that does not mean that its moral badness is unrelated to moral reasons for action.[16]

Despite this, it might be argued that what makes the belittling nonagential state morally bad is that it depicts its object as having *less value* and only because of *that* as being appropriately treated in morally bad ways, ways there are moral reasons against. Similarly, it could be claimed that conceptions of the moral value of persons or other sentient beings are not themselves views about moral reasons for action. We think that it is *because* of their distinctive moral value that other animals make moral demands on us or are worthy of our consideration. Their having this value cannot therefore be the very same thing as the reasons

14 See, however, Vranas (2005) for reasons for skepticism about overall character evaluations.

15 Hutcheson starts with an account of morally good motives and derives his account of moral choiceworthiness (and hence moral reasons to act) from that. Beginning with the idea that all morally good motives are instances of benevolence (desires for people's welfare), he concludes that actions are choiceworthy on the grounds of their contribution to welfare and that the most choiceworthy is the one "which procures the greatest happiness for the greatest numbers" (Hutcheson 2004, 125). All that matters for our purposes is that Hutcheson's approach agrees with that suggested here in holding that an account of morally good motives and moral reasons for acting must track together.

16 I have been helped here by discussion with Mario Attie.

for treatment they ground. Similarly, the idea that human or other rational beings have an inviolable dignity seems to be distinct from and to ground or explain our moral duties with respect to them.

Be that as it may, it is hard to see what *moral* value could possibly be if it is not intrinsically relevant to what moral agents have reason to *do*. Moral value seems importantly different in this way from, for example, aesthetic value. Different kinds of value manifest their distinctive natures in the different ways in which they are appropriately appreciated.

For example, aesthetic values show themselves through aesthetic contemplation. This is very different from the moral values that persons or other sentient beings might have. Even if the relevant attitudes here include forms of appreciation, like what Iris Murdoch calls "loving attention," that are not themselves agential; nonetheless, these seem intrinsically related to registering moral reasons for action (Murdoch 2014, 34, 36).[17] Similarly, something would not count as care or benevolent concern, for example, unless it had a motivational component of taking some being's good or welfare as giving one reason to act.

Or consider the idea of the dignity of persons. Even if human dignity is distinct from and explanatory of the moral demands of action it places upon us, it would seem that there is no way of understanding the former independently of its power to ground the latter. The appropriate appreciative response to dignity is (recognition) respect (Darwall 1977). And being disposed to regulate conduct by the demands a person's dignity places upon us is part of what it is to respect it.

Perhaps the most challenging ethical idea to locate in a Gibbardian framework is that of a morally good or desirable *outcome*. However, the situation here is structurally the same as the one we just considered. Just as Kantians take the dignity of persons to be a fundamental ground of moral reasons for acting, so also do consequentialists take the goodness of outcomes or states of affairs in what Parfit calls the "impartial-reasons-implying sense" to ground claims about what we have moral reason to do (Parfit 2011, v. I, 168). Just as dignity is taken to be distinct from the duties and reasons it grounds, so also is the goodness of outcomes taken to be different from the reasons for action it grounds. Even so, the connection is close enough that we can generate a Gibbardian account of the relevant value or good from an account of moral reasons. What makes dignity a *moral* value is that it grounds the relevant moral duties and practical moral reasons. Similarly, what makes, say, impersonal or impartial good a morally good outcome is that it grounds moral reasons to aim at it.

It is worth noting that the idea of an impersonally or impartially good outcome is not an exclusively moral idea. Suffering would arguably be impersonally or impartially bad, something we have reason to prevent, even if there were no such thing as morality or distinctively

17 In "Iris Murdoch and the Domain of the Moral," Lawrence Blum, in the course of an argument that loving attention should be seen as a *moral* attitude, says that it should be seen as a "legitimate source of reasons for action" (The context makes clear that by "reasons," he means "moral reasons," reasons that are, as he says, in "the domain of the moral" (Blum 1986, 345).)

moral evaluation of actions and agents. In Gibbard's terms, impersonal good would be intrinsically relevant to morality in the wide sense regardless of narrow morality. Parfit discusses a view he calls "impartial-reason act consequentialism," which he purposely characterizes in nondeontic terms: "What we have the strongest impartial reasons to do is whatever would make things go in the way in which we all have the strongest impartial reasons to want things to go" (Parfit 2011, v. I, 168). This is different from act-consequentialism as a theory of morally right action, which holds that it is morally wrong to do anything other than whatever would bring about the best (most impartially desirable) outcomes. About consequentialism, formulated in nondeontic terms, Parfit writes that it "may be better regarded, not as a moral view, but as being, like Rational Egoism, an external rival to morality" (Parfit 2011, v. I, 168). The idea that impersonally good outcomes are also *morally* good therefore requires some further connection to moral concepts than is given simply by their being impersonally good.[18] It is hard to see what else this could be other than that their impersonal goodness constitutes a moral reason to bring them about.

Conclusion

We began with Gibbard's distinction between wide and narrow conceptions of morality and with his idea that narrow morality is defined by the deontic moral notions of duty, obligation, right, and wrong that are conceptually related to actions that are apt objects of blame and guilt. As I prefer to formulate the idea: Necessarily, an action is morally obligatory if and only if it is a kind of action in the circumstances that would be blameworthy to omit without excuse. Since not all moral concepts are deontic, the challenge arises how to account for the full range of distinctively moral ideas within a Gibbardian narrow moral framework. As we have seen, the key to meeting this challenge is seeing how the idea of a moral reason for acting, though not itself a deontic idea, can nonetheless be understood in deontic terms as a pro tanto moral obligation-making consideration. Once this account of moral reason, and deriving from it, an account of moral choiceworthiness, is in place, it can be relatively easily extended to define the notions of morally good, or estimable, motive, character, and agent, in terms of dispositions to respond to moral reasons (whether or not these include deontic moral facts themselves).

We have seen also how ideas that are taken themselves to ground moral duties and reasons, whether of the value that persons and other sentient beings themselves have or of the impersonal goodness of outcomes, can also be conceived within a Gibbardian narrow moral framework. The latter is perhaps the most noteworthy. Since the idea of impersonal goodness is not itself an essentially moral idea, we can form the thought that an impersonally good outcome would be a *morally* good thing to happen only by relating it to essentially

18 Even "scalar" views of *moral* choiceworthiness go beyond the idea of impartial good and its relation to choiceworthiness *simpliciter* (e.g., Norcross 2006).

moral ideas in some way. The natural way to do this is to hold that its moral goodness consists in its being something there is moral reason to want and bring about, where the idea of a moral reason is itself understood in terms of its relation to the deontic moral standards that are the hallmark of narrow morality. In this way, even the idea of a morally good outcome can be understood as a narrow moral concept by virtue of its connection to justified holding-accountable attitudes like moral blame.[19]

References

Arpaly, Nomy (2004). *Unprincipled Virtue: An Inquiry into Moral Agency*. Oxford: Oxford University Press.

Baier, Kurt (1958). *The Moral Point of View: A Rational Basis for Ethics*. Ithaca, NY: Cornell University Press.

Blum, Lawrence (1986). "Iris Murdoch and the Domain of the Moral." *Philosophical Studies* 50: 343–67.

Butler, Joseph (2017). *Fifteen Sermons and Other Writings on Ethics*, edited by David McNaughton. Oxford: Oxford University Press.

Carbonell, Vanessa (2013). "*De dicto* Desires and Morality as Fetish." *Philosophical Studies* 163: 459–77.

Darwall, Stephen (1977). "Two Kinds of Respect." *Ethics* 88: 36–49.

——— (1995). *The British Moralists and the Internal "Ought": 1640–1740*. Cambridge: Cambridge University Press.

——— (2003). "Moore, Normativity, and Intrinsic Value." *Ethics* 113: 468–89.

——— (2006). *The Second-Person Standpoint: Morality, Respect, and Accountability*. Cambridge, MA: Harvard University Press.

——— (2010). "But It Would Be Wrong." *Social Philosophy and Policy* 27 (2010): 135–57.

——— (2013a). *Morality, Authority, and Law: Essays in Second-Personal Ethics I*. Oxford: Oxford University Press.

——— (2013b). *Honor, History, and Relationship: Essays in Second-Personal Ethics II*. Oxford: Oxford University Press.

——— (2016). "Making the Hard Problem of Moral Normativity Easier." In *Weighing Reasons*, edited by Errol Lord and Barry Maguire. Oxford: Oxford University Press.

——— (2017). "What Are Moral Reasons?," Amherst Lecture in Philosophy. http://www.amherstlecture.org/darwall2017/darwall2017_ALP.pdf.

Ewing, A. C. (1939). "A Suggested Non-Naturalistic Analysis of Good." *Mind* 48: 1–22.

Foot, Philippa (1972). "Morality as a System of Hypothetical Imperatives." *The Philosophical Review* 81: 305–16.

Frankena, William (1942). "Obligation and Value in the Ethics of G. E. Moore." In *The Philosophy of G. E. Moore*, edited by P. A. Schilpp, 93–110. LaSalle, IL: Open Court.

19 I am indebted to David Plunkett for helpful comments on an earlier draft.

——— (1977). *Perspectives on Morality: Essays*, edited by Kenneth Goodpaster. Notre Dame, IN: University of Notre Dame Press.

Gibbard, Allan (1990). *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

——— (2003). *Thinking How to Live*. Cambridge, MA: Harvard University Press.

Harman, Elizabeth (2016). "Morally Permissible Moral Mistakes." *Ethics* 126: 366–93.

Hume, David (1978). *A Treatise of Human Nature*, 2nd ed., edited by L. A. Selby-Bigge, with rev. P. H. Nidditch. Oxford: Oxford University Press.

Hutcheson, Frances (2004). *An Inquiry into the Original of Our Ideas of Beauty and Virtue*, Treatise II. Indianapolis, IN: *Liberty* Classics.

Kant, Immanuel (1996). *Groundwork of the Metaphysics of Morals*. In *Practical Philosophy*, translated and edited by Mary J. Gregor. Cambridge: Cambridge University Press.

Markovits, Julia (2010). "Acting for the Right Reasons." *The Philosophical Review* 119: 201–42.

Mill, John Stuart (1998). *Utilitarianism*, edited by Roger Crisp. Oxford: Oxford University Press. (Because there are so many editions, references are to chapter number.)

Moore, G. E. (1993). *Principia Ethica*, rev. ed. with the preface to the (projected) 2nd ed. and other papers, edited with an introduction by Thomas Baldwin. Cambridge: Cambridge University Press.

Murdoch, Iris (2014). *The Sovereignty of Good*. New York: Routledge.

Norcross, Alasdair (2006). "The Scalar Approach to Utilitarianism." In *The Blackwell's Guide to Mill's Utilitarianism*, edited by H. West, 217–32. Oxford: Oxford University Press.

Parfit, Derek (2011). *On What Matters*, 2 vols. Oxford: Oxford University Press.

Pufendorf, Samuel (1934). *On the Law of Nature and Nations*, translated by C. H. Oldfather and W. A. Oldfather. Oxford: Clarendon Press.

Ross, W. D. (2002). *The Right and the Good*, edited by Philip Stratton-Lake. Oxford: Oxford University Press.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

——— (2014). *Being Realistic about Reasons*. Oxford: Oxford University Press.

Scheffler, Samuel (1982). *The Rejection of Consequentialism*. Oxford: Clarendon Press.

Sellars, Wilfrid (1962). "Truth and 'Correspondence.'" *Journal of Philosophy* 59: 29–56.

Sidgwick, Henry (1967). *The Methods of Ethics*, 7th ed. London: Macmillan.

Smith, Michael (1994). *The Moral Problem*. Oxford: Blackwell.

Strawson, P. F. (1968). "Freedom and Resentment." In *Studies in the Philosophy of Thought and Action*. London: Oxford University Press.

Vranas, Peter (2005). "The Indeterminacy Paradox: Character Evaluations and Human Psychology." *Noûs* 39: 1–42.

Williams, Bernard (1985). *Ethics and the Limits of Philosophy*. Cambridge, MA: Harvard University Press.

7

MORALITY AND THE BEARING OF APT FEELINGS ON WISE CHOICES

Howard Nye

1. Introduction

Most of the time we assume that we should be moral—we think that there are good practical reasons for us to do what morality requires us to do. But, especially when morality's demands seem onerous, we can be tempted to waver in this belief, and crave a philosophical account of *why* there are such practical reasons to be moral. It is often assumed that this explanation will have to invoke substantive normative considerations, and many normative ethicists argue that their views provide the best explanation of why we should be moral. Contractarians like David Gauthier (1986), Kantians like Christine Korsgaard (1996), Contractualists like Tim Scanlon (1998), and Consequentialists like Peter Railton (1986) contend that if—but only if—their accounts of the demands of morality are correct, we can explain our reasons to be moral in terms of our allegedly more obvious practical reasons to promote our own self-interest, consistently value agency, act on principles no one can reasonably reject, or promote the well-being of others.

In this paper I argue to the contrary that the best explanation of why we should be moral is actually neutral about the content of morality. I contend that it is a conceptual truth that, if an act is required or recommended by morality, there are genuine practical reasons to perform it. I show how this follows from the best fitting attitude analyses of our moral concepts and a general relationship between fitting motives and practical reasons. While this account of the existence of practical reasons to be moral is neutral about what morality requires, it has important implications for normative ethics. By removing the explanation of why we

should be moral as a desideratum on normative ethical theories, it improves the prospects of theories like Rossian Pluralism, which seem ill placed to give a unified morality-independent explanation of why we should do what they say we are morally required to do.

My conceptual explanation of why there are practical reasons to do what morality requires or recommends consists of two main parts: (1) an analysis of an act's being required or recommended by morality in terms of the appropriateness or fittingness of having a particular kind of attitude towards that act, and (2) an account of the relationship between the fittingness of motivational states and what there is reason to do. Following authors like A. C. Ewing (1939) and Allan Gibbard (1990), I argue that analyzing moral concepts in terms of the fittingness of moral emotions like guilt and anger provides the best explanation of what distinguishes moral judgments from other normative judgments. But I contend that an act's being required or recommended by morality is best analyzed in terms of the fittingness of forward-looking feelings of obligation to perform it, which involve motivation to perform the act.

Next, I argue that the existence of practical reasons to perform an act is a matter, not of the act's contributing to an end that one *is* motivated to pursue, but to an end that *it is fitting* to be motivated to pursue. Since an act's being required or recommended by morality conceptually entails the fittingness of motivation to perform it, and the fittingness of this motivation conceptually entails that there is reason to perform it, it is actually a conceptual truth that there are reasons to do what is required or recommended by morality, whatever that turns out to be.

Finally, I show how my account can explain why, although moral considerations are not always overriding, we necessarily have *conclusive* reasons to do what morality requires. I contend that an act counts as morally required only if the reasons to feel obligated to perform it are conclusive, which entails that it is unfitting to be most strongly motivated not to perform it. This, together with my account of the connection between fitting motives and practical reasons, entails that whatever considerations are weighty enough to make the act morally required are conclusive reasons to perform it. This account goes deeper than that of authors like Stephen Darwall (2006) and Douglas Portmore (2011), who claim that there are conceptual connections between (1) an act's status as wrong and its status as blameworthy, and (2) its status as blameworthy and the existence of conclusive reasons not to perform it. My account shows how an act's status as morally wrong *explains* both the existence of conclusive reasons to perform it and the connection between blame and practical reasons to which such authors appeal.

While it is the business of substantive ethical theorizing to tell us which considerations if any are weighty enough to make acts wrong, I argue that our reasons to believe that certain acts (like inflicting massive suffering on innocents just for fun) are wrong are just as good as our reasons to believe any other substantive normative claim (like that the fact that an act would cause oneself pain is a reason to avoid it). What my metaethical account gives us is an explanation of why our excellent reasons to believe that certain acts are wrong constitute equally excellent reasons to believe that the considerations that make them wrong are

conclusive reasons not to perform them—without recourse to any further substantive ideas about the nature of practical reasons.

2. Moral Emotions and the Attractions of Fitting Attitude Analyses

Perhaps the greatest initial opposition to my approach will stem from the *judgmentalist* (or "cognitivist") view that emotions like guilt and feelings of obligation themselves involve judgments about moral blameworthiness or wrongness, and thus cannot be used to informatively analyze moral judgments.[1] I thus begin by considering some problems with judgmentalism and some virtues of my alternative approach to the relationship between moral emotions and moral judgments.

A widely discussed problem for judgmentalism is the phenomenon of *recalcitrant emotions*, or emotions we feel but think unfitting or inappropriate. For instance, one can feel guilt in spite of the fact that one believes that one has done nothing blameworthy, and one can feel outrage or resentment towards someone despite the fact that one judges that she has done nothing wrong. Judgmentalism is committed to the view that having a recalcitrant emotion involves making a judgment in conflict with other judgments one holds. But merely having a recalcitrant emotion seems not to have to involve such a conflict in judgment. Suppose, for example, that I feel guilt for knocking over and breaking a friend's lamp, though I exercised all due caution and think that I did nothing at all wrong or blameworthy. A conflict in judgment about whether I had done something culpable would involve such things as conflicting tendencies to draw inferences about the moral status of similar acts, conflicting views about whether I deserve reproach, and conflicting views about whether something is wrong with me for feeling what I feel.[2] But it seems that I can feel guilt about breaking the lamp and judge that the guilt makes no sense without any of these kinds of conflicts.

In response to this feature of recalcitrant emotions, those with judgmentalist leanings sometimes opt for a *quasi-judgmentalist* (or "perceptualist") view according to which moral emotions like guilt involve "moral evaluations" that are something less than full-blown judgments, but still attribute moral properties like *blameworthiness* and *wrongness*.[3] The

1 For proponents of this view see, for example, Solomon (1976, 1988), Sabini and Silver (1982), and Foot (1959).

2 Cf. D'Arms and Jacobson's (2003, 129–30) discussion of the difference between merely fearing flying and judging flying dangerous, and the relevance of this claim to the judgmentalist's need to posit inconsistent judgments wherever there are recalcitrant emotions. They discuss how those with phobic fears of flying "are typically well aware that [flying] is safer than activities they do not fear, such as driving to the airport . . . they do not worry when their friends fly, or buy insurance when forced to fly themselves," concluding in their footnote 7 that "the great challenge for judgmentalist accounts of recalcitrant emotion is that the behavioral evidence supporting the attribution of the evidentially suspect belief is problematic."

3 For examples of quasi-judgmentalists treatments of this kind see, for instance, Roberts (1988) and Greenspan (1988). For criticisms of quasi-judgmentalism related to (as well as distinct from) those I present here, see Gibbard (1990, 39–40, 129–32) and D'Arms and Jacobson (2003).

quasi-judgmentalist idea seems to be that moral emotions involve states that are more akin to "moral perceptions" than moral judgments.[4] In general, perceptual states might be distinguished from beliefs or judgments in that they are more "domain-specific" or less sensitive to learning and multiple sources of information, more quickly instanced, and incapable of being consciously inferred.[5] But perceptual states also play an important role in inference processes by contributing their contents as "starting points" or data, which can of course be debunked by theories that best explain the totality of such contents, but in favor of which a burden of proof is set in inquiry. To the extent that we have "moral perceptions" that play these roles, I think that we may know them as "moral intuitions."

The problem for quasi-judgmentalism, however, is that just as merely having a recalcitrant emotion seems not to have to involve conflicting moral judgments, it seems not even to have to involve a conflict between moral judgment and moral intuition. Return again to the guilt I feel for breaking my friend's lamp, despite judging that I have done nothing blameworthy. It certainly seems that I can feel and judge this way without my having an intuition to the contrary—that is, without my having even spontaneous appearances to the effect that I deserve reproach or that there would be something wrong with me were I to fail to feel the guilt I do, and without any tendency to set a burden of proof in inquiry in favor of the view that my conduct was wrong.

But (quasi-) judgmentalists may face a problem even deeper than those posed by the phenomenon of recalcitrant emotions. This is that they must explain what these moral judgments *are*, which feelings like guilt, outrage, and resentment supposedly involve, without reference to these feelings themselves.[6] In light of the wide diversity of things that people have coherently (though in many cases quite falsely) believed to be morally blameworthy or wrong, this seems to be a very difficult task. Some of these things include: inflicting harms upon others, failing to prevent harms to others, defecting in the presence of collective action problems, and failing to respect the autonomy of other agents. But they would also include all manner of apparently miscellaneous behavior, including sexual practices, drug use, violations of etiquette, "playing God" by engaging in cloning or genetic modification of organisms (quite apart from its effects on any individual's well-being), failures to adhere to certain religious practices, stringing together certain phonemes (in the form of curse words), and so on. It should be emphasized that these kinds of apparent miscellany can and have been coherently thought to be intrinsically wrong or blameworthy quite apart from beliefs about their contribution to anyone's well-being or autonomy.

Thus, an analysis of MORAL BLAMEWORTHINESS or WRONGNESS in terms of an act's failing to maximize happiness, or being hated by deities, or violating autonomy, or possessing any

4 For instance, Roberts (1988, 187–88) contends that such emotions involve "construing something in terms of a concept," which he explains by reference to how ambiguous images, like the duck-rabbit, give rise to different perceptual states depending upon which concepts are tokened.

5 See, for example, Zimbardo and Weber (1997, chapter 5, especially 177–97).

6 Cf. Gibbard (1990, 130): "Anyone who claims that anger includes a judgment of moral transgression needs to explain the judgment."

other substantive features would fail to account for how diverse coherent moral judgments can be, and what is at issue between people with rival moral views.[7] Perhaps, however, we can explain what is common to all moral judgments by reversing the judgmentalist order of explanation and analyzing moral judgments as judgments of the fittingness of moral emotions. For instance, following such figures as John Stuart Mill and A. C. Ewing, Gibbard (1990, 40–45, 126–27) proposes the following analysis of the concept of MORAL BLAMEWORTHINESS:

> **Fitting Attitude Analysis of Moral Blameworthiness**: To judge that what someone has done is morally blameworthy is to judge that it is fitting for her to feel guilty for having done it, and fitting for others to be angry at her for doing it.

All of the above coherent judgments that acts are blameworthy do seem to involve judging them to befit guilt and anger, and it seems difficult to identify anything else that they have in common.

One might wonder, however, what we gain by saying that common to all coherent moral judgments and disputes are views and disputes about which moral emotions are fitting, as opposed to saying that moral concepts simply resist being informatively understood in *any* further terms. What we seem to gain is an explanation of what moral judgments have in common with other normative judgments, like those concerning the fittingness, rationality, or appropriateness of desires, beliefs, and non-moral emotions. Common to all of these are views that a certain attitude is favored by reason, and the attitudes held to be favored by reason in the case of moral judgments are moral emotions like guilt and anger. These judgments about the fittingness of attitudes share such features as a wide diversity of things that can be coherently thought to befit them, our attempting to determine which of these coherent positions are correct via *a priori* methods of philosophical argument, and our conclusions about which responses are fitting exerting direct (non-behavior-mediated) causal pressure on our coming to have them.[8] By subsuming these phenomena that we see in the case of moral judgments in relation to moral emotions under those of judgments of the fittingness of attitudes generally, we gain an (at least partial) explanation of them.

---

7 I should perhaps emphasize that by itself this in no way entails the falsity of any *substantive view* about which acts are morally wrong or blameworthy, including the utilitarian claim that the morally wrong acts are all and only those that fail to maximize happiness. The point here is simply that the utilitarian cannot intend her view as a conceptual analysis of WRONGNESS; the denial of the position is coherent even if false, and we need an understanding of WRONGNESS that can capture the substantive dispute between the utilitarian and her rivals.

8 On its being a distinguishing characteristic of judgments that attitudes are fitting (as opposed to supported by non-fittingness considerations) that they are capable of directly causing us to have them, see Gibbard (1990), Parfit (2001), Hieronymi (2005), and Raz (2009). As D'Arms and Jacobson (2009) argue, mere direct causal influence may be insufficient to distinguish fittingness from non-fittingness judgments in all cases. But as I suggest in the text (and argue in more detail elsewhere—see Nye 2009, chapter 6), judgments of fittingness are best understood as part of a psychological process characterized by several functional roles, which include but are not limited to directly influencing our attitudes.

3. Feelings of Obligation, Moral Wrongness, and Moral Reasons

In order to see how fitting attitude analyses can shed light not only on the nature of our moral concepts but also on the connection between morality and practical reasons, I believe that we must turn our attention from aretaic or hypological concepts like BLAMEWORTHINESS to deontic concepts of MORAL WRONGNESS and MORAL REASONS. After proposing to analyze judgments of moral blameworthiness as judgments about the fittingness of anger and guilt, Gibbard noted some ways in which blameworthiness and wrongness can come apart. For instance, if one lashes out in grief at a friend offering condolences, one's conduct may be wrong but nevertheless exculpated by one's overwhelming grief (Gibbard 1990, 44).

Gibbard concludes that "we need a distinct concept of wrong . . . as opposed to blameworthy," noting that while the concept of blameworthiness is *retrospective* in character, the concept of wrongness is *prospective*. I think that the best way to understand this forward-looking character of the concept of moral wrongness is to see that it is concerned, not with the fittingness of guilt and anger towards what has already been done, but with the fittingness of the agent's *feeling obligated* to do or avoid doing various things that it is open to her to do.

Feelings of obligation are, as Richard Brandt (1959, 117–18) observed, what you have when you see someone in trouble and feel like you "just can't" leave her. J. S. Mill (1863) described the feeling as an "internal sanction of duty . . . a feeling in our own mind . . . attendant on violation of duty, which in properly cultivated moral natures rises, in the more serious cases, into shrinking from it as an impossibility," and "a mass of feeling which must be broken through in order to do what violates our standard of right." The phenomenology of feeling obligated not to do something is similar to that of feeling guilt for having done it, but whereas guilt is retrospective, feeling obligated not to do something involves a kind of prospective guilt-tinged aversion to doing it.[9]

As associated as these feelings of obligation may be with judgments that one is morally obligated to do something, it is possible to feel obligated recalcitrantly, or to feel obligated not to do things that one judges not to be wrong. For example, a man from a background with restrictive views about sexual morality might feel obligated not to engage in certain sexual practices even though he now thinks them perfectly morally permissible. Or a woman in an abusive relationship might feel obligated not to leave her partner, but be thoroughly convinced that she is in no way morally required to stay with him. As with our discussion of recalcitrant guilt above, it seems that the man and woman could in this way recalcitrantly feel obligated without any of the conflicting inferential tendencies, views about appropriate

9 It is important, however, to emphasize that feeling obligated not to do something involves an aversion to *doing it*, not to the prospect of feeling guilt for having done it. If you saw someone in need of help but had on hand a pill that would prevent you from feeling guilt for failing to help her, your feeling that you "just can't" leave her (unlike an aversion to feeling guilt) would motivate you to help her and generate no motivation at all to take the pill.

conduct, or views about their own responses required for conflicting judgments about what they are morally obligated to do. It seems, moreover, that the man and woman could recalcitrantly feel obligated without any of the spontaneous appearances and tendencies to set burdens of proof in inquiry required for an intuition or sub-judgmental moral evaluation in conflict with their judgments about their moral obligations.

For reasons similar to those that favor analyzing judgments of moral blameworthiness as judgments about the fittingness of guilt and anger, I think that the content and normative force of judgments that acts are wrong or opposed by moral reasons are best captured by analyzing them as judgments about the fittingness of feeling obligated not to perform them. For instance, what seems distinctive about viewing the fact that doing *A* will save someone's life as a *moral* reason to do *A* is one's taking this consideration to count in favor of feeling obligated to do *A*.[10] Similarly, what seems distinctive about thinking that the fact that doing *A* would kill someone makes it morally wrong or forbidden (as opposed to just unreasonable) to do *A* seems to be one's taking this consideration to make it, on balance, fitting for you to feel obligated not to do *A*.

This supports the following analyses of our concepts of MORAL REASONS and MORAL WRONGNESS:

Fitting Attitude Analysis of Moral Reasons: To judge that *R* is a moral reason for agent *X* to φ is to judge that *R* is a fittingness reason for *X* to feel obligated to φ, and

Fitting Attitude Analysis of Moral Wrongness: To judge that it is morally wrong for *X* to ψ is to judge that it is, on balance, fitting for *X* to feel obligated not to ψ.[11]

---

10 The best alternative proposal about what is distinctive about viewing this as a moral reason is something like that it involves one's taking it to be a reason that one has simply because one's act will promote the well-being of the individual in question. But it is surely *coherent* to think that there are distinctly moral reasons to do things other than promote well-being: with some plausibility one can think there are intrinsic moral reasons to respect autonomy and keep promises, and we know only too well what someone is thinking when she takes the alleged fact that an act is "unnatural," "against tradition," or "against God's will," to be an intrinsic moral reason against doing it. Moreover, although most of us are decent enough to accept a substantive principle of beneficence according to which there is intrinsic moral reason to promote the well-being of every individual capable of well-being, it is, sadly, coherent to think otherwise. The view that there are individuals whose well-being there is no intrinsic moral reason to promote (although perhaps still some intrinsic non-moral reason to promote) has been coherently entertained, for instance, by some who take exalted views of the moral relevance of such factors as retribution, autonomy, promise-keeping, supernatural wills, and group-loyalty.

11 Of course, we can think it perfectly fitting for someone to *experience* no feelings of obligation to refrain from doing things we think wrong if she is already sufficiently motivated not to do them. In most cases we would never even consider doing things that would kill others, and if we do, care for those others and fear of punishment are almost always sufficient deterrents. Although we think it would be wrong for us to kill in such cases we surely do not think it inappropriate that we experience no feelings of obligation to refrain from doing so. Moreover, as I discuss in more detail below in section 5, there is a sense in which we can

Just as judgments about the blameworthiness of actions have the central normative feature of guiding feelings of guilt and anger towards them, judgments about wrongness and moral reasons seem to have the central normative property of guiding feelings of obligation. These fitting attitude analyses of moral judgments can explain their ability to generate motivation to act out of feelings of obligation as a special case of the ability of judgments that attitudes are fitting to directly guide us into having them.[12]

Finally, these analyses can help explain the gap Gibbard noted between judging an act wrong and judging it blameworthy. Combining them with Gibbard's existing analysis of blameworthiness, we can understand, say, thinking that someone's lashing out in grief was wrong but not blameworthy as a thought to the effect that although it isn't fitting for us to feel angry at the person who lashed out and it isn't fitting for her to feel guilt for lashing out, it still was the case that before she lashed out she should have felt obligated not to do it.[13]

---

think it fitting on balance to feel obligated to do things that we do not think it wrong to fail to do. It seems perfectly fitting for someone who goes above and beyond what morality requires, say by getting killed to save a younger stranger from death, to feel obligated to do what this.

To clarify my proposal, it is important to note first that talk of feeling emotions, like talk of desiring or preferring, is ambiguous between an occurrent and a dispositional sense. Occurrent feelings and preferences exert causal pressure on one's behavior at the moment, and (at least typically) involve phenomenal experiences, while dispositional feelings and preferences merely have the disposition to become occurrent in certain circumstances. Thus, one can dispositionally feel obligated not to push one's friends out of windows in the same way one can dispositionally feel anger at one's father even while one is enjoying his company and experiencing no negative emotions. Second, as I discuss below in section 5, it is important to note that a response's being "fitting on balance" is ambiguous between (1) the response's being *mandatory*, in that there is no alternative response that is *as* strongly supported by fittingness reasons, or (2) the response's being *justified*, in that there is no alternative response that is *more* strongly supported by fittingness reasons.

In more detail, then, my proposal is that to think it morally wrong for X to ψ is to think that it is mandatory for X to have at least a dispositional feeling of obligation not to ψ (and mandatory for X to have an occurrent feeling of obligation not to ψ only if X is not already sufficiently motivated not to ψ). The sense in which one can judge it "fitting on balance" for X to feel obligated to φ when one takes X's φ-ing to be supererogatory is that one thinks X's feeling of obligation is justified but not mandatory.

12 To appreciate the centrality of this attitude-guiding role of moral judgments, suppose that someone were to label as "morally wrong" all those things we would call wrong, but took this to have no significance for what it was appropriate to feel obligated to do and consequently had no propensity to feel obligated not to perform the acts in question. It seems that by "wrong" she would not really mean wrong. On the other hand, if someone were to label as "morally wrong" precisely those things we think permissible, she would still seem perfectly intelligible as thinking that those things are wrong so long as she thought it was fitting to feel obligated not to perform them.

13 One might be wondering, however, why there is not in addition to a coherent wrongness-blameworthiness gap a similar coherent blameworthiness-wrongness gap. As we have seen, it seems perfectly coherent to think that it is fitting to feel obligated not to perform an act but that it is also unfitting to feel guilt or for others to feel outrage at one for performing it. But it seems incoherent to think that it is fitting to feel guilt or for others to feel outrage at one for performing an act if it was not fitting for one to feel obligated not to perform it in the first place. One thing I should point out is that conceptual connections between the fittingness of different moral emotions are already an issue for fitting attitude analyses like Gibbard's analysis of moral blameworthiness, in that their proponents need to explain why it seems incoherent to hold that

4. Fitting Attitudes and Reasons to Act

I have thus argued that to judge an act wrong or opposed by moral reasons is to judge that there are considerations that make it fitting to feel obligated not to perform it. Since a judgment's truth entails the truth of its analysans, this means that it is a conceptual truth that (1) *R is* a moral reason for *X* to φ iff *R is* a fittingness reason for *X* to feel obligated to φ, and (2) *X*'s ψ-ing *is* morally wrong iff it *is* fitting for *X* to feel obligated not to ψ.[14] I will now argue that these analyses, together with general facts about the relationship between fitting attitudes and reasons to act, explain why an act's deontic status entails the existence of practical reasons to perform or avoid performing it.

The basic idea here is that what there is reason for us to do is determined by what aims there is reason for us to have, and the question of what aims there is reason to have is identical to that of what it is fitting to be motivated to do. Since attitudes like feeling obligated to φ (or to avoid ψ-ing) involve motivation to φ (or avoid ψ-ing), the fittingness of these attitudes

---

acts can befit outrage on the part of others without befitting guilt on the part of their performers, or that acts can befit guilt on the part of their performers without befitting outrage on the part of others. I think that the answer in both cases is that it is a conceptual truth about these attitudes that states would not count as guilt, outrage, or feelings of obligation unless their fittingness was interrelated in these ways.

Although the details are beyond the scope of this paper, here very briefly is how I think we came to have emotion concepts like this. For evolutionary reasons, our ancestors tended to feel guilt and outrage only towards acts that were such that their performers would tend to feel obligated not to perform them the first place—but to tend not to feel outrage or guilt towards all such acts the performers of which would tend to feel obligated not to perform (when the performance was, e.g., due to overpowering impulses). When the governance of emotions by norms came on the scene, similar evolutionary pressures favored our ancestors' accepting systems of norms that prescribed feeling guilt and outrage only towards acts they also prescribed feeling obligated not to perform in the first place (but not vice versa). This was such a central feature of our ancestors' systems of norms for emotions that the folk psychological theory that came to be true of us was thus one according to which the states that played the guilt, outrage, and feeling of obligation roles were such that the first two were prescribed by norms only when the last was (but not vice versa, and also such that each of the first two were prescribed by norms only when the other was). Because such a folk theory was true of us (and we weren't too dim), it was the folk theory we came to have, and from which our folk emotion concepts of guilt, outrage, and feelings of obligation were extracted via the Ramsey-Carnap-Lewis method of analyzing theoretical concepts (see, for instance, Lewis (1970, 1972). For more on this suggestion about the origin of our emotion concepts as it relates to conceptual connections between the fittingness of different emotions, see Nye (2009, chapter 5).

14 Compare: if judging someone to be a bachelor amounts to judging him to be a male who is not in a romantic relationship but in a position to enter one, then it is a conceptual truth that someone is a bachelor iff he is a male who is not in a romantic relationship but in a position to enter one. Because analyses of one kind of judgment into another in this way support analytic relationships between the facts the judgments represent, I will slide rather freely between talking about what it is to make a certain kind of judgment ("to judge an act wrong is to judge that it is fitting to feel obligated not to perform it") and talking about the analytic relationships between the facts they represent ("it is a conceptual truth that an act is wrong iff it is fitting to feel obligated not to perform it").

entails the fittingness of this motivation, which entails the existence of reasons to φ (or to avoid ψ-ing).

The first part of this connection between fitting attitudes and reasons to act can be stated as a

> **Warrant Composition Principle** [WCP]: Let *M* be a mental state that involves mental state *M'* as an essential component. If *R* is a fittingness reason to be in *M*, then *R* is a fittingness reason to be in *M'*.

WCP simply states that if there is reason to be in a mental state, then necessarily there is reason to be in all that the state essentially involves. For instance, if one acknowledges my claim that part of what it is to feel obligated to φ is to be motivated to φ, it would seem incoherent to hold that a consideration (like φ's relieving someone's pain) counts in favor of feeling obligated to φ but does not count in favor of being motivated to φ. Since having the motivation is simply part of what it is to have the feeling of obligation, a consideration cannot make the feeling of obligation fitting without making the motivation fitting as well.

The second part of this connection between fitting attitudes and reasons to act is the relationship between what it is fitting to be motivated to do and what there is reason to do, which can be stated as a

> **Motivations-Actions Principle** [MAP]: Let φ-ing be an action. If *R* is a fittingness reason to be motivated to φ, then *R* is a reason to actually φ.

Just as the consideration that φ-ing would relieve someone's suffering cannot make it fitting to feel obligated to φ without making it fitting to be moved to φ, so too it seems this consideration cannot make it fitting to be moved to φ without actually counting in favor of φ-*ing*.[15]

It is intuitive that what there is reason to do is determined by what aims there is reason to have. I think that the best theoretical explanation of MAP is that, because practical reasoning governs our actions by means of governing our motives, the process of determining what aims to have—and thus what to do—is essentially a process of determining what intrinsic motives to have. As authors like Michael Bratman (1987, 54) and Thomas Scanlon (1998, 20–21) have argued, because our practical reasoning controls our actions by controlling our intentions to perform them, reasons to perform an action just are reasons to intend to perform it. But it must be clarified that reasons to do *A* are identical to *fittingness* reasons to intend to do *A*. As Kavka's (1983) toxin puzzle illustrates, merely pragmatic reasons to intend to do something (like the reason to intend—or get oneself to intend—to drink a toxin

15 WCP and MAP closely resemble John Skorupski's principles FDF and FDD, the conjunction of which he referred to as the "Feeling / Disposition Principle" (1999, 38, 63, 131, 174 n24) and more recently as the "Bridge Principle" (2010, 265–67).

tomorrow constituted by the fact that a reliable mind reader will pay you now if you intend this) need not be reasons to actually do it.

Moreover, in light of the role intentions play in realizing the objects of our desires and other valenced attitudes (like feelings of obligation), there are similar reasons to think that, because reason ultimately governs our intentions by governing these attitudes, the fittingness of intentions is itself determined by the fittingness of these other motives. As Bratman argues, the role of intentions is not to supply an utterly new source of motivation that conflicts with the motives involved in our valenced attitudes, but to help cognitively limited agents like us realize the objects of these motives over time. This role of intentions entails that their normative assessment must be tied closely to that of the valenced attitudes they serve.

Although Bratman often speaks as though practical reasoning must simply take our intrinsic valenced attitudes as given,[16] it seems clear that we can assess them as reasonable or unreasonable by determining through philosophical reasoning whether they are fitting or unfitting. Moreover, as we have seen, it is characteristic of these fittingness assessments that they directly guide our attitudes. For instance, one might start out with much weaker feelings of obligation to avoid inflicting harms of a given size (like a given amount of suffering), independent of their further consequences, upon non-human animals than upon humans. But one might then reflect upon how bare biological species membership amounts merely to something like a shared history of phylogenetic descent, phenotype-independent genotype, or psychology-independent morphology, and how, with reference to profoundly intellectually disabled humans, someone's lesser intellectual ability does not seem to justify lesser concern with her equally-sized interests. As a result, one might come to judge one's initially weaker feelings of obligation to avoid inflicting given-sized harms on non-human animals to be unfitting. This judgment that one's feelings are unfitting tends directly to change them, and to alter one's intentions from carrying out one's previous aim of, say, avoiding the infliction of given-sized harms on non-human animals only when one finds this relatively convenient (consistent, for instance, with one's previous intentions to continue to consume animal products) to carrying out one's new aim of avoiding inflicting such harms on non-human animals with as much priority as one gives to the avoidance of inflicting them upon humans (issuing for instance, in a new general intention to consume a vegan diet, and a new particular intention to buy vegan food at the grocery store).

Thus, because reason governs motives other than intentions through determinations of their fittingness, and intentions are simply a means of achieving the objects of these motives, fittingness reasons for these other motives are identical to fittingness reasons for intention. The role of intentions is primarily to enable us to settle in advance what future courses of action will best achieve the ends that it is fitting to be most motivated to achieve (like being

16 While Bratman often speaks as though our intrinsic desires or pro-attitudes are *themselves* normative reasons for intention and action, he makes it clear that he actually wishes to remain neutral between this view and the view that our intrinsic desires can be assessed as reasonable or unreasonable (1987, 22).

vegan to minimize harm to non-human animals), and to pick from among the many courses of action that often have equally good prospects of doing this (like going to one of two otherwise equally choiceworthy grocery stores to purchase vegan food). Together with the above observation that reason governs our actions through determinations of the fittingness of the intentions that lead us to perform them, this entails that, because reason ultimately governs our intentions and actions by determining the fittingness of the motives they seek to serve, fittingness reasons to be motivated to do something are identical to fittingness reasons to intend to do it and practical reasons to do it.[17]

Having thus argued in favor of my fitting-attitude analyses of moral wrongness and reasons, WCP, and MAP, I can use them to give the following explanation of why, if a consideration is a moral reason to do something, then it is also a practical reason to do it, and why, if an act is morally wrong, then one has practical reason not to perform it. Since feeling obligated to φ essentially involves motivation to φ (and feeling obligated not to ψ essentially involves motivation not to ψ), it follows from (1) and (2) together with the WCP that (1') if *R* is a moral reason to φ, then *R* is a fittingness reason to be motivated to φ, and (2') if ψ-ing is morally wrong, then it is fitting to be motivated not to ψ. Moreover, it follows from (1') and (2') together with the MAP that (1*) if *R* is a moral reason to φ, then *R* is a genuine practical reason to φ, and (2*) if ψ-ing is morally wrong, then there are genuine practical reasons not to ψ.

5. Conclusive Reasons Not to Do Moral Wrong

If my strategy for explaining moral concepts and their relation to reasons for action is correct, we are thus guaranteed a conceptual connection between an act's being favored by moral reasons and our having practical reason to perform it. Our having moral reason to do something seems to entail that we have some practical reason to do it, but not conclusive

17 It is important to clarify that neither MAP nor this explanation of it makes what there is reason for us to do dependent upon what attitudes or motives we actually happen to have. The idea is that what there is reason for an agent to do is what would serve the objects of fitting intrinsic motives, by which I mean the intrinsic motives that it *would be fitting* for her to have, *whether she has them or not*. On this view, if a consideration (like a policy's preventing suffering) is a fittingness reason for an agent to be motivated to do something (like vote for the policy if she can), then it is a reason for the agent to do this regardless of whether she ever has or comes to have any actual motivation to do it.

The point of the appeals to how actions and intentions are governed by reason is simply to establish that, because (in the absence of something going wrong—as when we do the right thing for the wrong reasons) we can only respond to *genuine* normative reasons to act and intend by our motives first responding to these considerations, fittingness reasons to be motivated to do something are fittingness reasons to intend to do it and practical reasons to do it. The underlying idea is that if *R* is a genuine normative reason for us to respond in way *W*, and (absent something going wrong) we can only have *W* in response to *R* by having (and because we have) *W** in response to *R*, then *R* is a reason to *W**, and its status as such explains its status as a reason to *W*.

reason. Some acts, like getting oneself shot in the head to save a slightly younger stranger from drowning, seem to be favored by moral reasons, but seem not to be morally required, or are not morally wrong to fail to perform. In the case of such an act, the moral reasons to perform it (viz., that it would save the stranger's life) seem to constitute practical reasons to perform it. But it at least seems coherent to think that one lacks conclusive practical reason to perform such an act, or that one has sufficient practical reasons not to perform it (constituted, e.g., by the fact that the act will get you killed).

But the notion of MORAL WRONGNESS or MORAL REQUIREMENT (where what is morally required is just what it is morally wrong not to do) seems to be different in this respect. There seem to be genuine problems with the coherence of thinking that one's doing something would be morally wrong but that one has sufficient reason to do it anyway. As such, a strong version of what Stephen Darwall (1997, 306) calls the thesis of "morality-reasons internalism" might seem to be true of moral requirement, namely "if *S* is morally required to do *A*, then necessarily there is *conclusive* reason for *S* to do *A*."

It would be odd, however, if this strong thesis were to be explained solely in terms of the weightiness of the considerations that make acts morally required or wrong not to perform. Such considerations—for instance that *I have promised to be across town* and that *she will die if I don't stay and help*—can be brought into conflict without necessarily giving rise to rational dilemmas,[18] and it seems at least coherent to think that they are at times outweighed by non-moral reasons like *getting across town will get me killed*.[19] As such, I think that a much more attractive explanation of the strong thesis is that whether an act gets to count as falling under our concept of MORAL REQUIREMENT—unlike, say, our concept of simply being favored by moral reasons—is itself sensitive to whether or not the reasons in favor of performing it are actually conclusive. That is, as W. D. Falk (1948, 124) suggested, "our very thinking that we ought [that is, are morally required] to do some act already entails that, by comparison, we have a stronger reason in the circumstances for doing it than any other."

I think that the strategy I have been pursuing for analyzing moral concepts and explaining their connections to reasons for action can help explain why Falk's kind of account of the necessary conclusivity of reasons not to do moral wrong is in fact correct. First, I should clarify the kind of fittingness reasons for feeling obligated not to do things with which the fitting attitude analysis identifies their moral wrongness. The idea is not that an act is morally wrong if one is simply justified or rationally permitted to feel obligated not to perform it in the same way in which one is justified or permitted to feel angry at actors whose conduct is

18 By which I mean situations in which whatever one does is irrational or other than one has most reason to do. I, for the record, do not think that it is conceptually possible for there to be such situations.

19 If the reader thinks that duties to oneself render this a moral reason, I invite her to consider whether there is some degree of trivialness of promise and some degree of harm that will befall one if one keeps it such that it is at least coherent to think that: (1) were it not for the harm to oneself one would be morally required to keep the promise, but (2) given the harm one would incur by keeping the promise it is rationally permissible to break it, yet (3) one does not "owe it to oneself" to prevent the harm to oneself by breaking the promise.

blameworthy. That an act is blameworthy does not entail that others are necessarily feeling something unfitting if they fail to feel angry with the blameworthy actor. Especially if the transgression is slight, it might be perfectly fitting for others to fail to feel such anger if they have more important matters to tend to, if the blameworthy actor is very remote, or if the blameworthy actor has done her best to make amends for what she has done or enough time has passed.[20] The kind of reasons one has to feel obligated not to perform morally wrongful acts, however, are not so easily overridden. Rather, to think an act morally wrong seems to involve thinking that, unless one is already going to refrain from performing it (for instance, one is sufficiently motivated not to perform it, or simply not motivated to perform it), one has *conclusive* fittingness reason to feel obligated not to perform it, in the sense that it would be unfitting for her not to feel so obligated.[21]

Next, I would argue that while conclusive reasons to be in motivational states like emotions need not always amount to conclusive reasons to act out of them, conclusive reasons to feel obligated to do things are atypical in this respect. Consider, for instance, our reasons for wanting and against eating good tasting but unhealthy foods. That the foods taste so good seems to make it fitting to want to eat them. That the foods are unhealthy seems to make it fitting to be averse to eating them, and to be a practical reason to refrain from eating them. But the unhealthiness of the foods does not seem to make it unfitting to want to eat them at all. In some cases, we take the latter set of reasons to be weightier, and think that, all things considered, we should not eat the foods. But since these reasons can leave intact our reasons to want to eat the foods, it seems that in such cases we can have that (1) it is fitting to be somewhat motivated to eat the foods, (2) it is fitting to be more strongly motivated not to eat them, and (3) there is most practical reason to act out of our motives not to eat the foods and refrain from doing so.

What this kind of case suggests is that we need to distinguish between having conclusive fittingness reason to have *some* motivation to perform an act and having conclusive fittingness reason to be more strongly motivated to perform an act than any of its alternatives. Let us call the former states of having some motivation "gradational motivations" to perform the act, and call the latter state of strongest motivation a state of being "most motivated" to perform it. States of being most motivated to perform an act are those that arise as a result of the combined strengths of one's gradational motivations to perform it being greater than the combined strengths of one's gradational motivations not to perform it. The connection between conclusive reasons for motivation and action suggested by the above example, then,

20 Cf. Gibbard (1990, 126–27).

21 Similarly, I should think that the reasons a morally blameworthy actor has to feel retrospective guilt for what she has done are not so easily overridden as those of others to feel angry at her. To think an act blameworthy seems to involve thinking that, at least until its performer has made amends or enough time has passed, and unless significantly more pressing matters arise (which must be more pressing than those minimally necessary to permit others not to feel angry at her), its performer has conclusive reason to feel guilt for performing it.

is that while one can have conclusive fittingness reason to be gradationally motivated to do something without having conclusive reason to do it, one's having conclusive fittingness reason to be most motivated to do it entails that one has conclusive reason to do it. Call this the most-motivation-action principle:

> **Most-Motivation-Action Principle**: If *S* has conclusive fittingness reason to be most motivated to do *A*, then S has conclusive practical reason to do *A*.

We have thus seen how for some gradational motivations, fittingness reasons to be in conflicting motivational states do not themselves constitute fittingness reasons against having these gradational motivations at all. It is in this respect, however, that feelings of obligation to do something seem to be different. Fittingness reasons to be motivated not to do something actually *do* seem to count against the fittingness of feeling obligated to do it. For instance, consider a situation in which one must break a promise in order to save someone's life. The fact that one has promised to do something is a reason to feel obligated to do it. In this case, however, we would seem to take the fact that one must not do what one has promised to do to be a fittingness reason, not only in favor of being most motivated not to keep one's promise, but indeed against feeling obligated to keep it under the circumstances.[22]

---

22 It is important here to distinguish the fittingness reasons to feel obligated to do things of which I am speaking from some closely related phenomena. To borrow (and use for slightly different purposes) an example from D'Arms and Jacobson (1994, 742–43), one's mother might deeply fear being put in a nursing home, though given one's inability to care for her and the costs to other family members one has most reason to put her in a home. In such a case, it might seem consistent with thinking that one has conclusive reason to put mother in the home to think that there is something wrong with one if one does not feel obligated to omit putting her in the home. Similarly, it might seem consistent with one's having conclusive reason to put her in the home that it would be unfitting for one to feel no kind of reluctance towards putting her in the home, or to be able to put her in the home "with perfect equanimity" (my thanks to Stephen Darwall for this way of putting the intuition).

I contend, however, that these are not thoughts about the consistency of thinking that one has conclusive reason to put one's mother in the home with thinking that one has conclusive fittingness reason to feel obligated not to do so. The first is most likely a thought that it is morally bad or disestimable to fail to feel obligated to omit putting her in the home. Such a judgment may easily be mistaken for a judgment that one has conclusive fittingness reason to feel so obligated because the former resembles the latter in two important respects: (1) it entails that there is reason to be motivated to do what one would do if one felt the aversion, and (2) for reasons discussed by Velleman (2002), it can translate into feeling obligated to omit putting mother in the home without one's having to do anything to bring this about (though in a way that is dependent on judging the disesteem fitting).

The second thought is most likely a thought that one has conclusive fittingness reason to feel an attitude that we might call compunction, which bears some similarities to but can still be distinguished from the feelings of obligation the conclusive fittingness of which I am claiming we can understand MORAL WRONGNESS in terms of. Phenomenally, compunction might also seem in a sense to be "guilt tinged," but feeling compunction towards performing an act seems to involve something more like a feeling of hesitancy about, being unsettled about, or reluctance about performing it. What I have been calling feelings of obligation not to perform an act do not seem so aptly characterized in these ways—they seem to involve something

This apparent fact that any considerations that count in favor of being most motivated not to do something must also count against feeling obligated to do it would entail that one's reasons to feel obligated to do something can only be conclusive if they outweigh the reasons in favor of being most motivated not to do it. This in turn entails the following thesis:

Contour Thesis: If *S* has conclusive fittingness reason to feel obligated to do *A*, then *S* has conclusive fittingness reason to be most motivated to do *A*.[23]

We can now combine these considerations in favor of the most-motivation-action principle and the contour thesis with the fitting attitude analysis of moral wrongness to vindicate the strong morality-reasons internalism thesis about moral requirement. Given our clarification of the fitting attitude analysis, if it would be morally wrong for *S* to fail to do *A*, then either *S* is already sufficiently moved to do *A* or *S* has conclusive fittingness reason to feel obligated to do *A*. If *S* is already sufficiently moved to do *A*, then either *S* is in a fitting state of being most motivated to do *A*, or *S*'s state of being most motivated to do A is unfitting. In the latter case (in which, for instance, *S* is only motivated not to do *A* because she irrationally believes that she will be punished for doing *A*), reason requires that *S* cease to be in this state and thus be such that she has conclusive fittingness reason to feel obligated to do *A*. By the contour thesis, if *S* has conclusive fittingness reason to feel obligated to do *A*, then *S* has conclusive fittingness reason to be most motivated to do *A*. Thus, if it would be morally wrong for *S* to

more like a feeling that one "just can't" or "just can't bring oneself" to perform the act. Compunction seems to be more closely associated with going back and forth about or checking and re-checking to make sure about what it would be wrong for one to do.

23 One might wonder why the contour thesis is, as I have argued it seems to be, true, or what makes (or guarantees that something makes) it true. I think that it is a conceptual truth about feelings of obligation that reasons to act contrary to such feelings count against having them at all. Were an attitude to be otherwise similar to these feelings in terms of phenomenology, attention direction, and motivation, but were to lack this feature, I think it would still fail to count as our feeling of obligation, conclusive reasons for which constitute the moral requirement to perform (or the moral wrongness of failing to perform) its object.

Although the details are again beyond the scope of this paper, I think that we came to have an emotion concept like this for much the same reasons as those I mentioned in footnote 13. That is, for evolutionary reasons, our ancestors came to tend to feel this kind of guilt-tinged aversion as an adaptive inhibition to defecting, where it was important to our genes (as it were) that we were conclusively so deterred. See, for example, Kitcher (1998, esp. 299–303) for a discussion of how fragile cooperation can be when inhibitions against defection are just one motive in the "internal melee" among many, as with our evolutionary cousins, the chimpanzees. When the governance of emotions by norms came on the scene, there were similar evolutionary pressures for our ancestors to come to accept norms that required feeling obligated to do something only when they required no stronger (or equally strong) motives to the contrary. This was such a central feature of their systems of norms that the folk psychological theory that came to be true of us was one where the states that played the feeling of obligation role were ones that were prescribed by norms only when no stronger (or equally strong) motives to the contrary were prescribed. Because it was true of us and we picked up on it, this was the folk psychological theory we came to have and over which we Ramsified to arrive at our folk concept of FEELINGS OF OBLIGATION. For more on this see Nye (2009, chapter 5).

fail to do *A*, reason will only allow *S* to be most motivated to do *A*, so *S* has conclusive fittingness reason to be so motivated. Finally, by the most-motivation-action principle, if *S* has conclusive fittingness reason to be most motivated to do *A*, then *S* has conclusive reason to do *A*. Thus, if it would be morally wrong for *S* to fail to do *A*, which is to say that *S* is morally required to do *A*, then *S* has conclusive practical reason to do *A*.[24]

6. Conclusion

So if one thinks that one is morally required to do *A* and still wonders, "why should I do *A*?," the answer is simply: "whatever makes it the case that you are morally required to do *A*." What my account gives us is an explanation of *why* the fact that these considerations are moral-requirement-makers guarantees that they are also conclusive practical reasons to act as one is morally required to act.

It may be important to conclude by emphasizing what my kind of vindication of reasons to be moral does and does not show. On my account, we have conclusive practical reason to do what we are morally required to do only because: (1) for *S* to be morally required to do *A*, it must be the case that, absent a sufficient motivation or tendency to do *A* anyway, *S* has conclusive fittingness reason to feel obligated to do *A*, and (2) in contrast to reasons for some other motivational states, if the fittingness reasons that count in favor of *S*'s feeling obligated to do *A* are not sufficiently weighty to determine that she has conclusive reason to act out of them, they are also insufficiently weighty to determine that she has conclusive fittingness reason to have the feeling of obligation. Like other kinds of reasons, an agent's reasons to

24 I should point out that if one doubts the truth of the contour thesis for reasons related to cases like that discussed in footnote 22, there are still at least two other ways in which my approach to moral concepts and their connection to reasons for action can help to vindicate Falk's kind of account of why we have conclusive reason not to do what is morally wrong. One might wish to insist that it is consistent with thinking that one has conclusive reason to put mother in the home that one has conclusive reason to feel (not just compunction but) *some* feeling of obligation not to do so. It could be argued, however, that this is only consistent with the thought that one has conclusive reason to put mother in the home because one also thinks that one has reason to feel an even stronger feeling of obligation *to* put her in the home (my thanks to Stephen Darwall for making me aware of this option). It could moreover be argued that the thought that one should feel most strongly obligated to put mother in the home entails both that one is morally required to do so and (due to the sensitivity of the fittingness of *strongest* feelings of obligation to countervailing considerations) that one has conclusive reason to be most motivated to put her in the home.

If, however, one is reluctant to accept my defense of the original contour thesis in relation to cases like putting mother in the home, one might be apt to object to the above modification in slightly altered cases for similar reasons. For instance, suppose that one has made a reasonably important promise to a friend that turns out to be extremely personally costly to keep. One might think that it is consistent with thinking that in such a case one has most reason not to keep the promise that one still has conclusive reason to feel (not just compunction but) some feeling of obligation to keep it, and that this is the only thing it makes sense to feel obligated to do in the circumstances. I would still at least contend that the thought that one's feelings of obligation to keep the promise are rationally overpowered by motivations to break it entails that it would not be morally wrong to break the promise.

do things out of feelings of obligation can be overwhelmed by other considerations; it is simply that when they are, they are overwhelmed on the front of determining what it is fitting for the agent to feel and hence (given the fitting attitude analysis of moral wrongness) are no longer sufficient to make it morally wrong for the agent to do otherwise, or morally required for her to do what they are reasons to do.

My vindication of the strong morality-reasons internalism thesis thus does not give any conceptual guarantee that any given consideration—even *I have promised* or *she will die if I don't help her*—is either a genuine moral-requirement-making feature or a weightier reason than any other. It would show, however, that if we are as a matter of substantive normative fact morally required to do something, then we have conclusive practical reason to do it. I think that our reasons to believe that we are morally required to do things are just as good as our reasons to believe that other normative concepts are instantiated, and that this consists in the best unification and explanation of what substantive normative principles are most plausible after ideal reflection on what they are really saying.[25] To the extent that our best such normative theories tell us that we are morally required to do something, I think we should conclude that we are, which I have argued entails that we have conclusive practical reason to do it.

What my conceptual account gives us is an explanation of why our rock-solid reasons to believe that certain acts are wrong constitutes equally rock-solid reasons to believe that the considerations that make them wrong are conclusive practical reasons not to perform them. As such, we do not need an independent substantive account of why we should care about what morality requires us to do. We do not have to rely upon substantive morality-independent intuitions to the effect that we have practical reasons to promote our own self-interest, consistently value agency, act on principles no one can reasonably reject, or promote the well-being of others to justify our practical reasons to be moral if but only if morality requires us to do what Contractarians, Kantians, Contractualists, or Consequentialists say it does. Quite apart from any such substantive morality-independent intuitions about our practical reasons, we are conceptually guaranteed that there is decisive practical reason to do whatever morality requires us to do. So even if what morality requires us to do does not correspond to any morality-independent intuitions about what there is practical reason to do, as might be true according to moral theories like Rossian Pluralism (see, e.g., Ross 1930), we have a perfectly good explanation of why we have conclusive practical reason to do these things.

I suspect that a great deal of the attraction of some moral theories, like Contractarianism, Kantianism, Contractualism, and Consequentialism, relies upon their apparent ability to explain why we should care about morality and do what morality requires by appealing to substantive morality-independent intuitions about practical reasons. I have in effect argued that these explanations of why we should be moral are misguided. As such, I believe

25 See Nye (2015).

that my conceptual account of the connection between morality and practical reasons may significantly diminish the case in favor of these moral theories, and indirectly contribute to the relative attractiveness of moral theories like Rossian Pluralism, which, if we remove the desideratum that a moral theory should correspond to a substantive account of why we should be moral, may give a more plausible account of the content of morality.

References

Brandt, Richard B. (1959). *Ethical Theory*. Englewood Cliffs, NJ: Prentice Hall.

Bratman, Michael (1987). *Intention, Plans, and Practical Reason*. Cambridge, MA: Harvard University Press.

D'Arms, Justin and Daniel Jacobson (1994). "Expressivism, Morality, and the Emotions." *Ethics*, 104: 739–65.

——— (2003). "The Significance of Recalcitrant Emotion (or, Anti-Quasijudgmentalism)." *Philosophy: The Journal of the Royal Institute of Philosophy* 52 (suppl.): 127–45.

——— (2009). "Demystifying Sensibilities: Sentimentalism and the Instability of Affect." In *The Oxford Handbook of Philosophy of Emotion*, edited by P. Goldie. Oxford: Oxford University Press.

Darwall, Stephen (1997). "Reasons, Motives, and the Demands of Morality." In *Moral Discourse and Practice*, edited by S. Darwall, A. Gibbard, and P. Railton, 305–12. New York: Oxford University Press.

——— (2006). "Morality and Practical Reason: A Kantian Approach." In *The Oxford Handbook of Ethical Theory*, edited by D. Copp, 282–320. Oxford: Oxford University Press.

Ewing, A. C. (1939). "A Suggested Non-Naturalsitic Analysis of Good." *Mind* 48: 1–22.

Falk, W. D. (1948). "'Ought' and Motivation." *Proceedings of the Aristotelian Society*, 48: 111-138.

Foot, Philippa (1959). "Moral Beliefs." *Proceedings of the Aristotelian Society* 59: 83–104.

Gauthier, David (1986). *Morals by Agreement*. Oxford: Clarendon Press.

Gibbard, Allan (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Oxford: Clarendon Press.

Greenspan, Patricia (1988). *Emotions and Reason: An Inquiry into Emotional Justification*. London: Routledge & Kegan Paul.

Hieronymi, Pamela (2005). "The Wrong Kind of Reason." *Journal of Philosophy* 102: 437–57.

Kavka, Gregory S. (1983). "The Toxin Puzzle." *Analysis* 43: 33–36.

Kitcher, Philip S. (1998). "Psychological Altruism, Evolutionary Origins, and Moral Rules." *Philosophical Studies*, 89: 283–316.

Korsgaard, Christine (1996). *The Sources of Normativity*, edited by O. O'Neill. New York: Cambridge University Press.

Mill, John Stuart (1863). *Utilitarianism*. London: Parker, Son, and Bourn.

Nye, Howard (2009). "Ethics, Fitting Attitudes, and Practical Reasons." PhD Dissertation, University of Michigan.

——— (2015). "Directly Plausible Principles." In *The Palgrave Handbook of Philosophical Methods*, edited by C. Daly, 610–36. London: Palgrave MacMillan.

Parfit, Derek (2001). "Rationality and Reasons." In *Exploring Practical Philosophy: From Action to Values*, edited by D. Egonsson, B. Petersson, J. Josefsson, and T. Ronnow-Rasmussen. Aldershot: Ashgate.

Portmore, Douglas (2011). *Commonsense Consequentialism: Wherein Morality Meets Rationality*. Oxford: Oxford University Press.

Railton, Peter (1986). "Moral Realism." *The Philosophical Review* 95: 163–207.

Raz, Joseph (2009). "Reasons: Practical and Adaptive." In *Reasons for Action*, edited by D. Sobel and S. Wall. Cambridge: Cambridge University Press.

Roberts, Robert C. (1988). "What an Emotion Is: A Sketch." *The Philosophical Review* 97: 183–209.

Ross, W. D. (1930). *The Right and the Good*. Oxford: Clarendon Press.

Sabini, John, and Maury Silver (1982). *Moralities of Everyday Life*. Oxford: Oxford University Press.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge: Harvard University Press.

Skorupski, John (1999). *Ethical Explorations*. New York: Oxford University Press.

——— 2010. *The Domain of Reasons*. Oxford: Oxford University Press.

Solomon, Robert C. (1976). *The Passions*. Garden City, NY: Doubleday/Anchor.

——— (1988). "On Emotions as Judgments." *American Philosophical Quarterly* 25: 183–91.

Velleman, J. David (2002). "Motivation by Ideal." *Philosophical Explorations*, 5: 90–104.

Zimbardo, Philip G., and Ann L. Weber (1997). *Psychology*, 2nd ed. New York: Longman.

III

EXPRESSIVISM, NORMATIVE LANGUAGE, AND SEMANTICS

8

EXPRESSIVISM WITHOUT MINIMALISM[1]

Tristram McPherson

Introduction

Expressivism is one of the most influential research programs in contemporary metaethics. On first approximation, the expressivist claims that indicative normative sentences such as 'I ought to resist racism' semantically express psychological states that are broadly desire-like, where indicative nonnormative sentences such as 'I am late for class' semantically express beliefs.[2] It is a striking fact that many of the most recently influential expressivists in metaethics have embraced minimalist accounts of words such as 'truth,' 'fact,' and 'property.'[3] Roughly, minimalism is the thesis that the meaning of each of these words is exhausted by certain equivalences, which deprive these words of the distinctive metaphysical significance they are sometimes taken to have.

Expressivism and minimalism are distinct hypotheses about different parts of language. An expressivist might embrace certain minimalist theses simply because she finds those

1 I am grateful for illuminating discussion of issues related to this paper at Latrobe University, the Melbourne Moral Rationalism Workshop, and the University of Melbourne. The last section of the paper benefited from discussion of material extracted from that section on the Pea Soup blog. Additional thanks to Derek Baker, Gunnar Björnsson, Billy Dunaway, David Faraci, Eric Hubble, Robert Kraut, David Plunkett, Michael Smith, Pekka Vayrynen, and an anonymous referee for this volume, for comments and conversations that improved my understanding of the issues in this paper.

2 In this paper, single quotation marks (e.g., 'cat') are used strictly to mention linguistic items. Double quotation marks (e.g., "cat") are used for a variety of tasks including quoting others' words, scare quotes, and mixes of use and mention. Terms in small caps (e.g., OUGHT) pick out concepts.

3 See, for example, Blackburn (1993), Gibbard (2003, 2012), and Timmons (1999).

theses independently plausible. Or she might be interested in exploring the consequences for expressivism of embracing such minimalism (compare Allan Gibbard's cautious embrace of minimalism about truth (2003, 18, 182–3)). However, embracing minimalism is also widely taken to help make expressivism more credible as an account of our actual normative thought and talk.[4] Indeed, some philosophers have recently suggested that marrying expressivism to minimalism is crucial to the plausibility of the expressivist's research program.

This paper argues that this rationale for marrying expressivism to minimalism is a mistake: expressivists have weak dialectical reasons to marry their view to minimalism and strong reason to avoid the marriage. I begin by more carefully introducing the would-be spouses (§1) and the case for marriage that I aim to rebut (§2). I then argue that the costs to the expressivist of this marriage are higher than has typically been appreciated (§3). This would cast doubt on the viability of the expressivist research program, if expressivism were implausible without minimalism. I argue that it is not, in two stages (§4). First, I motivate the idea that every account of normative thought and talk will likely have to abandon some of what ordinary speakers tend to associate with such thought and talk. I then sketch a schematic theory of error for some of the sorts of data that have seemed to force the expressivist to embrace minimalism. It is worth emphasizing that this paper focuses on expressivism about normative thought and talk. However, expressivist accounts have been proposed for a variety of parts of language (such as epistemic modals, probability ascriptions, and epistemic and semantic talk).[5] To the extent that such accounts face similar dialectical pressures, some of the lessons of this paper may generalize to those accounts.

1. Noncognitivism, Expressivism, and Minimalism: What They Are

Suppose that you notice growing racism in your community, and deliberate about how to respond. You might determine along the way that morality requires you to resist the racism, while prudence dictates that you do not: resistance is (unfortunately) risky. In response to this perceived conflict, you might conclude your deliberation by thinking: *I ought to resist this racism*. The OUGHT concept that you deploy in this thought appears to wear a distinctive normative authority on its sleeve.[6] In Gibbard's phrase, this sort of thought—unlike narrowly moral or prudential ought-thoughts—constitutes a distinctive "flavorless endorsement" of an action (1990, 7). One striking feature of thoughts like this one is that we expect them to have an especially intimate connection to action. Studied practical indifference to this conclusion is more puzzling than parallel indifference to conclusions about what the law, or

4 Blackburn (1993, 3–5) embraces both the independent plausibility and the importance for expressivism of ideas related to what I am calling minimalism.

5 For example, see Field (1998), Gibbard (2003, Ch. 11; 2012), and Yalcin (2011, 2012).

6 For a more detailed discussion of this OUGHT concept, and a defense of a cognitivist-friendly gloss on it, see McPherson (2018).

etiquette, or even morality require.[7] Noncognitivism provides an elegant explanation of this striking feature of our actual OUGHT thoughts.

My preferred way of understanding noncognitivism begins by dividing psychological states according to their functional roles. The crucial division is well-introduced by an example lightly adapted from Elizabeth Anscombe (1957, 56): imagine a detective hired to follow a shopper around a grocery store and record what he purchases. Suppose that both the detective and the shopper have lists of groceries. The shopper's list functions to guide his behavior: when all goes well, if 'butter' is on the list, he puts butter in his cart. The shopper's aim is to make the contents of his cart conform to the contents of his list. The detective's list, by contrast, functions to represent part of the world (the contents of the cart): when all goes well, if there is butter in the cart, the detective adds 'butter' to her list. Her aim is to make her list conform to the contents of the shopper's cart.

The functionalist proposes that psychological states can be understood in similar functional terms. A *cognitive* or belief-like psychological state is one that functions like the detective's list: when all goes well, one will believe that there is butter in the shopper's cart only when there is. A *noncognitive* or desire-like psychological state functions like the shopper's list: when all goes well, the desire to put butter in the cart will tend to cause one to put butter in the cart. Part of all's going well here is the presence of true relevant means-end beliefs. For example, the belief that picking up the butter and dropping it this way will get it into the cart. If the shopper instead believes that the way to get butter into the cart is to throw it at the wall, he will not tend to get butter in his cart, despite his desire. (For the locus classicus of this sort of functionalism in the metaethical context, see, e.g., Smith 1994a, 10, 111–29).[8]

As I will understand it, noncognitivism is the thesis that normative thoughts such as *I ought to resist this racism* have the noncognitive functional role. That is, at least in paradigmatic cases,[9] they function to guide our behavior, rather than to represent the world.[10]

---

7 The contrast with morality is especially clear for someone who believes (perhaps incorrectly) that genuinely moral standards are normatively insignificant, perhaps because they believe that one ought always to be prudent and morality does not always pay, or because they think that moral standards themselves are best understood as pernicious in some way.

8 This paragraph and the last largely duplicate a portion of Faraci and McPherson (2017).

9 For an argument for the qualification that concludes this sentence, see Björnsson and McPherson (2013). This qualification is inessential to the discussion of this paper.

10 The term 'noncognitivism' has been used in several ways distinct from this one. For example, Horgan and Timmons (2006) offer a holistic characterization of cognitivism, which pulls it apart from the functional properties mentioned in the text. Importantly for this paper, noncognitivism has sometimes (especially in less recent work) been thought to involve a denial of *fact-statingness* or *truth-assessability*, for example, in Gibbard (1990, 8) and Smith (1994b, 1), respectively. For our purposes, it is important to set aside this understanding, which would make noncognitivism trivially incompatible with minimalism about truth- and fact-talk. In my view, the characterization of noncognitivism offered in the text is the most useful because it carves at the joints of relevant contemporary metaethical debates. While Gibbard does not endorse the way of characterizing noncognitivism suggested here (cf. also 2003, 183–84), he does endorse a strong modal connection between normative judgment and motivation (2003, 152–58). For a helpful recent

According to the noncognitivist, this functional role is what explains the striking feature of normative thoughts mentioned above: the especially intimate connection to action we expect such thoughts to have. One compelling example of noncognitivism as I am understanding it is developed in Gibbard's (2003), which identifies normative thoughts as *planning states*.

Pure expressivism is a broad theory about the relationship between normative thought and talk.[11] On this view, normative sentences bear a systematic expression relationship to noncognitive mental states. And the meanings of normative sentences are identical to, or grounded in, the mental states they bear this expression relation to.[12]

Together, noncognitivism and pure expressivism provide the heart of a powerful research program for understanding our actual normative thought and talk. (Contrast the *revisionary expressivist's* project of advocating for revising our normative thought and talk to make it expressivist.[13]) Rather than talking about 'non-cognitivist pure expressivism about actual normative thought and talk' in what follows, I refer to this research program simply as 'expressivism.'

Some influential recent expressivists have tended to wed their expressivism to *minimalist* theses about certain of our vocabulary.[14] I will introduce the minimalist idea by illustration, focusing on the most familiar example: the sentential truth predicate. To begin, consider:

---

discussion of noncognitivism, see Bedke (2017); what I call 'noncognitivism' corresponds to what Bedke calls 'psychological noncognitivism.'

11 In this paper, I set aside hybrid forms of expressivism (for discussion, see, e.g., Ridge 2014; Schroeder 2009; Toppinen 2013, 2017). Many such views avoid commitment to minimalism, seeking to achieve similar explanatory aims in different ways.

12 This characterization of expressivism is close to that offered by Camp (2017, 87). It allows that expressivism can be developed both as a semantic thesis (a systematic thesis about what the meanings of normative sentences are) and as a metasemantic thesis (a thesis about what grounds the facts about what the meanings of normative sentences are). For the metasemantic gloss, see Chrisman (2012, Sec. 4) and Ridge (2014, Ch. 4). My gloss on expressivism is silent on the semantics of nonnormative sentences. However, there are strong reasons to incorporate expressivism within a unified theory of both normative and nonnormative language (Schroeder 2008, 5). A natural way to do this is to insist on a mentalist (meta-)semantics across the board: the idea is that the meanings of both normative and nonnormative sentences are identical to, or grounded in, the mental states they express. On this way of developing expressivism, the crucial difference between normative and nonnormative sentences is that the former express desire-like states, while the latter express beliefs (compare Gibbard 2003; Schroeder 2008).

13 For discussion of revisionary expressivism, see, for example, Köhler and Ridge (2013) and Svoboda (2017).

14 This paper could have been framed in terms of the relationship between expressivism on the one hand and either *anti-realism* or *quasi-realism* on the other. If one rejects minimalism about truth- and fact-talk, it will be natural to conclude that (noncognitivist, pure) expressivism entails that there are no normative truths or facts. And this is naturally described as an anti-realist view. By contrast, wedding expressivism to minimalism allows the expressivist to vindicate as felicitous many realist-sounding assertions about the normative (such as the assertion that there are normative facts). It is common to follow Simon Blackburn's infectious terminology and call this *quasi-realism* (e.g., his 1993). However, since the term 'quasi-realism' is very closely associated with Blackburn's specific version of the sort of research program I am exploring, I will stick to talking about the relationship between expressivism and minimalism. (Compare Gibbard 2003, 18–19, for a prominent example of associating 'quasi-realism' specifically with Blackburn's program.)

Trivial The sentence 'Cats are mammals' is true just in case cats are mammals.

Trivial first *mentions* a sentence of English—'Cats are mammals'—and then *uses* that same sentence. It states an equivalence between the truth of the mentioned sentence and what the use of the sentence expresses. Trivial is hard to intelligibly deny. And it does not appear to depend upon anything special about the sentence 'Cats are mammals.' This has led many philosophers to be attracted to a universal generalization of this sort of claim. The idea is that every well-formed indicative sentence 'S' satisfies the following schema:

T-Schema 'S' is true just in case S.

Because of its role in certain semantic paradoxes, the universal applicability of this schema is *not* trivial. (I return to this in §§3 and 4 below.)

To help bring out the distinctive character of minimalism, it will be useful to contrast it with a familiar competitor, the correspondence theory. Suppose that we accept the universal generalization concerning T-Schema. Consider the simplest form of a correspondence theorist's explanation of *why* this generalization is true. She suggests that when we *use* an indicative sentence, we attempt to *describe* (part of) how the world is. For example, one interesting feature of the world is that the species *felis catus* belongs to the class *mammalia*. When we use 'cats are mammals' on the right-hand side of Trivial, we describe this feature of the world. In virtue of its meaning, the English sentence 'cats are mammals' is *about* that very feature of the world, and the truth of that sentence consists in the *correspondence* between what that sentence means and this feature of the world. The correspondence theorist insists that similar explanations hold of truths generally.

The minimalist denies this account, at least as an illuminating explanation of the general applicability of T-Schema.[15] Instead, the core minimalist idea is that T-Schema itself (perhaps together with similar treatments of other contexts in which 'true' occurs) provides a complete *analysis* of the meaning of the predicate 'true.' However, this is not enough to secure minimalism, because a word can refer to a kind whose nature is not revealed by the analysis of the word. For example, 'water' refers to a kind that has a distinctive molecular nature. A correspondence theorist could insist analogously that correspondence was part of the nature of truth. To rule this out, the minimalist must insist that the nature of truth itself is exhausted by what can be gleaned from the T-Schema (for this point, compare Cuneo 2013, 231).[16]

---

15 This qualification is important, because the minimalist may find herself in a position to sincerely utter the correspondence theorist's speech if she accepts minimalism about sufficient further vocabulary. Cf. §3.3 below.

16 Like 'noncognitivism,' 'minimalism' is used to refer to many different views. One important such view is simply the view that no metaphysical account of truth *follows from the analysis of* 'true' (cf. Soames 1999, 231, on this sort of "deflationism"). However, if truth had a hidden nature (like water), then minimalism in

It might seem puzzling that we would have a truth predicate if its meaning was so minimal. However, minimalists point out that such a predicate can simplify communication. For example, suppose that I want to communicate that for every indicative sentence '*S*', if Sally says '*S*', then *S*. I can more clearly and efficiently communicate this by uttering 'Everything Sally says is true.'

It is worth emphasizing at this point that the distinction between minimalist and correspondence theories of truth is not exhaustive. There are many other types of theory of truth, which compete with both of these classes of theories. The dialectical significance of this broader range of theories of truth is beyond the scope of this paper.

As the correspondence theory discussed above illustrates, 'true' is an example of vocabulary whose use can easily be understood as bringing metaphysical commitments with it. Minimalism, as I am understanding it, is one strategy for interpreting such vocabulary in a way that avoids ascribing such metaphysical commitments. Minimalist glosses can be offered for other seemingly metaphysically committal vocabulary, by following the same recipe: to identify certain schematic equivalences. For example:

Fact It is a fact that P just in case P.
Property x has the property of being F just in case x is F.

As with T-Schema, there is a reading of each of these equivalences on which their collected instances constitute striking facts in need of explanation. The minimalist instead claims that these equivalences themselves tell you all there is to know about the nature (or lack thereof) of facts and properties.

To sum up: expressivism is a distinctive family of views about the nature of normative thought, talk, and their relation to each other. Minimalism is a distinctive family of views about the nature of certain seemingly metaphysically significant vocabulary such as 'true,' 'fact,' and 'property.' It is an interesting and difficult question whether expressivism is true, and the same holds for various minimalist theses. This paper does not aim to answer either of these questions. Instead, my aim is more modest: to understand the dialectical relationship between these two views: specifically, is expressivism more or less plausible when married to minimalism? To begin, I explain why some prominent philosophers have thought that the expressivist is dialectically committed to this marriage.

---

this sense would be insufficient to make expressivism compatible with the felicity of claims like 'It is true that I ought to resist racism.' Another deflationist view that contrasts with the sort of minimalism spelled out here gives a broadly *expressivist* gloss on the metaphysical vocabulary itself (see, e.g., Kraut 1993). Some of the older literature concerning whether noncognitivism and expressivism are *compatible* with minimalism (Divers and Miller 1994; Horwich 1993; Smith 1994b; Wright 1992) is driven in part by ways of formulating both of these theses that contrast with the formulations given in the text. I largely ignore this literature in what follows. In the text, I adopt what I take to be the most dialectically relevant formulation.

2. The Case for Marriage

As I noted in the Introduction, many influential contemporary expressivists have embraced minimalism. Some philosophers have argued that embracing minimalism is crucial to expressivism's plausibility as a theory of our actual thought and talk. In this section, I distinguish two broad rationales for thinking that the marriage to minimalism makes expressivism more plausible. These appeal respectively to the need to explain certain plausible linguistic felicity judgments and to the desirability of preserving parts of apparent common sense.

First, note that most normative words appear to play perfectly ordinary syntactic roles in indicative sentences that include words like 'true,' 'fact,' 'property,' and so forth. This contrasts strikingly with other forms of words which we might be tempted to think function to express our noncognitive attitudes. Consider sentences like 'Hooray for bears!' or 'Shut the door!' You cannot felicitously reply to utterance of either of these sentences by saying 'That's true' or 'That's false' or 'That's a fact.' But you can felicitously reply in these ways to 'I ought to resist racism.' Such felicity data provide apparently powerful evidence for an account of normative thought and talk. Minimalism allows the expressivist to vindicate such felicity data. The minimalist can argue that the contrasting felicity data is explained by the fact that 'I ought to resist racism' possesses—and 'Hooray for bears!' lacks—the indicative syntactic structure required by the T-Schema for felicitous application of the truth predicate. The ability of minimalism to vindicate such felicity judgments provides the expressivist with what I call the *felicity rationale* for adopting minimalism.

Second, consider conjoining expressivism with nonminimalist theories of truth and fact. Absent some further explanation, this might appear to commit the expressivist to claims like *there are no normative facts*.[17] For if normative thoughts consist simply in certain desire-like states, the background functional account suggests that they do not *represent* any normative reality. But the denial of normative facts has potentially embarrassing instances, like *there is no fact of the matter about the wrongness of child abuse*. Some philosophers have thought that philosophy could never put us in a position to accept claims like these, which arguably fly in the face of common sense (for discussion, see McPherson 2009). By embracing minimalism, the expressivist avoids being forced to reject such seemingly commonsensical claims on the basis of her philosophical arguments. As Simon Blackburn memorably puts the idea, by embracing minimalist resources, he claims to vindicate the view that there is "nothing illegitimate in our ordinary practice and thought" (1993, 216). Call this the *commonsense rationale* for adopting minimalism.[18]

---

17 Whether there is such a commitment will depend in part on the relationship between facthood and the sort of representational function I associated with cognitivism in §1. Not all nonminimalist theories will connect these.

18 These motivations are not exhaustive. Consider, for example, a theoretical analogue of the commonsense motivation: one might become convinced that both expressivism and contemporary variants of metaethical realism have captured a crucial part of the truth about the normative, and that an expressivism-minimalism

The felicity rationale and the commonsense rationale are independent: for example, one can imagine an expressivist who accepts the felicity rationale but takes normative "common sense" to be a tissue of errors. And both rationales provide some apparent reason for the expressivist to embrace minimalism. To some philosophers, one or the other of these rationales might appear *decisive*. For example, many contemporary philosophers share the inclination that Kit Fine reports, to "doubt that philosophy is in possession of arguments that might genuinely serve to undermine what we ordinarily believe" (2001, 2). Since arguments for expressivism are presumably philosophical, this inclination would require that a credible expressivism avoids conflict with common sense. Similar views about the felicity rationale have been prominently endorsed. James Dreier notes that "To save the phenomena, then, expressivism needs a conception of truth that doesn't commit to anything metaphysically heavy-duty" (2010, 170). Dreier argues that in light of this point, it is *no dialectical cost at all* for the expressivist to appeal to minimalist resources in addressing an independent challenge (2010, §4.1). Huw Price (2015, §1.2) argues on very similar grounds that an anti-realist alternative to adopting minimalistic resources would constitute a *reductio* of expressivism. If these assessments were correct, the expressivist would face a shotgun wedding with minimalism. This would in turn make trouble for expressivism, because, as I now argue, the dialectical costs of marriage to minimalism are considerable.

3. Costs of the Marriage

This section introduces several costs to the expressivist of marrying her view to minimalism. I begin by suggesting some straightforward points, before turning to the significance of the now-familiar "creeping minimalism" challenge. I argue that the dialectical significance of this problem for the marriage of expressivism and minimalism has not been fully appreciated.

3.1 Basic Dialectical Costs

As I emphasized at the conclusion of §1, expressivism and minimalism are distinct families of views about different elements of thought and talk. Because minimalist theses are controversial, the expressivist would ideally hope to defend her view in a way that was maximally *neutral* concerning such theses. Marrying expressivism to minimalism fails to maintain such neutrality. To put the basic worry crudely, the credence one should have in the truth of the expressivism plus minimalism package is, roughly, the product of the credence it is reasonable to have in expressivism independently of issues about truth and the credence it is reasonable to have in minimalism. Because minimalism is a highly controversial claim

hybrid is the right way of integrating the insights of these views. I take it that this is one way of understanding Gibbard's (2011) talk of aiming to "emulate" or "mimic" a certain type of realism by use of minimalist resources.

about truth, the credence one should have in the package is likely small. (This crude worry is intended only to point the reader in the right direction. The idea itself is obviously too crude. For example, our evidence for expressivism and minimalism might overlap, and (for reasons sketched in the previous section) expressivism might be implausible given the rejection of minimalism.) To illuminate the costs of the marriage, I briefly illustrate a small part of the controversy concerning the most deeply explored minimalist thesis, that concerning the word 'true.'

On a brief introduction, truth minimalism can easily appear to be uncomplicatedly plausible. However, this should not mislead us when we seek to understand the commitments the expressivist takes on by endorsing minimalism. Summing up the history of technical work on truth and paradox, Timothy Williamson says:

> One clear lesson is that claims about truth need to be formulated with extreme precision . . . because in practice correct general claims about truth often turn out to differ so subtly from provably incorrect claims that arguing in impressionistic terms is a hopelessly unreliable method.
>
> (2007, 281)

In short, truth is an area of philosophy where the apparent plausibility of a claim, roughly sketched, is likely to be a poor guide.

Consider one example of this point. I presented minimalism about truth as claiming that the T-Schema tells us everything that we need to know about truth. It might seem obvious that because we all understand the T-Schema, an attractive aspect of minimalism is that it explains what our mastery of 'true' consists in. However, Anil Gupta has shown that this is far from clear, once we ask what generalization or inference rule we need to accept in order to have such mastery. On the one hand, such a rule must be adequate to explain our mastery of a minimalistically understood 'true.' On the other hand, it must be a rule that we plausibly possess. For example, we clearly don't need to know all of the acceptable instances of T-Schema in order to master 'true.' Gupta shows that it is extremely challenging to identify a mastery condition that meets these two constraints. So it is unclear that minimalism has an advantage over other theories of truth in illuminating mastery of 'true' (1993, §§IV–V).

The most familiar problem for the idea that the meaning of 'true' is exhausted by T-Schema is different. It arises straightforwardly from the fact that some instances of T-Schema are likely false. The unrestricted applicability of the T-Schema plays a role in generating the Liar and related paradoxes. Many philosophers have responded to these paradoxes in part by restricting the T-Schema.[19] If these philosophers are correct, it suggests that the nature of

19 The ur-strategy for restriction, due to Tarski (1958 [1935]), posits of hierarchy of metalanguages, and allows a truth predicate to apply only to sentences at a lower level in the hierarchy (the base language having no truth predicate). There are other replies to the semantic paradoxes, however, which retain an unrestricted

truth is *not* transparently available to anyone who reflects on the T-Schema. And this calls into question the core minimalist doctrine.

There is also reason to doubt that minimalism can adequately capture our ordinary commitments about 'true.' Recall that, according to the minimalist, truth is not an interesting property; rather, the word 'true' is best understood as a useful linguistic device. Next recall:

Trivial The sentence 'cats are mammals' is true just in case cats are mammals.

In contexts like Trivial, the minimalists suggest, 'true' simply allows us to map a certain relation between use and mention of a single well-formed indicative sentence. One worry about this idea is that the following sort of explanatory claim appears very plausible:

Explanatory The sentence 'cats are mammals' is true because: cats are mammals (together with the fact that the sentence 'cats are mammals' means that cats are mammals).

Daniel Stoljar and Nic Damjanovic call a similar thesis (absent the parenthetical addition) the "correspondence intuition" (2014, §7.2). It is not straightforward for the minimalist to vindicate Explanatory. One way to see this is to suppose that 'because' here claims a metaphysical explanatory relation, such as a grounding relation. There is a clear danger that such an explanation will be backed in a standard way, by an account of the nature of truth. (Compare: on a standard picture, every fact of the form [P&Q] is grounded by its conjuncts, because of the nature of conjunction.) And this looks like a natural way of developing a correspondence theory of truth.[20]

These brief sketches provide only a preliminary sense of a few of the many challenges to minimalism about truth.[21] My aim is not to suggest that these challenges cannot be met. Instead, I seek to make vivid the fact that minimalism about truth is a highly controversial thesis about a central philosophical topic.

Next consider the expressivist who has accepted minimalism about truth, and considers whether to adopt minimalism about other apparently metaphysical language. On the one hand, there are differences among the different pieces of seemingly metaphysical vocabulary, and among the most plausible equivalence schemas for each one. So we can expect that there will be interesting differences among the challenges facing minimalist theories of, for example, 'fact' or 'property.' Commitment to a conjunction of such minimalisms thus

T-schema. For an introductory discussion of the liar paradoxes and contemporary responses, see Beall et al. (2017).

20 Stoljar and Damjanovic suggest a different worry about the compatibility of minimalism and the correspondence intuition in their 2014, §7.2.

21 For surveys of the dialectic concerning truth-minimalism, see Armour-Garb and Beall (2005) and Stoljar and Damjanovic (2014).

commits the expressivist to the failure of each objection to each such minimalist view. This suggests that caution is warranted in generalizing one's embrace of minimalism.

On the other hand, there are at least three reasons that it is natural for the expressivist who has embraced minimalism about truth to adopt further minimalist interpretations. First, such further minimalist theses may seem to share the initial plausibility of truth minimalism. Second, such minimalisms are arguably very similarly motivated. To use Mark Johnston's apt phrase, the minimalist thinks that, quite generally, our ordinary thought and talk has "given no hostages to metaphysical fortune" (1992, 590). Further, adopting minimalism about at least some vocabulary beyond 'true' also satisfies the commonsense and felicity rationales for marrying expressivism to minimalism. For example, one can felicitously reply to a normative claim with 'That's a fact,' and it is arguably commonsensical that there is a fact of the matter about the wrongness of child abuse. These points suggest that minimalism about 'true' alone will fail to fully answer these rationales. In short, if the expressivist adopts minimalism about truth, she faces significant dialectical pressure to embrace minimalist accounts of other apparently metaphysical vocabulary.

3.2 Creeping Minimalism

If minimalism is correct about apparently metaphysical vocabulary quite generally, this raises a further complication for the expressivist program. This is whether "creeping minimalism" robs the expressivist of the ability to distinguish her view from competing views, and thereby undercuts the distinctive appeal of the expressivist approach.

Much of the distinctive appeal of the expressivist research program lies in the combination of three apparent virtues. First, as emphasized in the Introduction, we take normative judgments to play a distinctive practical role. We expect that if I think *I ought to resist racism*, then (other things being equal) this will tend to structure my actions to oppose racism. By contrast, it is an unfortunate but familiar fact that a nonnormative belief like *there is a lot of racism in my community* can be combined with enthusiastic embrace of such racism or indifference, as easily as it can with opposition. The expressivist's characteristic noncognitivism about ethical judgment appears able to explain this sort of distinctive tie between ethical thought, motivation, and the production of action (see Faraci and McPherson 2017 for discussion). Second, it appears that possession of OUGHT is compatible with a striking variance in facts about what use of this concept appears to track, both across individuals and linguistic communities, and it appears that *genuine disagreement* is possible among those who possess OUGHT (Björnsson and McPherson 2014; Gibbard 1990; Horgan and Timmons 1991). Many have thought that the expressivist is well-poised to explain these facts about concept possession and disagreement. Third, the expressivist program promises to offer a broadly *naturalistic* explanation of ethical thought and talk, requiring neither commitment to mysterious 'nonnatural' properties, nor puzzles about how we could know—or even think about—such properties (for a more detailed discussion of naturalism in the metaethical context, see McPherson 2015).

With a wide range of minimalist theses in hand, however, the expressivist may find that her normative views commit her to a striking range of realist-sounding theses. For example, the expressivist may find herself committed to there being mind-independent, irreducible ethical facts. And now the contrast between expressivism and views that do not appear to share the appealing features just mentioned—such as nonnaturalistic realism—might seem to be threatened.[22] This problem of "creeping minimalism" is arguably a problem for everyone who embraces minimalism about a wide range of metaphysical vocabulary, but here I am focused on its significance for the expressivist.

Various ways of saving the differences have been proposed, but none are without difficulties.[23] However, even if expressivism can be distinguished from (e.g.) nonnaturalistic realism, it is not clear that this contrast will suffice to protect the minimalist expressivist from the sorts of epistemological and metaphysical challenges that many take to render the nonnaturalist's position unattractive. For example, once committed to talk of normative properties, I might find it difficult to explain why my normative judgments reliably represent them, and why those normative properties supervene on the nonethical properties (for discussion, see Dreier 2012, 2015; Gibbard, 2011; Golub 2017b; Street 2011).

3.3 The Dialectical Implications of Stopping the Creep

In this section, I briefly explain two ways in which the problem of creeping minimalism complicates the apparent virtues of marrying expressivism with minimalism, even if the problem of creeping minimalism can be solved in a way that retains the distinctive appeal of expressivism.

In order to see the first problem, consider the broad strategy of stopping the creep suggested by Gibbard and Dreier. In Gibbard's words, the strategy is to explain belief in normative facts, "without helping ourselves to normative facts at the outset" (2003, 183). In Dreier's exposition of this strategy, the idea is to focus on constitutive explanatory claims, such as claims about what it is to think an action is wrong. The idea is that the expressivist and nonnaturalistic realist will offer contrasting patterns of constitutive explanation. On this gloss, roughly, the expressivist but not the realist can insist that there is nothing to being a normative thought over and above a certain noncognitive functional role (Dreier 2004, 39).

This strategy blocks the minimalist interpretation in only one place—the context of constitutive explanatory claims. One virtue of this strategy is that it allows the expressivist to embrace minimalism about a wide range of apparently metaphysical vocabulary. As noted

22 Notice that these apparent metaphysical commitments can be limited in two ways, exemplified by Gibbard (2003). First, Gibbard only embraces minimalism about some metaphysical vocabulary. Second, he appeals to further metaphysical theses—such as an intensional criterion of property identity—that limit the apparent ontological and ideological commitments of his account. Even so, many commentators have taken Gibbard to be committed to realism.

23 See, for example, Asay (2013), Dreier (2004, 2018), Dunaway (2016), Gibbard (2003, 184–93; 2012, Ch. 10), Golub (2017a), and Simpson (2018).

above, the motivations for such minimalisms overlap, so this is in one respect an attractive result. For example, it is hard to deny that 'It is an objective fact that I ought to oppose racism' is a felicitous English sentence. It is a virtue of the marriage that it allows the expressivist to vindicate this apparent fact.

In another respect, however, this way of stopping the creep is awkward for the minimalist expressivist. Consider the fact that it is natural to develop and motivate expressivist views as *competitors* to views that claim that there are objective, mind-independent facts about what we ought to do.[24] This fact arguably casts doubt on minimalism as an interpretation of our apparently metaphysical thought and talk. If the folk are all implicitly minimalists, why is it so natural to use folk metaphysical talk to attempt to discuss metaphysically committal views? Setting aside its implications for the credibility of minimalism, the *expressivist* might have hoped to block the creep before it vindicated trivial inferences to claims that are natural ways of describing competing views.

My second worry about the significance for the expressivist of creeping minimalism is less familiar. The worry is that—once the creep is stopped—this will weaken the commonsense and felicity rationales for embracing minimalism. I will assume Dreier's way of stopping the creep in what follows, but I take the point to generalize.

Recall the functionalist account of cognitivism and noncognitivism in §1. On this account, cognitive states (but not noncognitive states) bear a certain functional relationship to their contents. This relationship is like the relationship of Anscombe's detective's list (and not the shopper's list) to the contents of the cart. Call this broad functional relationship the *represent*$_F$ relation. On expressivism, we can extend this relation to sentences: sentences that express cognitive states represent$_F$ the contents of those states.

This puts me in a position to introduce a functionally characterized truth predicate by stipulation:

Functional What it is for a sentence 'P' to be true$_F$ is for it to be the case (1) that the sentence 'P' represents$_F$ that P and (2) that P.

Functional is, in essence, just a stipulatively introduced correspondence truth predicate.

According to the expressivist, normative indicative sentences express noncognitive states. And these states do not represent$_F$ anything. So, no indicative normative sentence will be true$_F$.

Next, consider the range of folk speakers who are disposed to assert sentences of the form: 'It is true that I ought to do A.' Imagine equipping those speakers with substantial understanding of true$_F$. How many of those speakers would also assent to parallel sentences

24 Gibbard (1990) is an especially powerful instance of this approach. Note that if one was instead motivated by the idea that there were important truths in realism, that it was important for the expressivist to capture or mimic, then this might not be a problem at all.

of the form 'It is true$_F$ that I ought to do A'? I conjecture that the answer is: *a lot*. Indeed, I conjecture that many (most?) of them would take themselves to have been asserting ordinary normative truth claims (and finding them commonsensical) all along *because* they implicitly took themselves to be asserting (and finding commonsensical) the related functionalist truth claims.

One piece of evidence for this is the plausibility of an explanatory claim that exemplified one of the difficulties for minimalism about truth mentioned in §3.1 above. This claim can be generalized this way:

Generalized Explanatory For all true sentences 'S,' 'S' is true because (1) S, and (2) the meaning of 'S' (where (2) alone may be explanatorily sufficient in some cases).

Note that if we implicitly assume that 'true' were synonymous (or close to synonymous) with 'true$_F$,' this would neatly explain why we find Generalized Explanatory compelling. For it would suggest that the truth of a sentence is a matter of a match between the representational purport of that sentence (which I assume here to be a consequence of its meaning) and what the sentence purports to represent.

The plausibility of Generalized Explanatory might be evidence against minimalism about truth. However, I am making a different point here, which holds even if minimalism about truth is correct. This begins with the conjecture that, for many ordinary speakers, the practice of asserting normative truth claims, and taking them to be felicitous, is tied to (implicit) assumptions incompatible with expressivism, such as the truth$_F$ of certain normative sentences. Given minimalism, expressivism can explain why such truth claims are felicitous. But it is unclear whether this makes expressivism more plausible, if nonminimalist assumptions largely explain why such truth claims are being uttered and taken to be felicitous in the first place. Compare: an expressivist theory of the semantics of 'God' might be able to account for the range of felicitous utterances using this word. But it might nonetheless be a poor theory of actual theological thought and talk, if what almost always explains why people say things like 'God exists' is that they believe a certain supernatural entity exists.

To conclude this section, let us review the costs to the expressivist of marriage to minimalism. First, to the extent possible, there are principled dialectical reasons for the expressivist to avoid saddling herself with controversial commitments about topics beyond the core commitments of her view. Minimalism about truth is such a controversial commitment. Second, the rationales for minimalist arguments generalize, introducing the threat of creeping minimalism. The creep must be stopped, on pain of collapsing differences between expressivism and realist views that expressivists typically take to be implausible. Third, it is unclear whether the creep can be satisfactorily stopped, and the best proposal for doing so stops it far beyond the place where the expressivist might hope it would be stopped, committing the minimalist expressivist to a host of realist-sounding claims. Fourth, some have

argued that even if the creep can be stopped, it will still saddle the expressivist with explanatory burdens parallel to those faced by her most salient metaethical opponents. And finally, if we can stop the creep, this will permit us to introduce nonminimalist versions of the relevant metaphysical vocabulary. Doing so will plausibly reveal that much of the thought and talk that the marriage is supposed to accommodate turns out to reflect commitments inconsistent with expressivism.

4. Expressivism without Minimalism

I have just been emphasizing the dialectical costs to the expressivist of marriage to minimalism. However, as I explained in §2, there are significant reasons for thinking that divorce will come at the unacceptable cost of failing to make expressivism compatible with elements of commonsense, obvious felicity judgments, or both. Together, this might seem like it constitutes an argument against the plausibility of expressivism: the expressivist must embrace minimalism, but doing so comes at a striking cost to the plausibility of the view.

In this section, I argue that it would be a mistake to jump to this conclusion: expressivists can afford to reject minimalism. I argue in two stages. First, I show that normative thought and talk present philosophers with a hard interpretive problem. Hard interpretive problems motivate the search for theories of error concerning some of the considerations that generate them. I then sketch an appealing theory of error concerning the sorts of considerations that can seem to force the expressivist to accept minimalism.

Normative thought and talk presents philosophers with what I will call an *interpretive* problem: the phenomenon of normative thought and talk is all around us, and the problem is how to best understand it. Contrast this with philosophical questions like: *Does God exist?* or *Are properties part of the fundamental structure of reality?* Even if interpretive work is important in answering these questions, the questions themselves do not concern how to understand our own thought and talk. An interpretive problem is hard, in the sense I am after, if skillful and reasonable investigators do not tend to converge on an understanding of the phenomenon, despite largely sharing the available relevant evidence.

Normative thought and talk present us with an interpretive problem that is hard in the sense just glossed. Consider one familiar dimension of the considerations that make it so. On the one hand, as I have explained in this paper, much of what we know about normative judgments suggests that they are noncognitive. On the other hand, normative judgments appear to behave in many ways like ordinary, context-insensitive beliefs. For example, for most of us, the epistemology of the normative appears to be the attempt to *discover* something independent of our own perspective, as opposed to the attempt to *introspect* or to *decide*. One vivid way of bringing this out is that it seems appropriate to many of us to worry that we might be *mistaken* about fundamental normative matters, in ways that merely improving our coherence would not correct (Egan 2007; Köhler 2014). Further, normative disagreements seem *substantive* in ways that disagreements that might be thought

to express contrasting noncognitive attitudes do not. (E.g., "Broccoli is tasty." "No it isn't.")[25] Finally, of course, there is the very data that the marriage seeks to accommodate: normative terms behave felicitously and commonsensically in ordinary discourse much like terms that we ordinarily take to be associated with cognitive contents, and to enter felicitously into sentences featuring words like 'true' or 'fact.'

It is not easy to provide a consistent theory that offers a satisfying explanation of these and other important apparent features of normative thought and talk. Evidence for this difficulty is provided by the practice of metaethics. It is not difficult to get competent users of normative concepts puzzled by the problem of how to interpret normative thought and talk. And once puzzled, they reach for a wide variety of different solutions. If the folk were straightforwardly and smoothly interpreted as unconfused minimalist expressivists, it would be hard to see why alternative views would tend to find traction. And such diverse reactions are not restricted to the barely tutored: philosophers have been systematically investigating normative thought and talk for generations, without marked evidence of convergence.

The fact that normative thought and talk present a hard interpretive problem is compatible with the existence of an interpretation of such thought and talk that completely vindicates everything we tend to find plausible about the topic. However, it does make the existence of such an interpretation seem highly unlikely.

Note next that with a noninterpretive philosophical question, such as whether properties are part of the fundamental structure of reality, persistent intelligent disagreement might be explained by the fact that our evidence concerning this question is impoverished. By contrast, it is much less plausible that contemporary metaethicists have impoverished evidence concerning the nature of their own normative thought and talk. It is more plausible that the persistence of intelligent disagreement concerning normative thought and talk is partly explained by the fact that some of our apparent evidence about normative thought and talk is misleading.

This methodological observation does not license simply rejecting whatever apparent evidence is inconvenient to one's favored theory. Rather, that a theory rejects any initially plausible feature of normative thought and talk constitutes a defeasible dialectical cost to that theory. However, we can mitigate such costs by providing a principled *theory of error*: a theory that purports to explain why some of the claims that help to generate the interpretive problem are accepted despite being false. The methodological observation just offered makes it credible that some such theory of error is true in metaethics, and thus motivates the search for such a theory.

In what follows, I briefly sketch one such theory of error, which provides the expressivist with an attractive way of undercutting the felicity rationale for embracing minimalism. (I take it to be possible to extend the theory to also undercut the commonsense rationale, but

25 For one ambitious attempt to develop this into an argument for realism, see Enoch (2011, Ch. 2).

I do not do so here.) I begin by introducing a different theory of error, which serves as a useful model for the theory I offer on behalf of the expressivist.

As with normative thought and talk, alethic thought and talk plausibly presents a difficult interpretive problem, with many competing pressures and accounts. To see one dimension of the problem, consider the Liar sentence:

Liar The sentence named *Liar* is not true.

Liar is associated with a *paradox*: there are seemingly compelling arguments focused on this sentence (and a collection of related sentences) that conclude in contradiction. Matti Eklund (2002) defends two central claims about this paradox:

(1) The false premise in the paradoxical arguments is the assumption that all instances of T-Schema are true.
(2) We are disposed to accept this false premise because we are (defeasibly) disposed to accept arbitrary instances of T-Schema in virtue of being competent with 'true.'

If Eklund's first central claim is correct, all instances of T-Schema being true of an appropriately expressively rich language would render that language *inconsistent*. If his second claim is correct, we are (defeasibly) committed to this inconsistency-making claim by our semantic competence.

I reject Eklund's second claim: I don't think we should take competence *per se* to dispose users to accept falsehoods.[26] To see why, consider an attractive alternative to Eklund's claim (2):

(2*) We are disposed to accept the false premise that generates the Liar because the disposition to accept any instance of T-Schema is ordinarily strongly and nonaccidentally correlated with competence with 'true.'

If true, (2*) does what Eklund wants (2) to do: it purports to explain why so many of us find plausible the false premise of the Liar paradox. But (2*) is weaker than (2) precisely in not making competence itself the grounds for the disposition to accept falsehoods.

Thesis (2*) can in turn be explained by a plausible conjecture about how we come to be competent with 'true.' On this conjecture, we typically learn to use terms felicitously by encountering others' presumptively felicitous uses of those terms, and by getting felicity feedback on our own utterances. While doing so, we form something like an implicit model of how to use the relevant terms felicitously (perhaps aided by innate structure that narrows

26 Eklund himself notes relevant complexities at 2017, 90–91.

the range of candidate hypotheses). For those of us who do not wrestle with how to interpret sentences like Liar at an early age, T-Schema is plausibly a *highly elegant* inference schema to acquire for 'true.' If this is right, the process by which almost all of us become competent with 'true' also disposes us to accept arbitrary instances of T-Schema. This explains why, even if such a disposition is not necessary for competence with 'true,' it is nonaccidentally correlated with competence with 'true,' given how that competence is ordinarily acquired.

Why prefer (2*) over (2)? Because plausibly, one can be competent without possessing the relevant disposition. Consider a parent who comes to reject the universal applicability of T-Schema, in virtue of its paradox-inducing role when applied to sentences like Liar. Such a speaker will believe that T-Schema is correctly applicable to only some restricted range of sentences. If that speaker then taught her child to speak English in a way that made the Liar paradox salient very early on, I find it extremely plausible that the child could become fully competent with 'true' without having any disposition to accept the paradox-inducing instances of T-Schema. So I don't think that competence with 'true' requires such a disposition.

Together, (2*) and its explanation provide a natural *theory of error* for the Liar paradox. It explains how a natural story about the normal ways we become competent with certain vocabulary could involve our acquiring dispositions to form false semantic beliefs about that very vocabulary. This discussion is useful in two ways. First, it shows that we can offer plausible theories of error for semantic beliefs: we should not simply assume that these are the least vulnerable of our apparent metaethical evidence.[27] Second, it provides a model for the theory of error I propose on behalf of the expressivist. Like the disposition to accept arbitrary instances of the T-schema, I will suggest that the disposition to accept the truth-aptness of normative sentences is an *easily acquired overgeneralization*.

To begin developing this theory, consider a familiar and compelling motivation for views like expressivism: normative thought and talk is functionally powerful as a tool for planning and social coordination.[28] The expressivist suggests that noncognitive states are the crucial psychological locus for such a tool, for we want our planning and coordination to issue smoothly in *action*, and noncognitive states are functionally crucial for the production of action.

In order for normative thought and talk to play these roles, it is plausible that there must be structurally rational relations among normative thoughts and implication relations among normative sentences. To see why, imagine that there were no such relations.

27 On the Pea Soup blog (McPherson et al. 2017), Frank Jackson points out that indicative conditionals are another important philosophical topic where folk ascriptions of truth-aptness are challenged. Here nontruth-conditional accounts are motivated in part by the difficulty of finding systematic theories of truth-conditions for such conditionals that are consistent with the conditionals that we accept. This illustrates that folk judgments about the felicity of truth-ascriptions are vulnerable to varying sorts of interesting challenge.

28 See, for example, Blackburn (1993, 168ff.), Gibbard (1990, 26ff.), and Bjornsson and McPherson (2013, 17ff.).

If thinking *I ought to resist racism* was not in rational tension with thinking *I ought not to resist racism*, then it is hard to see how I could reason my way to conclusions about what I ought to do. Similarly, implication relations between sentences in normative discourse enable us to identify relationships between normative views, make valid arguments, and so forth. For example, if we want to coordinate on a project, it will be useful to be able to debate how we ought to proceed, and to keep track, within this debate, of the implications among competing proposals. (It is controversial whether we can explain inferential and implication relations on a noncognitive basis. But if we cannot, expressivism is hopeless for reasons that have nothing to do with minimalism, so I set this aside here.[29])

In nonnormative discourse, aptness for inferential and implication relations is tightly correlated with indicative syntactic form (where this form differs from, e.g., the typical syntax of interrogatives or imperatives). If normative talk—as the expressivist understands it—in fact arose as a linguistic vehicle to enable expression of inferential and implication relations, it did so via adopting the same indicative syntactic form. In nonnormative sentences, indicative form is highly correlated with truth-aptness. A flat-footed explanation of this is that nonnormative indicative sentences function to describe the world. But a universal generalization on the correlation between indicative form and truth-aptness would be part of the most elegant implicit model for the felicity of 'true,' which we would expect language learners to develop as they acquire English. The expressivist can propose that, as with its apparent applicability to Liar sentences, the apparent applicability of 'true' to normative sentences is an easily acquired elegant overgeneralization. Even if normative sentences are not truth-apt, then, this theory of error can explain why being disposed to take them to be truth-apt is strongly and nonaccidentally correlated with becoming competent with 'true' in the normal way.[30]

It is worth emphasizing two virtues of this theory of error. First, recall:

Generalized Explanatory For all true sentences 'S,' 'S' is true because (1) S, and (2) the meaning of 'S' (where (2) alone may be explanatorily sufficient in some cases).

This theory of error allows that the natural functional interpretation of Generalized Explanatory is in fact true, while minimalism was in tension with that interpretation. Now consider:

29 It has been proposed that minimalism itself can help the expressivist with these problems (see, e.g., Horwich 1993, 74–76). I take Dreier (1996) to have refuted this proposal.

30 As I noted in the introduction, I have focused here on expressivism about the OUGHT of flavorless endorsement. It is an interesting question how well this error theory "ports" to other parts of thought and talk for which expressivism is a live hypothesis. I am optimistic about its extension to moral and epistemic vocabulary, but I am less certain about other cases.

Normative Explanatory The sentence 'Tristram ought to resist racism' is true because (1) Tristram ought to resist racism and (2) the meaning of 'Tristram ought to resist racism'.

It is plausible that some of us will tend to accept Normative Explanatory, while others will not (even if they think Tristram ought to resist racism). Why? A plausible expressivist explanation for the tendency to rejection would appeal to some of us having a firmer implicit grasp of the expressivist character of normative talk. But with our theory of error in hand, we can also explain why many people are inclined to accept Normative Explanatory. The explanation is that Normative Explanatory is a trivial consequence of the combination of Generalized Explanatory and the general applicability of T-Schema. The error theory we have just offered suggests that instances of T-Schema involving normative sentences are false. This means that we can explain the intuitive plausibility of Normative Explanatory as a predictable consequence of our theory of error.

Second, one of the vices of the marriage to minimalism highlighted in the discussion of creeping minimalism is that the wedded expressivist faces dialectical pressure to accept a host of realist-sounding claims that she might otherwise hope to eschew. Take a simple case: the claim that whether I ought to resist racism is a matter of *fact*. This is the sort of claim that generates striking controversy in introductory metaethical discussions. But on a minimalist account of 'fact,' this is a triviality, secured by an insubstantive biconditional. Why do competent speakers not recognize it as such? This is a puzzle for the minimalist. The style of theory of error I have sketched can promise to explain why this claim is *more* controversial among competent speakers than claims of normative *truth*-aptness. The idea is that, even if truth ultimately amounts to correspondence, competence with 'true' is perhaps to a large extent a matter of keeping score of the status and relation among *sentences*. And this makes the pattern of correlation between inferential significance and truth-aptness very salient to someone building an implicit theory of the meaning of 'true' on the basis of her fellow speakers' usage in these contexts. Talk of 'facts,' on the other hand, arguably has its most natural home in folk metaphysical discourse, broadly construed. It is thus natural for some competent speakers to resist the claim that whether I ought to resist racism is a matter of fact, exactly in light of their implicit grasp of the expressivistic character of normative thought and talk.[31]

This section has sketched one way in which expressivism might be true without minimalism. The picture has two parts: an argument that we should expect to need a theory of error

31 On the Pea Soup discussion (McPherson et al. 2017), Jack Woods points out that the folk are more reluctant to ascribe truth to (e.g.) matters of taste than to normative matters. This hesitancy is arguably even more pronounced with facthood. This fits well with the picture being sketched here. Matters of taste do not pose the same hard interpretive problem as normative thought and talk: there are fewer apparent pressures toward realist interpretation, and hence it is easier for the folk's implicit grasp of the anti-realist character of the discourse to lead them to restrict the elegant generalization.

somewhere in our theorizing about normative thought and talk, combined with the provision of a proposed theory of error that specifically targets the apparent evidence concerning the felicity of ascribing 'true' to normative sentences. This puts us in a position to step back and consider the overall dialectic facing the expressivist.

Conclusions

This paper has argued that expressivists about normative thought and talk can and should resist marrying their views to minimalist theories of apparently metaphysical discourse. I have argued on two fronts. First, I showed that there are weighty reasons for expressivists to seek to avoid the marriage: It saddles their view with additional controversial commitments; it may threaten the distinctiveness of expressivism as a metaethical view, or undercut its apparently distinctive dialectical virtues. And if these worries can be fended off, the reconstructed distinctions will likely show that much of the thought and talk that the marriage is supposed to accommodate reflects commitments inconsistent with expressivism. Second, I have argued that we have principled reasons to seek theories of error in our theorizing about normative thought and talk, and I have sketched an attractive such theory of error to explain the felicity judgments that otherwise threaten to make minimalism indispensable to the expressivist.

It may seem that, in presenting that theory of error, above, I have exhibited the very vice that I claimed the expressivist should seek to avoid: I have committed the expressivist to a contentious theory concerning 'truth'-talk. However, I do not claim that the expressivist must adopt the theory of error just sketched. Rather, that theory of error illustrates the credibility of the broad strategy of challenging the probative force of the relevant felicity judgments. This in turn motivates the idea that expressivism need not be wedded to *any* particular view about truth. Perhaps some alternative way of interpreting this data is ultimately preferable: either a different theory of error or some alternative theory of truth. Indeed, perhaps some form of minimalism about truth is correct after all. My point is that expressivists can afford to be *uncommitted* about this topic.

Given the considerable costs of commitment to minimalism, this agnosticism is the most credible attitude for expressivists to take to this topic. In closing, it is worth emphasizing further reasons to adopt this sort of agnosticism. For brevity here, I have tended to contrast minimalism with a correspondence theory of truth. However, there are many nonminimalist theories of truth worth exploring. Some of these theories may, like minimalism, be able to combine with expressivism to vindicate felicity and commonsense judgments about normative sentences. Two salient examples are pluralist theories of truth, such as Lynch's (2013), and expressivist theories of truth, such as Schroeder's (2010). This complicates the upshot of this paper in two ways. On the one hand, it may suggest more reason for optimism that the correct theory of truth does not, in Jamie Dreier's words, "commit to anything metaphysically heavy-duty" (2010, 170). On the other, it reinforces the general point of this paper that there is little reason for the expressivist to be committed to minimalism about truth.

References

Anscombe, G. E. M. (1957). *Intention*. Cambridge, MA: Harvard University Press.

Armour-Garb, Bradley, and J. C. Beall (2005). "Introduction." In *Deflationary Truth*, edited by B. Armour-Garb and J. C. Beall, 1–29. Chicago, IL: Open Court Press.

Asay, Jamin (2013). "Truthmaking, Metaethics, and Creeping minimalism." *Philosophical Studies* 163, no. 1: 213–32.

Beall, Jc, Michael Glanzberg, and David Ripley (2017). "Liar Paradox." In *Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2017/entries/liar-paradox/.

Bedke, Matthew S. (2017). "Cognitivism and Non-Cognitivism." In *The Routledge Handbook of Metaethics*, edited by Tristram McPherson and David Plunkett, 292–307. New York: Routledge.

Björnsson, G., and T. McPherson (2014). "Moral Attitudes for Non-Cognitivists: Solving the Specification Problem." *Mind* 123, no. 489: 1–38.

Blackburn, Simon (1993). *Essays in Quasi-Realism*. Oxford: Oxford University Press.

Camp, Elisabeth (2017). "Metaethical Expressivism." In *The Routledge Handbook of Metaethics*, edited by Tristram McPherson and David Plunkett, 87–101. New York: Routledge.

Chrisman, Matthew (2012). "On the Meaning of 'Ought'." In *Oxford Studies in Metaethics*, vol. 7, edited by Russ Shafer-Landau, 304–32. Oxford: Oxford University Press.

Cuneo, Terence (2013). "Properties for Nothing, Facts for Free?" In *Oxford Studies in Metaethics*, vol. 8, edited by Russ Shafer-Landau. Oxford: Oxford University Press.

Divers, John, and Alexander Miller (1994). "Why Expressivists about Value Should Not Love Minimalism about Truth." *Analysis* 54, no. 1: 12–19.

Dreier, James (1996). "Expressivist Embeddings and Minimalist Truth." *Philosophical Studies* 83: 29–51.

——— (2004). "Metaethics and the Problem of Creeping Minimalism." *Philosophical Perspectives* 18, no. 1: 23–44.

——— (2010). "When Do Goals Explain the Norms That Advance Them?" In *Oxford Studies in Metaethics*, vol. 5, edited by Russ Shafer-Landau. Oxford: Oxford University Press.

——— (2012). "Quasi-Realism and the Problem of Unexplained Coincidence." *Analytic Philosophy* 53, no. 3: 267–87.

——— (2015). "Explaining the Quasi-Real." In *Oxford Studies in Metaethics*, vol. 10, edited by Russ Shafer-Landau. Oxford: Oxford University Press.

——— (2018). "The Real and the Quasi-Real." *Canadian Journal of Philosophy* 48, no. 3–4: 532–47.

Dunaway, William (2016). "Expressivism and Normative Metaphysics." In *Oxford Studies in Metaethics*, vol. 11, edited by Russ Shafer-Landau. Oxford: Oxford University Press.

Egan, Andy (2007). "Quasi-Realism and Fundamental Moral Error." *Australasian Journal of Philosophy* 85, no. 2: 205–19.

Eklund, Matti (2002). "Inconsistent Languages." *Philosophy and Phenomenological Research* 64, no. 2: 251–75.

——— (2017). *Choosing Normative Concepts*. Oxford: Oxford University Press.

Enoch, David (2011). *Taking Morality Seriously*. Oxford: Oxford University Press.

Faraci, David, and Tristram McPherson (2017). "Ethical Judgment and Motivation." In *The Routledge Handbook of Metaethics*, edited by Tristram McPherson and David Plunkett, 308–23. New York: Routledge.

Field, Hartry (1998). "Epistemological Non-Factualism and the a Prioricity of Logic." *Philosophical Studies* 92, no. 1/2: 1–24.

Fine, Kit (2001). "The Question of Realism." *Philosophers' Imprint* 1, no. 1: 1–30.

Gibbard, Allan (1990). *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

——— (2003). *Thinking How to Live*. Cambridge, MA: Harvard University Press.

——— (2011). "How Much Realism? Evolved Thinkers and Normative Concepts." In *Oxford Studies in Metaethics*, vol. 6., edited by Russ Shafer-Landau, 33–51. Oxford: Oxford University Press.

——— (2012). *Meaning and Normativity*. Oxford: Oxford University Press.

Golub, Camil (2017a). "Expressivism and Realist Explanations." *Philosophical Studies* 174, no. 6: 1385–1409.

——— (2017b). "Expressivism and the Reliability Challenge." *Ethical Theory and Moral Practice* 20, no. 4: 797–811.

Gupta, Anil (1993). "A Critique of Deflationism." *Philosophical Topics* 21, no. 2: 57–81.

Horgan, Terence, and Mark Timmons (2006). "Cognitivist Expressivism." In *Metaethics after Moore*, edited by Terence Horgan and Mark Timmons, 255–98. Oxford: Oxford University Press.

Horwich, Paul (1993). "Gibbard's Theory of Norms." *Philosophy & Public Affairs* 22: 67–78.

Johnston, Mark (1992). "Reasons and Reductionism." *Philosophical Review* 101: 589–618.

Köhler, Sebastian (2014). "What Is the Problem with Fundamental Moral Error?" *Australasian Journal of Philosophy* 93, no. 1: 161–5.

Köhler, Sebastian, and Michael Ridge (2013). "Revolutionary Expressivism." *Ratio* 26, no. 4: 428–49.

Kraut, Robert (1993). "Robust Deflationism." *Philosophical Review* 102, no. 2: 247–63.

Lynch, Michael (2013). "Expressivism and Plural Truth." *Philosophical Studies* 163, no. 2: 385–401.

McPherson, Tristram (2009). "Moorean Arguments and Moral Revisionism." *Journal of Ethics and Social Philosophy* 3, no. 2: 1–24.

——— (2015). "What Is at Stake in Debates among Normative Realists?" *Nous* 49, no. 1: 123–46.

——— (2018). "Authoritatively Normative Concepts." In *Oxford Studies in Metaethics*, vol. 13, edited by Russ Shafer-Landau, 253–77. Oxford: Oxford University Press.

McPherson, Tristram, Derek Baker, Jamie Dreier, Eric Hubble, Frank Jackson, Jussi Suikkanen, and Jack Woods (2017). "Expressivism without Minimalism." Blog post with discussion. *Pea Soup* blog, October 2017. http://peasoup.us/2017/10/expressivism-without-minimalism/.

Price, Huw (2015). "From Quasi-Realism to Global Expressivism—and Back Again?" In *Passions and Projections*, Edited by Robert Johnson and Michael Smith, 134–52. Oxford: Oxford University Press.

Ridge, Michael (2014). *Impassioned Belief*. Oxford: Oxford University Press.

Schroeder, Mark (2008). *Being For*. Oxford: Oxford University Press.

——— (2009). "Hybrid Expressivism: Virtues and Vices." *Ethics* 119: 257–309.

——— (2010). "How to Be an Expressivist about Truth." In *New Waves in Truth*, edited by Nikolaj Jang Pedersen and Cory Wright, 282–98. New York: Palgrave MacMillan.

Simpson, Matthew (2018). "Solving the Problem of Creeping Minimalism." *Canadian Journal of Philosophy* 48, no. 3–4: 510–31.

Smith, Michael (1994a). *The Moral Problem*. Oxford: Blackwell.

——— (1994b). "Why Expressivists about Value Should Love Minimalism about Truth." *Analysis* 54, no. 1: 1–11.

Soames, Scott (1999). *Understanding Truth*. Oxford: Oxford University Press.

Stoljar, Daniel, and Nic Damjanovic (2014). "The Deflationary Theory of Truth." In *The Stanford Encyclopedia of Philosophy* (Fall 2014 Edition), edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2014/entries/truth-deflationary.

Street, Sharon (2011). "Mind-Independence without the Mystery." In *Oxford Studies in Metaethics*, vol. 6, edited by Russ Shafer-Landau, 1–32. Oxford: Oxford University Press.

Svoboda, Tony (2017). "Why Moral Error Theorists Should Become Revisionary Moral Expressivists." *Journal of Moral Philosophy* 14, no. 1: 48–72.

Tarski, Alfred (1958 [1935]). "The Concept of Truth in Formalized Languages." In *Logic, Semantics, Metamathematics: Papers from 1923 to 1938*. Oxford: Oxford University Press.

Timmons, Mark (1999). *Morality without Foundations*. Oxford: Oxford University Press.

Toppinen, Teemu (2013). "Believing in Expressivism." In *Oxford Studies in Metaethics*, vol. 8, edited by Russ Shafer-Landau. Oxford: Oxford University Press.

——— (2017). "Hybrid Accounts of Ethical Thought and Talk." In *The Routledge Handbook of Metaethics*, edited by Tristram McPherson and David Plunkett, 243–59. New York: Routledge.

Williamson, Timothy (2007). *Philosophy of Philosophy*. Oxford: Oxford University Press.

Wright, Crispin (1992). *Truth and Objectivity*. Cambridge, MA: Harvard University Press.

Yalcin, Seth (2011). "Nonfactualism about Epistemic Modality." In *Epistemic Modality*, Edited by Andy Egan and Brian Weatherson, 295–332. Oxford: Oxford University Press.

——— (2012). "Bayesian Expressivism." *Proceedings of the Aristotelian Society* CXII, no. 2: 123–60.

9

METASEMANTIC QUANDARIES

Nate Charlow

> *Traditional interest theories hold that ethical statements are descriptive of the existing state of interests—that they simply give information about interests . . . It is this emphasis on description, on information, which leads to their incomplete relevance. Doubtless there is always* some *element of description in ethical judgements, but this is by no means all. Their major use is not to indicate facts, but to* create an influence. *Instead of merely describing people's interests, they* change *or* intensify *them. They* recommend *an interest in an object, rather than state that the interest already exists.*
>
> (Stevenson 1937)

> *Conversation, then, is far more than a carrier of information. In talk we work out not only what to believe about things and events and people, but how to live. We work out how to feel about things in our lives, and in the lives of others.*
>
> (Gibbard 1990)

1. Introduction

If an expression's semantic value or content is sometimes fixed in virtue of certain features of the context in which that expression is uttered, does it follow that that expression's semantic value, relative to the context of utterance, is *always* fixed in virtue of such features? If the semantic value of a sentence containing such an expression, relative to the context of utterance, can depend on the choice of a value for that expression, does it follow that the semantic value of such a sentence, relative to the context of utterance, *always* depends on such a choice? Although such questions are not often registered in the philosophy of language,

a great many philosophers of language—this paper will refer to them as "Referentialist" Metasemanticists—do seem to behave if the answer to both of these questions is a clear "yes."[1]

This paper proposes a different direction. It advocates a form of Expressivism as a strategy for resolving certain metasemantic puzzles about identifying the semantic value of a context-sensitive expression in context. On this version of Expressivism, while some utterances containing expressions in the target class do aim to proffer a proposition in the discourse in which they occur, such uses should be thought of as a kind of special interpretive case. Puzzles arising from the pressure to say what a putatively context-sensitive expression "refers to" in contexts that do not seem to specify a referent dissolve, once we appreciate that such attempts were ill-placed to begin with.

The version of Expressivism defended here will not deny that expressions in the target class *can* contribute a contextually determined semantic content to the computation of the semantic value of a larger syntactic constituent in which they occur. It instead says that expressions in the target class *need not* deliver a contextually determined semantic value as the input to semantic computation. Computation of a semantic content in context can avail itself of a variety of compositional mechanisms (e.g., λ-abstraction) in order to generate *nonpropositional* (more specifically, *prescription-type*) semantic contents, whose features are appropriate to realistic conversational/communicative aims of speakers.

Here is the plan. Section 2 reviews the basic metasemantic challenge raised by uses of context-sensitive expressions that do not appear to be specifically referential—that is, uses of such expressions in contexts in which they appear not to contribute a single entity of the relevant semantic type to semantic computation. Section 3 reviews a recent proposal for representing this phenomenon in the domain of gradable adjectives—the Expressivist account of MacFarlane (2016). It argues that there are *two* types of failures of specific referentiality that must be distinguished: those arising from semantic indeterminacy and those arising from broadly "nondescriptive" or "expressive" communicative aims (while also arguing that MacFarlane's account blurs these). While the former kind of failure can be modeled simply by allowing our theory of content-assignment to go indeterminate for certain kinds of context, the latter kind of failure cannot. Section 4 identifies some (at least prima facie) shortcomings with a metalinguistic treatment of the phenomenon of nondescriptive uses. Section 5 suggests an alternative model, on which nondescriptive uses are represented as semantically expressing a kind of *prescription*—a prescription that is semantically derivable, using familiar compositional machinery, from the standard semantic representation of,

1 Some linguists have been more circumspect (see, e.g., Dowty 1985, Jacobson 1999, and the literature on Direct Compositionality that grows out of this work). Theorists who have entertained the "no" answer to such questions (including Barker 2002; MacFarlane 2016) have tended to work with a (in my view) distorted perception of what the theoretical terrain will look like (and what the relevant modeling options are), once we've liberated ourselves of the demands of a Referentialist Metasemantics. Much more on this below.

for example, a gradable adjective in the positive form. Section 6 situates this proposal within the metasemantic framework of Gibbard (1990), while identifying one significant respect in which this proposal differs from Gibbard's.

2. Metasemantic Puzzles

According to our shared Kaplanian folklore, there are two ways a linguistic context can fix a semantic value for a context-sensitive expression. Some such expressions (e.g., the so-called "pure" indexicals, like "I") seem to come with lexically encoded rules (characters) that suffice to identify their semantic values in any context. Other such expressions (what King 2014 calls the *supplementives*) seem to be such that their characters do not suffice to fix their semantic values for any context whatever; in Kaplan-speak, such expressions require an associated "demonstration" (understood here to involve some type of indication of the speaker's referential intention that is supplementary to the utterance itself) in conjunction with the information provided by the utterance context and the expression's lexical character, in order to be assigned a semantic value in context. King argues that the class of natural language supplementives includes demonstratives, certain pronouns, modals, conditionals, gradable adjectives, and more besides (see also his 2017, 2018).

It is evident that the expressions that King understands to be supplementives exhibit *some form* of context-dependence (more precisely, that they *can receive* context-dependent semantic values). A speaker who says "Steph Curry is tall" can be interpreted as making a different claim depending on the relevant comparison class (e.g., American adults or NBA players). What the speaker says seems false if the speaker is comparing Steph Curry, who is just under 2 meter tall, to other NBA players, true if they are comparing him to American adults. A speaker who says "Bond might be in Zurich" can be interpreted as characterizing or describing a feature of their own knowledge, or what is known by a relevant group (what "we" know). What the speaker says seems false if the speaker is trying to characterize or describe what the group knows (and someone in the group knows Bond is not in Zurich). But it seems true if the speaker is trying to characterize what *they* know (and it is compatible with what they know that Bond is in Zurich).

None should dispute that supplementives exhibit context-dependence in this (quite thin) sense: for each supplementive, it is clear that *there are* contexts in which its semantic value is a function of that context. It does not, however, follow from this that supplementives exhibit context-dependence in a more theoretically robust sense. In particular, from the fact that there are contexts in which the semantic value of a supplementive is a function of that context, it does not follow that the semantic value of a supplementive is a function of context in *any* context.[2]

2 Throughout I bracket bound occurrences of supplementive expressions (e.g., pronouns).

Thank goodness for that! Suppose it were the case that, for *any* context *c*, the semantic value of a supplementive (tokened as part of a larger clause with a semantic value at *c*) was, in fact, a function of *c*; suppose, that is to say, that context always fixed the semantic value of a supplementive (when it occurred within a larger clause with a semantic value). Assigning a semantic value to the larger clause would *commit* the theorist to assigning a semantic value to the supplementive as a function of context. And this would in turn give rise to metasemantic questions about what features of the context the supplementive's semantic value was a function of.

Metasemantic questions like these have a tendency to become metasemantic quandaries. As King has stressed, supplementive expressions exhibit a strong degree of tolerance for **contextual nonspecificity**.[3] Even when no relevant facts about the context seem to settle which interpretation of the supplementive is "the" interpretation of the supplementive in that context, speakers still make liberal use of supplementive expressions (and addressees generally have no difficulty interpreting such utterances, assessing them for truth or falsity, affirming or denying the semantic content of those utterances, etc.). We might imagine, for example, that the speaker is[4] *indifferent* to (and therefore undecided about) whether, in saying "that is a beautiful car," they mean to be talking about the car type or token. An addressee who is, by stipulation, omniscient about the speaker's referential intentions will have no way of determining what "that" designates, but no difficulty in figuring out how to update on the speaker's utterance. The speaker appears to say something with content, even though the semantic value of the supplementive "that" is, almost by stipulation, not a function of the relevant context.

How should a theorist react to the phenomenon of nonspecificity? I can see three options:

1. One might think that there is pressure (from, e.g., compositionality) to hold that the semantic value of a supplementive at *c* (tokened as part of a larger clause with a semantic value at *c*) is *always* a function of *c* (if it is ever a function of *c*). Such pressures render metasemantic quandaries theoretically unavoidable.
2. One might also think that sentences with nonspecifically referential supplementives can receive *pragmatically workable interpretations* in context, but that such interpretations are not semantic values in the ordinary sense (e.g., because such interpretations arise as a kind of indirect speech act).

---

3 King's term is "underspecification" (see King 2017, 2018). "Nonspecificity" is clumsier, but it avoids the implication, which I will be denying here, that specificity of semantic value is the default/normal state of affairs for a supplementive in context.

4 Typically, I would think, speakers *are undecided* on whether they mean to be referring to the car type or car token. In most contexts, it just is unimportant, given the communicative aims of the speaker, for the speaker to be precise or decided in this fashion.

3. One might posit some sort of semantic distinction between two tokens of the same sentence type relative to a context c_1 in which the semantic value of the supplementive is specified and a context c_2 in which no semantic value for the supplementive is specified.

The next sections will examine these options, and identify some reasons for developing a theory in the general mold of (3). After that, I try to lay down a few semantic and pragmatic cornerstones for theorizing in this direction. The theory I will outline is in certain respects just a "generalization" of Gibbard's Expressivism to non-practical language. The driving idea is that Gibbard's theory of practical claims is a branch of a larger theory of "cognitive prescriptions", on which speakers express cognitive prescriptions – *ways* of thinking/representing on some at-issue matter, which they have *and* in some sense expect their interlocutors to share – by expressing properties of (contextually free) semantic parameters. (Practical claims are a special case, in which the contextually free semantic parameter is a "normative system" or "planning state".) This paper will ultimately register some (significant) differences with Gibbard's theory, while also taking issue with the, we might say, overly "literal" Gibbardianism of theorists like MacFarlane (2016). But the theory is, in its essence, an Expressivist theory – it presupposes the approach to thinking about thought and content that Gibbard pioneered.

3. Referential Metasemantics

It is evident that supplementives sometimes receive semantic values as a function of context (and that, when they do, their semantic values combine with the semantic values of sister nodes in the ordinary compositional fashion). In type-theoretic frameworks (like Heim and Kratzer 1998), the default mode of semantic composition is Functional Application.[5]

Functional Application (FA)
If α is a branching node whose daughters are β and γ, then, when defined, $[\![\alpha]\!]^c = [\![\beta]\!]^c([\![\gamma]\!]^c)$ or $[\![\alpha]\!]^c = [\![\gamma]\!]^c([\![\beta]\!]^c)$.

Suppose γ is a supplementive expression whose semantic value in c, $[\![\gamma]\!]^c$, is fixed as a function of c. If $[\![\beta]\!]^c$ is a function whose domain includes $[\![\gamma]\!]^c$, semantic composition proceeds in the usual fashion, and the semantic value of α at c, $[\![\alpha]\!]^c$, is simply $[\![\beta]\!]^c([\![\gamma]\!]^c)$.

This "familiar" state of affairs can come to have the air of obligation, if our only mode of semantic composition is FA: unless $[\![\gamma]\!]^c$ is defined (and of the right type to allow it to combine with $[\![\beta]\!]^c$), semantic composition (via FA) seems to break down. So, for any context c

5 If the reader is worried that the content of a complex constituent, relative to a context, is not compositional (see, e.g., Lewis 1980), they may read these claims as claims about the semantic value (or extension) of a complex constituent, relative to a context.

in which semantic composition appears not to break down for α—if, say, α is a sentence that appears to say something at c—there is pressure to hold that $[\![\gamma]\!]^c$ is a value of the requisite semantic type. And so there is compositional pressure to articulate what I'll term a **Referential Metasemantics** for supplementive expressions (by which I will mean an account of how, or in virtue of what, a supplementive expression receives *a semantic value of the ordinary type at any context* in which computation of semantic value for a syntactic constituent that contains it succeeds).

3.1 MacFarlane on Nonspecificity

Though Referential Metasemantics is the theoretical default for work on the metasemantics of supplementives, nonspecific uses of supplementive expressions present an immediate, and serious, challenge to the view.

Let's start with the following remark by MacFarlane:

> We have plenty of . . . flexible expressions, whose extensions are to a great extent up to the speaker to determine. The most obvious examples are bare demonstratives like "this" and "that." In principle, I can use "that" to refer to any object. But with this freedom comes great responsibility. I must provide my hearers with enough cues to enable them to associate my use of "this" with the same object I do, or communication will fail . . . [I]n every case, *we're obliged to do whatever is required to get our hearers to associate the same object with the demonstrative that we do*. If we fail to do this, it will be sheer luck if they understand us.
>
> (MacFarlane 2016, 260–61, emphasis mine)

MacFarlane says, I think correctly, that successful communication in a context in which a speaker expresses a semantic value with a supplementive requires that speaker and addressee coordinate on that semantic value. Nonspecific uses of supplementive expressions present cases in which it appears *no* relevant facts about the context settle the expression's interpretation. In such contexts, a speaker *does not* (and, in many cases—e.g., in the case of specifying an exact comparison class for a gradable adjective—*cannot*) provide cues that allow her addressee to associate her use of a supplementive with the object, if any, she associates with the supplementive. If the supplementive contributed a semantic value of the ordinary type to the proposition expressed by the utterance, there would be no way for speaker and addressee to coordinate on the proposition expressed by the utterance. Since successful communication seems to demand such coordination, the prediction is that communication must *fail* in any such context.

We have seen that this is a false prediction. To avoid it, MacFarlane proposes denying, in at least some such cases, that a speaker expresses a proposition with her utterance that is a function of a contextually determined semantic value for the supplementive. MacFarlane's specific target is the contextually determined degree threshold invoked in the standard degree semantics for gradable adjectives (see a.o. Kennedy 2007). On that semantics, "Steph

is tall" is true at *c* just when Steph's degree of height, deg_{tall} (*Steph*), exceeds a *c*-determined threshold of height $\theta_c(deg_{tall})$ (with $\theta_c(deg_{tall})$ delivering a *minimum degree* of tallness—a threshold—above which someone's degree of tallness is sufficient to count as tall in *c*.

$$[\![Steph\ is\ tall]\!]^c = 1\ iff\ deg_{tall}\ (Steph) > \theta_c(deg_{tall})$$

MacFarlane takes this sort of semantics to be ruled out by the sorts of considerations described in the prior paragraph; coordinating on a precise degree threshold is not something that people are even ordinarily able to do in conversation.

Instead, MacFarlane claims, sentences like "Steph is tall" are semantically evaluated with respect to Gibbardian *hyperplans*, which are objects that, for any possible situation *s*, specify exactly which actions are forbidden/permitted in *s* (see esp. Gibbard 1990, 2003). A hyperplan is, inter alia, a plan for where to draw the line for the degree of height required to count as tall in *c*, for any context of utterance *c*. For MacFarlane, the semantic content of a sentence like "Steph is tall" at a context of utterance *c* is the set of hyperplans *h* such that Steph's height exceeds the degree of height required to count as tall in *c*, according to $h—\theta_h(deg_{tall})$:

$$[\![Steph\ is\ tall]\!]^c = \lambda h.deg_{tall}\ (Steph) > \theta_h(deg_{tall})$$

Equivalently, its content is a *property of planning states*—specifically, the property of planning to count anyone of at least Steph's height as tall. Contra Kennedy, it is not a contextually determined proposition. Proffering such a property in discourse amounts, *not* to proffering a proposition for addition to the Common Ground, instead to proffering to one's audience a (*practical*, rather than doxastic or epistemic) constraint on who *to count as* tall—namely, anyone of at least Steph's height.

3.2 Semantic Indeterminacy

Though I am broadly sympathetic to the conclusion—and will argue for a version of it later on—I believe the argument rests implicitly on a dubious contrast between bare demonstratives and gradable adjectives. MacFarlane draws the contrast as follows:

> While in using a bare demonstrative like "this" one must have a definite object in mind, and successful uptake requires recognizing what object that is, there are no analogous requirements for the use of "large." The speaker need not have in mind a particular delineation (even a "fuzzy" one), and the hearer need not associate the speaker's use with a particular delineation. What we get instead are constraints on delineations. In saying that apple C is large, I rule out certain ways of drawing a line between large and non-large apples, while leaving others open.
>
> (2016, 265)

Consider an utterance of "that's a beautiful car" in a context in which the speaker lacks a specific referential intention. Let us suppose further that it is common ground that a token of the car type is beautiful just when that car type is beautiful. Thinking this particular car beautiful is informationally equivalent, in this context, to thinking the car type beautiful. Communication seems to succeed here, even though the speaker does not provide the addressee the relevant cue about her referential intention (indeed, the speaker appears to lack any such intention to provide a cue *about*) (see again King 2017, 2018).

Now the fact that communication using a demonstrative can succeed, even absent specific referential intention, does *not* warrant the conclusion that the semantic value in context of a demonstrative-containing sentence like "that's a beautiful car" should be taken to be *anything other than* a proposition. Certainly, it does not warrant the (stronger and false) conclusion that the content of this sentence *is* a property of planning states (i.e., the property of planning to use "that" to refer to a beautiful car type or car token, but being indifferent between these[6]). One therefore wonders: what is supposed to differentiate the sort of semantic nonspecificity that warrants an alternative assignment of content (e.g., a property of hyperplans) from the sort of semantic nonspecificity that does not?[7]

Nor does this fact warrant even the weaker conclusion that the semantic value of "that's a beautiful car" is not a function of a contextually determined semantic value for the supplementive. "The" semantic value of this sentence could be *either* the proposition that cars of this type are beautiful or that this particular car is beautiful, in the sense that both propositions are semantically "eligible" in this context—there are contextually admissible resolutions of the demonstrative "that" that yield each of these propositions as the output of semantic computation. Either resolution of the demonstrative would yield a proposition with the right pragmatic profile (since, by assumption, updating on one yields the same informational change in this context as updating on the other).

We *could* ask which of these propositions is "the" semantic value of "that's a beautiful car" in context. But why? The speaker's utterance could express *either* of two propositions (and an interpreter may update on either of these two propositions, apparently without loss of information or understanding). There is no obvious cost to saying that the speaker expresses *both* semantic values in uttering this sentence in this way (compare King 2014, 106)—or, alternatively, that there is no determinate fact of the matter about which of these semantic values is the semantic value expressed by the utterance. Let us assume that, to realize their communicative aims without misunderstanding, speaker and addressee must

6 This seems to misconstrue the point of such an utterance. A speaker who says "that's a beautiful car" in the context we are imagining intends to express an aesthetic judgment about a car, not to express a constraint on one's plans for using the demonstrative "that."

7 MacFarlane writes that the standard picture of content as the proposition expressed by a sentence in context "assumes that speaker and hearer have shared knowledge of what it takes for the sentence to be true in the present context. When that assumption breaks down, truth-conditions lose their explanatory relevance" (MacFarlane 2016, 265).

be able to coordinate on a way of updating the context/their information. But this does not generally require that speaker and address be able to coordinate on a unique semantic value for the speaker's utterance. Semantic specificity beyond what is required for communicative aims is otiose (for the agents of a conversation, as well as for theorists trying to model their conversation). Why, then, would we posit it?

With a little effort, this strategy can be extended to MacFarlane's target, gradable adjectives. Someone who says "Steph Curry is tall" could, from the point of view of a referential metasemantics, express any of (continuum) many propositions—one for each way of drawing the threshold (compare Braun and Sider 2007; King 2014, 112). Do these propositions "carry" the same information in context? In one sense, obviously, no: the proposition that Steph's height exceeds θ_1 is, of course, distinct from the proposition that Steph's height exceeds θ_2; supposing θ_2 exceeds θ_1, the proposition that Steph's height exceeds θ_2 asymmetrically entails the proposition that his height exceeds θ_1. Still, if both θ_1 and θ_2 are candidate thresholds—both are contextually eligible ways of drawing the line between tall and non-tall—Steph's height must be assumed (by someone who says "Steph is tall") to exceed *both* θ_1 and θ_2: if Steph's height is not assumed to exceed θ_2, but θ_2 is regarded as an eligible threshold in the context, the utterance is marked.

A: Steph is 1.9m. Do you have to be 2m to be tall?
B: I'm not sure.
A: Is Steph tall?
B: ?? Yes, he is.

For any "contextually eligible" θ_1 and θ_2, someone who says "Steph is tall" will be taken to be committed to both the proposition that Steph's height exceeds θ_1 and the proposition that it exceeds θ_2. There is, again, no clear cost to saying that the speaker expresses *both* semantic values in saying "Steph Curry is tall"—or, alternatively, that there is no fact of the matter about which of the various compositionally eligible semantic values is *the* semantic value expressed by her utterance—since the speaker is taken as committed to any such value. Why, again, would a theorist demand semantic specificity, when the communicative aims of speakers and addressees *do not*?

Metasemantic quandaries dissolved? Expressivism circumvented? And this simply by introducing a plausible bit of indeterminacy into the semantico-pragmatic relation (*expression*) that relates speakers to the semantic values of their utterances?

3.3 Determinacy and Description

No, or so I will say. Semantic indeterminacy of the sort just described is a useful rubric for thinking about one kind of failure of specific referentiality (in the domain of gradable adjectives). But semantic indeterminacy does *not* help to account for another way in which speakers can use context-sensitive items nonspecifically.

A now common observation in the literature on gradable adjectives notes that they have two canonical "modes of use" (Barker 2002, 1ff.), one broadly *descriptive* in character (e.g., a speaker uses an utterance of a gradable adjective in the positive form to provide information about someone's height), another broadly *nondescriptive* in character. Barker contrasts these modes of use as follows:[8]

> Normally, (1) will be used in order to add to the common ground new information concerning Feynman's height:
>
> (1) Feynman is tall.
>
> But (1) has another mode of use. Imagine that we are at a party. Perhaps Feynman stands before us a short distance away, drinking punch and thinking about dancing; in any case, the exact degree to which Feynman is tall is common knowledge. You ask me what counts as tall in my country. "Well," I say, "around here, . . ." and I continue by uttering (1). This is not a descriptive use in the usual sense. I have not provided any new information about the world, or at least no new information about Feynman's height . . . All I have done is given you guidance concerning what the prevailing relevant standard for tallness happens to be in our community; in particular, that standard must be no greater than Feynman's maximal degree of height. (Barker 2002, 1–2)

It is *this*, apparently nondescriptive, use that MacFarlane proposes to model with hyperplan-type content (see, e.g., MacFarlane 2016, 256). That is to say, MacFarlane proposes to represent the speech act that Barker glosses as giving "guidance concerning . . . the prevailing relevant standard for tallness" as a speaker's proffering a (practical, not doxastic or epistemic) constraint on who *to count as* tall.

The question of whether an utterance expresses a specific semantic value in *c* is, however, distinct from the question of whether it functions to describe in *c*. That is to say, semantic determinacy is orthogonal to descriptiveness.

a. An utterance may exhibit semantic determinacy, while expressing an assertion that updates the context with a determinate proposition. (The "Familiar" Case)
b. No one proposition is the semantic value of the utterance, though the utterance is understood as expressing an assertion that updates the Common Ground with a determinate proposition (since the semantically eligible propositions are informationally equivalent in context). (most of King's Cases)
c. An utterance may exhibit semantic determinacy, while the utterance's force is to proffer some kind of cognitive property or constraint.

8 Barker glosses the nondescriptive use as "metalinguistic." On the limits of this gloss, see section 4 below.

d. An utterance may exhibit semantic indeterminacy, while the utterance's force is to proffer some kind of cognitive property or constraint.

As a specific illustration, a speaker can use epistemic "might" in any of the following ways:[9]

a. To describe what is possible at *c* given a salient body of information at *c*. (This is the "ordinary" or "familiar" case.)
[Context: A is gathering information about B's information, and this is common ground between A and B.]
A: Where might Bond be [given your information]?
B: He might be in Zurich.

b. To describe what is possible at *c* given the content of some or other body of information at *c*, while remaining undecided about which body of information. (King-type nonspecificity)
[Context: A is gathering information about C and D's information, and this is common ground between A and B.]
A: Where might Bond be?
B: He might be in Zurich [given what C or D believe, it doesn't matter].

c. To determinately constrain someone's information so that, if the addressee accepts her utterance, she will regard a determinate proposition *as possible* (compare Moss 2013, 2015; Swanson 2006, 2016).[10]
[Context: A and B are disagreeing about where Bond might be.]
A: Bond has to be in London!
B: No, Bond might be in Zurich.

d. To indeterminately constrain someone's information (in a context where satisfying one semantically eligible constraint is informationally equivalent to satisfying any semantically eligible constraint).
[Context: A/B could be referring to either the car type or token.]
A: That might be a Ferrari.
B: No, but it might be a Maserati.

I think it is apparent that stating a theoretically adequate metasemantics for semantic indeterminacy—one liberated from the assumption that the relation that holds between

9 N.B. I don't claim all of these functions are attested for all supplementive expressions (although for some such expressions, like epistemic "might," it does appear that all are attested).

10 It is standardly held that such an update cannot be generally modeled as updating on a proposition (Russell and Hawthorne 2016; Veltman 1996; Yalcin 2011). I here treat disagreement about *whether to treat the proposition that Bond is in Zurich as possible* as indicative of nondescriptive disagreement, that is, disagreement that is not well-represented as disagreement about features of the actual world. This is not to say that other models of this sort of disagreement are ruled out (see MacFarlane 2011, 2014 for one).

a speaker and the content of what she says (i.e., *expression*) must be a *function*—does not free the theorist from the need to explain how certain context-sensitive expressions receive nondescriptive interpretations in context (or from the need to explain what such interpretations consist in). These are just different tasks. The problem of semantic indeterminacy is resolved—to a first pass, anyway—by getting comfortable with the notion that certain semantic facts are indeterminate (or that the semantic relation of expression is one-many).[11] It is not to be resolved with planning content (unless the theorist is willing to hold—contrary to apparent fact—that all utterances exhibiting semantic indeterminacy express nondescriptive planning content).

Nor is the problem of modeling nondescriptive, or constraint-type, interpretations resolved by making semantic facts indeterminate. It is, I will ultimately argue, resolved by recognizing a distinctive kind of *prescriptive content*—similar to, but also distinct in important ways, from the sort of planning content envisioned by MacFarlane.

4. The Metalinguistic Strategy

Nondescriptive uses of context-sensitive expressions raise metasemantic issues similar to those raised by nonspecific uses. Specifically, both are prima facie counterexamples to a Referentialist metasemantics for such expressions: if the expression contributed a semantic value of the ordinary type to semantic computation for the sentence, the output of semantic computation for the sentence would be a proposition involving that semantic value, and this would apparently fix a descriptive use.

This section will describe one strategy for resisting this argument—the metalinguistic strategy suggested in Barker (2002) and refined in Plunkett and Sundell (2013). On this strategy, sentences in the target class generally express propositions, but this need not fix a descriptive use for the sentence. I will argue that nondescriptive uses are not generally well understood as advancing proposals bearing on how to use language.

In contexts that do not provide a unique threshold for counting as tall, Barker (2002) represents the content of a claim like "Feynman is tall" as a constraint on the tallness-thresholds that are compatible with the context of utterance; the force of uttering such sentences is to

11 I say "to a first pass" because a general theory of how indeterminate semantic values determine illocutionary force in context will also need to be stated. This will be a difficult project, since there are a range of relations a set of candidate semantic values may bear to the semantic value on which agents update in a context (e.g., sometimes agents update on a disjunction of candidate semantic values, other times they update on the strongest semantic value; for examples of this variability, see King 2017). The project is then to say how interpreters derive the content on which they update in such contexts. (Note, by the way, that this is distinct from the problem of saying which proposition interpreters update on in such contexts: the information interpreters glean from the utterance is generally clear from the context, but it does not appear to be a function of which propositions are expressed by the utterance.) I will table this problem here.

eliminate those thresholds that are incompatible with the utterance from eligibility in the context of utterance.

Barker (2002) offers a substantive characterization of the *speech act* associated with proposing or expressing this sort of update as metalinguistic in nature or aim: "My purpose in uttering ['Feynman is tall'] . . . would be nothing more than to communicate something about how to use a certain word appropriately" (2). At first blush, however, Barker's metalinguistic gloss on this update looks undermotivated: proposing to exclude a candidate threshold for tallness from eligibility is not (obviously, anyway) the same thing as proposing to tell one's addressee how to use "tall." What reason is there to *identify* these speech acts? More generally, why think that modeling the nondescriptive interpretation of, for example, a sentence containing a gradable adjective *requires* a metalinguistic analysis?

In the Stalnakerian framework Barker takes on (Stalnaker 1978, 1984), proposals to update the context that are modeled as *eliminative* (e.g., as the intersection of the set of thresholds compatible with the utterance with the set of thresholds antecedently eligible in the context of utterance) are linked to a particular *functional role*. The function of updating a context with a set of possible worlds is straightforward to characterize: the antecedently eligible worlds—those in the pre-update context set—represent ways the world could be, given the present state of the conversation. The functional role of refining this set is to eliminate certain worlds as *candidates for actuality* in the conversation.

This philosophical context in view, it seems fairly clear that this functional story does not extend to the elimination of thresholds from contextual eligibility. An eligible threshold is not a *candidate for actuality*—there is, I will take it, no "actual" threshold that we are trying to figure out in conversation when we use gradable adjectives (MacFarlane 2016 agrees and offers considerations in support of this claim).

The metalinguistic strategy posits a *novel functional role* for the act of eliminating a threshold from contextual eligibility. The idea is that a threshold θ for a gradable adjective α is *c-eligible* when it constrains appropriate use of α in c: an utterance containing α must be compatible with some c-eligible threshold if the utterance is appropriate in c. (Compare: in the standard Stalnakerian framework, the possible worlds proposition one asserts must be compatible with the context set for the assertion to be appropriate.) To eliminate a threshold from contextual eligibility for α is therefore to make a metalinguistic proposal: to constrain how one's interlocutors use α in the conversation.

If the metalinguistic strategy is the right one for modeling nondescriptive interpretations of, for example, sentences containing gradable adjectives, these sentences might semantically express propositions—as Plunkett and Sundell (2013) are happy to take them to do—even while their functional role is not to refine our mutual representation of ways the world could be. Though Plunkett and Sundell (2013) do not use this terminology, on such an account, nondescriptive interpretations will arise as *indirect speech acts* (see, e.g., Asher and Lascarides 2001; Searle 1975): by asserting the (typically true) proposition that is determined by *their* preferred way of resolving, for example, the degree threshold for a gradable adjective,

a speaker *also* "pragmatically advocate[s] for the parameter settings by virtue of which [that proposition is] asserted" (Plunkett and Sundell 2013, 15).[12]

One difficulty is that, in cases where a contextual "parameter setting" is semantically indeterminate, this sort of account will struggle to say which proposition is asserted. Since the account takes the form of an indirect speech act account, it will struggle to say precisely which indirect speech act is performed by the speaker in the context of utterance. This isn't to say that this challenge couldn't be somehow met. It is just to note that it is not met by the sort of account that is described in Plunkett and Sundell (2013).

Another, more empirical difficulty—this one arising from the attempt to place nondescriptive interpretations under the rubric of indirect speech acts—is that, with bona fide indirect speech acts, an utterance's literal semantic content remains accessible to downstream "relational speech acts"—speech acts that are, in an intuitive sense, *anaphoric* to other speech acts in the discourse. For example:

A: Can you pass the salt?
B: Yes, I can!/No, I can't!

B's reply makes no sense if interpreted as relating to A's request.

A: Please pass the salt.
B: #Yes, I can!/#No, I can't!

B's reply in the first dialogue is licensed by the fact that A *semantically expresses a question about B's abilities*, albeit in service of a further communicative aim: requesting that B pass the salt (for discussions of this phenomenon, see Asher and Lascarides 2001; Charlow 2011).

By contrast, what the metalinguistic analysis treats as an utterance's semantic content in cases of nondescriptive interpretations does not seem to be accessible to the expected array of relational speech acts. It is surprisingly difficult to target the claimed propositional content of a nondescriptive interpretation with, for example, relational affirmation or denial.

[Context: A and B agree that Feynman is 6 feet tall, but disagree about whether that makes Feynman tall.]
A: Feynman is tall.
B: ?? No, he is below the threshold.

12 To be fair, Plunkett and Sundell (2013) do their best to prescind from semantic debates: whatever one's semantic commitments, they argue, one will require a treatment of metalinguistic negotiation-type uses (see esp. their Section 6.1). My target here is the particular model of metalinguistic negotiation-type uses they use to illustrate their account.

[Context: A and B agree that A's information is compatible with Bond being in Zurich and B's information is not]

A: Bond might be in Zurich.

B: ?? No, the information rules out Bond being in Zurich.

The infelicity in both cases seems due to misunderstanding: when A says "Feynman is tall," A does not (ordinarily) mean to be interpreted as making a comparison between Feynman's height and "the" relevant threshold. When A says "Bond might be in Zurich," A does not (ordinarily) mean to be interpreted as making a claim about the properties of the relevant information.[13] (Intuitively, in the first case, A and B are not disagreeing about the truth of any proposition of the form *Feynman's height exceeds θ*: their disagreement appears to consist in the fact that A *regards* or *considers* Feynman's height as sufficient for being tall, while B does not. Intuitively, there is a disagreement in attitude, unaccompanied by any evident disagreement in fact.)

Until we see an explanation of why speech acts that are anaphoric to the claimed propositional content of A's utterances do not appear to be licensed in such cases, it will be unclear whether the alleged propositional content of "Feynman is tall" or "Bond might be in Zurich" has *any explanatory role to play* in accounting for nondescriptive interpretations of these sentences. If we can account for such interpretations, without appeal to such alleged contents, this is at least some reason to think that we should.

A final observation about the metalinguistic account: the action of eliminating, for example, a degree threshold from eligibility *need not be* understood metalinguistically (i.e., as targeting or aiming to constrain the *linguistic behavior* of the audience).[14] I would suggest that we do better to understand this action as an attempt to constrain the downstream *sortal attitudes* of the audience—as a kind of prescription (bearing on who to regard as tall, and who to regard as not-tall). Such a prescription no doubt bears indirectly on the use of words: to regard a set of degree thresholds as *c*-eligible might commit one to regarding *utterances* whose meanings are incompatible with at least one *c*-eligible threshold as inappropriate. But

13 Unsurprisingly, these dialogues sound a lot better when the context settles the referent of quasi-technical phrases like "the relevant threshold" and "the relevant information." If it is clear that *A is comparing Feynman's height to an explicit standard* (e.g., a line drawn on a wall), B can certainly reply to A's utterance "Feynman is tall" by saying "that's wrong, Feynman is below that line." If it is clear that B is seeking information from A about *where A thinks Bond might be*, B can certainly reply to A's utterance "Bond might be in Zurich" by saying "that is a damned falsehood, I know very well that your information rules this out."

14 This observation is distinct from a claim that Plunkett and Sundell (2013) are careful to rebut, namely that the action of eliminating, for example, a degree threshold from contextual eligibility *cannot* be understood metalinguistically. On this objection, metalinguistic moves in a conversation tend to be pointless—akin to merely verbal disagreement; we misconstrue the aims of speakers if we understand them as participating in such moves. But, as Plunkett and Sundell (2013) note, how we use words matters: whether or not Steph is regarded as "tall" will license (or not) a host of downstream actions (e.g., will we trade for Steph?). Metalinguistic judgments typically have nonmetalinguistic motivations and effects.

this does not exhaust the functional role of the state of representing a set of degree thresholds as *c*-eligible: such a set constrains the kinds of things that are eligible candidates for, for example, the property of *tallness*. Hence, the set can also be regarded as a constraint on which things are treated as eligible candidates for some task requiring tallness (e.g., selection for a game of pickup basketball, if that is the *c*-relevant task).[15]

I am thus inclined to think that the constraints on linguistic behavior that arise from representing a set of degree thresholds as *c*-eligible are better understood as a *consequence* of a more fundamental feature of this sort of representational state: representing a set of degree thresholds as eligible constrains who one *regards* or *considers as tall* in *c*, which subsequently constrains who one is able to appropriately *call tall* in *c*. On this model, the metalinguistic guidance that is provided by the elimination of a degree threshold from eligibility is not intrinsic to this kind of update; it is rather a kind of natural effect of adopting (in the sense of coming to satisfy) a constraint on who to regard as tall.

5. Prescription-Type Meanings

I will instead suggest that we recognize a distinction in *semantic type* between descriptive and nondescriptive interpretations. Descriptive interpretations have the usual propositional interpretation (with provisions for semantic indeterminacy). The nondescriptive interpretations in which I am interested here are *prescriptive* in nature (also with provisions for semantic indeterminacy). Drawing on the proposal for the semantics of imperatives—a dedicated clause-type for expressing prescriptions in natural language—developed in my earlier work (see, e.g., 2011, 2014, 2018), I will propose that these interpretations are semantically distinct (but semantically related). More concretely, when an expression of semantic type T is used nonspecifically (in the simple sense that it lacks a semantic value in context) its argument place can be optionally bound (by, e.g., λ-abstraction) to yield a characteristic function of objects of type T—that is to say, a *property* of objects of type T. Properties like this are well-suited to account for nondescriptive interpretations, I will here argue.

5.1 Prescriptions and Propositions

Imperatives encode/express constraints on states that have an *action-guiding* or *motivating* functional role (e.g., plans or preferences); in particular, they tell someone who accepts or updates on the imperative what their plans or preferences must be like (see Harris 2017a;

15 As noted above, Plunkett and Sundell (2013) utilize the fact that predicational questions (e.g., do we predicate "spicy" of this chili?) are linked to nonlinguistic questions (do we add more spice to the chili?) to argue that predicational questions are not merely verbal in import. While correct, this does not establish that the illocutionary point of saying "this chili is spicy" is to answer a predicational question (do we predicate "spicy" of this chili?). My eventual suggestion here will be that we do better to think of the resolution of such predicational questions as a natural effect of the resolution of nonlinguistic *normative questions* (e.g., *should we regard this chili as spicy*?).

Portner 2004, 2007, 2018; Roberts 2015, 2018; Starr 2020). Theorists differ about whether or not imperatives do this as a matter of their semantics: for Portner and Roberts, they do not, for Charlow, Starr, and Harris, they do. Here I will be assuming that the latter position is broadly correct.[16]

In my own account, an imperative like "Confess!" is *compositionally related* to a corresponding modal claim "you must confess" (it will take me a page or so to explain how). Consider the following (schematic, extensional) representation of a modal logical form:[17]

$$Modal_{f,g}(Restrictor)(Scope)$$

This representation is what we find in Kratzer (1977, 1981, 1991): modals are generalized quantifiers expressing a quantificational relation between (1) a domain of quantification jointly characterized by the Modal Base *f*, the Ordering Source *g*, and a (explicitly or implicitly provided) Restrictor, and (2) a set of possibilities characterized by the Scope.

Like other context-sensitive expressions, prioritizing (e.g., deontic) modals admit of descriptive and nondescriptive readings: "you must confess" can be used to describe what is required at *c*, given some or other body of norms or priorities made salient at *c*, but it can also be used to *tell someone to confess*. In such uses, the modal takes on a meaning that seems to be prescriptive, exhibiting many of the same features as the meaning of the corresponding imperative (e.g., infelicity when joined with denials that the relevant obligation will be discharged) (Ninan 2005; Portner 2007).

#You must go to confession, but you're not going to.
#Go to confession! You're not going to go to confession.

About the prescriptive meaning, Portner writes:

> Since it seems that *must* has an obligation-imposing function, in addition to a traditional truth-conditional semantics, as part of its conventional meaning, the next question is what the nature of this obligation-imposing reading is. Ninan (2005) proposes to model it in terms of the notion of To-Do List (Portner 2004). Thus, he explains ["Confess!"] as follows: uttering the sentence places the property of going to confession on [the addressee's] To-Do List. But one cannot place a requirement on someone's To-Do List while at the same time asserting that it will not be met, and therefore the sentence is anomalous. What's important here is that the

16 All that will turn on this is whether the sort of contents I describe below for constraint-type interpretations are assigned in the semantics or by some "Dynamic Pragmatic" theory equipped with a mechanism like abstraction. On the Dynamic Pragmatic program, see especially Portner (2018) and Roberts (2018). For arguments against, see Charlow (2018).

17 The notion of "logical form" invoked here is syntactically neutral: I do *not* assume that representing a variable in logical form means representing a variable in morphosyntax.

> ordinary, truth-conditional semantics for the modal does not play a role in explaining the patterns [above]. Rather, the independent imperative-like meaning does the job. (2007, 365–66)

Portner here assumes that any obligation-imposing (prescriptive) meaning carried by the modal would be "*independent*" of its "truth-conditional" (i.e., quantificational) semantics. Here, however, are two empirically viable possibilities for modeling the prescriptive meaning carried by prescriptive "must"; the truth-conditional dimension of the modal's meaning plays an essential role in both.

- **Performative**: Prescriptive interpretations consist in proposals to adjust a salient body of norms (or more generally a salient state with an action-guiding functional role) so that the modal proposition semantically expressed by the sentence is true relative to the adjusted body of norms.[18]

Like the metalinguistic account, a performative account makes central explanatory use of the proposition allegedly expressed by a sentence like "Confess!"; I will set such accounts to the side here (for arguments against performative accounts, see Charlow 2018).

- **Modally Derived**: Prescriptive interpretations of imperatives consist in proposals to adjust a salient body of norms (or more generally a salient state with an action-guiding functional role) so that it comes to satisfy a modally characterized property (see Charlow 2011, 2014, 2018).

The suggestion here is to represent the semantic content of both "Confess!" and "You must confess!" (on its prescriptive interpretation) at a context c with the same modally characterized property, namely:

$$\lambda g[must_{f_c,g}(addr_c \ goes\ to\ confession)]$$

This is the property an ordering source g has, iff all the g-best possibilities compatible with the c-relevant information f_c are possibilities where $addr_c$ goes to confession. Less technically, it is the property g has when g induces a *ranking* (more precisely, an ordering) on the possibilities compatible with f_c, according to which possibilities where the addressee of c goes to confession are *highest-ranked*. Less technically still, it is the property g has when, according to g, it is preferred/planned/required/ . . . that the addressee of c goes to confession.

18 More precisely, one proposes to alter a body of norms that characterizes a domain of quantification for the modal "must" so that all worlds compatible with the domain of quantification are worlds in which the prejacent proposition (e.g., that the addressee goes to confession) holds. For accounts of imperatives in this vein, see especially Lewis (1979) and Kaufmann (2012).

What links property-type contents of this sort to illocutionary forces or discourse moves? Such contents are properties that can, in a sense, be *instantiated* by motivating psychological states (e.g., an agent's plans).[19] It is natural to say that the force of an imperative is to propose a selected or salient motivating state—typically that of the addressee—come to instantiate this property.[20] The following Force Assignment principle gives the rough idea.

Force Assignment
The force of expressing a property of an ordering source at *c* is to propose that the action-guiding/motivational state(s) of one's addressee(s) at *c* satisfy this property.

"Satisfaction," in the relevant sense, amounts to *representability*: an agent's action-guiding or motivating state satisfies $\lambda g[must_{f_c,g}(R)(S)]$ just when that state is representable with some g such that $must_{f_c,g}(R)(S)$ is true. So, for example, the force of "Confess!" at *c* is to propose that the motivational state of one's addressee come to be representable with priorities that require going to confession. Imperatives conventionally express directives, in the following sense: imperatives are conventionally associated with attempts to get their addressee(s) to be motivated or adopt plans with specific modally characterized properties.

5.2 Extending the Account

The strategy described in the last section is easily extended beyond imperatives. In general, when a speaker means to *express a property of some parameter* (for the sake of *adoption* by some agent in the context), rather than to describe a feature of this parameter, we will say:[21]

- The **semantic value** expressed by the speaker is a λ-abstract over a free (contextually unbound) occurrence of that parameter.[22]

---

19 Strictly speaking, I do not want to assume that there is any kind of tight/conventional link between expressing a property of *an ordering source* and the imperatival speech act of direction. In contexts in which an ordering source represents a body of *expectations* (as opposed to a body of plans or intentions), expressing a property of an ordering source means proposing that the expectations of one's addressee satisfy this property. Ordering sources come in different "flavors" depending on what information they are used to represent, and the force of expressing a property of an ordering source will depend on what type of information the ordering source is used to represent. (This sketchy suggestion is meant to build on Kratzerian (1981) orthodoxy about "modal flavor" (i.e., polysemy). That said I am doubtful that we in fact have a good understanding of what modal flavor is/how it is determined relative to a context of utterance.)

20 For a more complete statement of the account, see Charlow (2018).

21 For a related view of semantic composition and its metasemantic implications, see discussion (and references) in Harris (2017b).

22 There are different compositional routes to this semantic value. One familiar possibility is to treat the parameter as a free variable and apply the Predicate Abstraction rule of Heim and Kratzer (1998, 114). A less familiar, but also attractive, possibility eliminates variables (and variable assignments) from the metalanguage, instead treating semantically underspecified expressions as semantic placeholders, whose values are resolved "post-semantically" (see especially the "Variable-Free" system of Jacobson 1999).

- The **illocutionary force** of expressing such a meaning is proffering this property for adoption by the addressee in the context.

Proffering a property for adoption by another agent is an *inherently prescriptive* act: it amounts to offering a kind of *cognitive advice* (adapting a phrase of Swanson 2006, 39). It is therefore important to distinguish this general notion of prescriptive force from the more particular notion of prescriptive force that is relevant to the analysis of imperatives. Imperatival prescriptive force is *practical* or *directive* in nature: it bears on an agent's motivating or action-guiding psychological states; it tells the agent how she should plan (and therefore how she should act). Prescriptive force need not be practical or directive, in this more particular sense: when a speaker proffers a property of a nonmotivating (e.g., doxastic) state for adoption, she is not telling the addressee what to do; she is telling the addressee what, say, her beliefs must be like.

MacFarlane, recall, treats gradable adjectives as encoding practical information: information that constrains an agent's plans, rather than some other (e.g., nonmotivating) state of mind. This is *optional* on the framework I am proposing: while MacFarlane and I agree that gradable adjectives (in cases of nondescriptive uses) proffer prescriptions, I have here distinguished two senses of prescription: practical and more broadly cognitive prescriptions. The question of whether nondescriptive interpretations of gradable adjectives are best represented with practical or cognitive prescriptions remains open.

Consider again "Steph is tall" as uttered a context in which the sentence receives a nondescriptive interpretation—that is, roughly, a context in which the speaker is encouraging her addressee to *think of Steph as tall* (while not attempting to offer any descriptive information about Steph's height). Is the speaker well-understood as suggesting that the addressee *adopt a plan for* thinking Steph tall?

I do not think so. For one thing, thinking Steph tall—in the stative, rather than eventive, sense—isn't under the addressee's voluntary control, and typically isn't the sort of thing that you can plan to do. (This is why an imperatival expression of this kind of planning content, like "regard Steph as tall!," is generally heard as marked.[23]) A theory should not blur the distinction between the type of prescriptive force relevant to the analysis of imperatives (which is well-modeled with planning content) and the type of prescriptive force associated with proffering a property for cognitive adoption.

23 MacFarlane tends to use eventive language (e.g., where to "*draw the line*" between tall and non-tall) to describe the kind of cognitive question that a gradable adjective in the positive form is meant to resolve. There is nothing grammatically marked about telling one's addressee *to draw the line* between tall and non-tall in such a way that Steph counts as tall. Still, how I draw this line isn't really up to me: how I draw the line determines who I think tall, and who I think tall isn't really up to me. It therefore seems like a distortion (prima facie) to represent the cognitive question that a gradable adjective in the positive form is meant to resolve as a question about where to plan to draw the line between tall and not.

A small revision of MacFarlane's semantics avoids this worry. Letting *c* be a context in which "Steph is tall" receives a nondescriptive interpretation, we will say:[24]

$$[\![\textit{Steph is tall}]\!]^c = \lambda\theta.deg_{tall}(Steph) > \theta(deg_{tall})$$

This is the property a higher-order threshold function has if it maps the degree measure deg_{tall} into a value that is lower than Steph's degree of tallness. This property is the sort of property an addressee might be encouraged to "cognitively instantiate" by a speaker (whose communicative aim is to get the addressee to share their appraisal of Steph as tall).

It is already common ground in this debate that agents cognitively represent entities of type θ—such entities are, after all, represented in the dominant theory of what a speaker's *semantic competence* with respect to gradable adjectives consists in.[25] What will be controversial between a proponent of the Kennedy (2007) view and myself is the *functional role* of this sort of representation.

According to Kennedy's own interpretation of his semantics for gradable adjectives (and any form of Contextualism about gradable adjectives with which I am familiar), representing some $\theta_c(deg_{tall})$ as eligible in a context *c* is to represent *c* as being a certain way—for example, as being such that $\theta_c(deg_{tall})$ is the *c*-relevant threshold for the minimum degree of tallness needed to count as tall (see Kennedy 2007, 17ff.). Given this account—which is

24 The context does not bind anything to its right in this semantic proposal; this is a simplification. Context can fix the values of various parameters relevant to evaluating a tallness claim (e.g., a relevant comparison class, a discourse task that would be resolved with a way of sorting individuals in the comparison class into tall and not-tall, etc.). According to the proposal being advanced here, we would not be surprised to observe nondescriptive uses targeting these contextual parameters (e.g., expressions of *properties of* comparison classes or *properties of* discourse tasks). My view is that such interpretations are likely attested. This is *not* to say that a speaker can freely express a property of any contextual parameter—in fact, speakers' freedom in this domain appears to be tightly constrained: when an expression of semantic type *T* is *indexical* in nature, speakers do not use sentences embedding such an expression to express properties of objects of type *T*. (As noted above, speakers do not seem to use sentences embedding "that" to express a view about the appropriate referent of "that" in their context.) Khoo (2017) introduces a model that provides an appealing explanation of this fact (as well as an appealing characterization of the phenomenon of indexicality in general). On Khoo's analysis, the values of indexical expressions (here understood to include demonstratives) are very tightly anchored to "objective" features of the context—features that are not freely modulated by speech acts. (As mentioned in section 3.2, it doesn't appear that a speaker can use "that's a beautiful car" to express a prescription governing the use of "that.") Nonindexical context-sensitive expressions (which Khoo dubs "quasi-indexical") work differently: speakers can use these expressions in a way that is not deferential to objective features of the context, as a way of modulating the features of the context toward which a referential use would ordinarily be sensitive.

25 It is unlikely that speakers have what Harris (2017b) calls "central access" ("central" in the sense of Fodor 1983) to this sort of entity, and so it is unlikely that speakers will have central access to -abstracts in which the -term binds variables over such entities. Nothing in my account assumes central accessibility, in the relevant sense (e.g., I do not assume that speakers have de dicto intentions to express the semantic contents that figure in my account).

very well-suited to accounting for Specific + Descriptive readings of gradable adjectives in the positive form (and, with a little tinkering, Nonspecific + Descriptive readings too)—it would hardly be extravagant to add that agents might also target candidate thresholds with *normative judgments*: agents might (indeed, obviously do) regard certain candidate thresholds as *appropriate* (or *inappropriate*) ways of drawing the boundary between tall and not-tall.

In general, we will say that an agent regards a threshold $\theta(\delta)$ **as appropriate** iff $\theta(\delta)$ is consistent with her sortal attitudes toward any object x such that $\delta(x)$ is defined—so, for example, an agent regards $\theta(deg_{tall})$ as an appropriate threshold for tallness iff, for any x with some degree of tallness, the agent regards x as tall only if $deg_{tall}(x) > \theta(deg_{tall})$. It seems unlikely that the state of regarding $\theta(\delta)$ as appropriate can be understood as representing some way the world or context could actually be: $\theta(\delta)$ is not plausibly treated as a candidate for actuality (here again see MacFarlane 2016). On the face of things, the functional role of this sort of state is exhausted by its characterization (or, perhaps, determination or regulation) of an agent's sortal attitudes.[26]

Let $\Theta_{x,c}$ designate the set of thresholds that are consistent with an agent x's normative judgments about what is an appropriate way of drawing the boundary between tall and not-tall in c.[27] I have claimed that, in a context in which "Steph is tall" receives a nondescriptive interpretation, a speaker semantically expresses the following property of higher-order thresholds: $\lambda\theta.deg_{tall}(Steph) > \theta(deg_{tall})$ (henceforth abbreviated λ_θ). It is straightforward to say how an addressee representable with $\Theta_{addr_c,c}$ will respond to a speaker who semantically expresses this

26 Similarly, in the modal/conditional domain, there is evidence that the functional role of attitudes toward modal/conditional sentences cannot be understood as a representation of some way things could be. Yalcin (2011, 2012) adduces cognitive evidence that the class of possible descriptive contents fails to cover a core range of uses of epistemic and probabilistic talk. Charlow (2016b) and Russell and Hawthorne (2016) muster evidence from formal epistemology (Triviality Results) to suggest that a core range of uses of conditional and modal sentences *cannot* be assigned a propositional semantic value in context. And, of course, there is the famous result of Gibbard (1981), which shows that, if the indicative conditional is a two-place operator that respects modus ponens and import-export, any proposition expressed by the indicative conditional would have to be equivalent (pace the facts) to the proposition expressed by the material conditional.

27 This is not intended as a thesis about the content of the state of regarding a certain threshold as appropriate or eligible (although I am sympathetic to such an account for prioritizing modals; see Charlow 2018, Sec. 5). The account given in this paper is meant to remain neutral on questions about how to represent the content of this sort of state of mind, as well as questions about how to formally represent the update an addressee performs when she updates on a normative judgment bearing on what is an appropriate way of drawing the boundary between tall and not-tall in c. I do not assume that update goes via intersection of set-theoretically represented contents (although I utilize intersective operations to represent certain features of such an update). More generally, I do not assume that it is the job of linguistic theorizing to characterize a way of updating on a piece of semantic content. As I have argued elsewhere, requiring a theory to characterize such an update function will ultimately mean writing a (epistemological) theory of rational attitude revision into our representation of semantic competence—something to be avoided (Charlow 2014, 2016a).

property (if she comes to accept the speaker's utterance): she will no longer regard candidate thresholds that fail to satisfy λ_θ as appropriate, and so the set of thresholds that are consistent with her normative judgments about what is an appropriate way of drawing the boundary between tall and not-tall in *c'* (where *c'* is a context posterior to *c*, in which $addr_c$ has come to accept the speaker's utterance *c*) will be given by:

$$\Theta_{addr_c,c'} = \Theta_{addr_c,c} \bigcap \lambda_\theta$$

As with practical prescriptions (as expressed by imperatives), cognitive prescriptions (as associated with a nondescriptive interpretation of "Steph is tall") are associated with the illocutionary act of proffering a property for cognitive adoption—a property that is not assumed to always correspond to a state of representational belief in the truth of some proposition. States that do not correspond to states of representational belief divide into practical (planning, action-guiding) states and nonpractical states (e.g., regarding some threshold as an appropriate way of sorting agents into tall and not-tall) that play a *determinative role in fixing an agent's sortal attitudes*. States of the latter kind do not play a determinative role with respect to an agent's practical states, e.g., her intentions (although of course the sortal attitude associated with thinking Steph tall can *cause* an agent to have intentions, e.g., the intention to pick Steph in our pickup game).

This section has argued that speakers can, and do, semantically express properties of contextual parameters, and that this kind of *locutionary* act can be married to a satisfying account of the *illocutionary* function of such locutionary acts—that is, proffering this property for adoption by the addressee in the context. The account was illustrated with gradable adjectives but could be extended with a bit of effort to the "informational" parameter against which modals are semantically evaluated.

Because semantic determinacy is orthogonal to prescriptiveness, the model I have described here does not bear directly on the issues of indeterminacy canvassed above. This is deliberate.[28] Speakers can, for this reason, semantically express a *range of candidate properties* for adoption by their addressees. Recall the following case:

[Context: A/B could be referring to either the car type or token.]
A: That might be a Ferrari.
B: No, but it might be a Maserati.

The account here will analyze this case as one in which B semantically expresses more than one candidate property for adoption by A—the property of not ruling out possibilities in

28 This is not to say that these issues are unconnected: handling the problems arising from the phenomenon of semantic underspecification *and* those arising from nondescriptive uses will require admitting exceptions to Referentialist metasemantics.

which the car token is a Maserati, as well as the property of not ruling out possibilities in which the car's type is the type *Maserati*. Assuming that the car token is a Maserati iff its type is *Maserati*, it seems B's communicative aims are realized regardless of which property A adopts. I see no reason to say that the metasemantics of natural language demands that B express one of these properties but not the other, when the realization of B's communicative aims does not.

6. Coordination and Expressivism

I have argued that speakers semantically can use semantically "underspecified" language to express properties of semantic parameters (like states of information and degree thresholds), and thereby express an attitude — loosely, the attitude of (being representable as) satisfying this property. How and why does *language* provide for this sort of thing?

So far as the *function* of language is concerned, I would not try to improve on the story told in Gibbard (1990) (which is similar to, but in certain respects more general than, the story told in Lewis 1969).[29] These introductory remarks give a good flavor of the account:

> The need for complex coordination stands behind much of the way language works in our thoughts, in our feelings, and in social life. It figures centrally in our emotional dispositions, especially for such morally significant emotions as outrage, guilt, shame, respect, moral admiration, and moral inspiration. Matters of coordination, in the picture I shall sketch, stand squarely behind the psychology of norms, and hence behind what is involved in thinking something rational or irrational. Primitive human life is intensely social. In the conditions under which we evolved, anyone's prospects for survival and reproduction depended crucially on the beneficial human bonds he could cultivate. Human cooperation, and coordination more broadly, has always rested on a refined network of kinds of human rapport, supported by emotion and thought.
>
> (Gibbard 1990, 26)

The practice of proffering properties (of semantic parameters) for cognitive acceptance or adoption—regardless of whether the property is one whose psychological instantiation is equivalent to representing the world a certain way—is a practice that plausibly facilitates coordination in causally and behaviorally significant features of informational, sortal, and motivational psychological states. It is no surprise that, if the account of this paper is on the right track, language users would avail themselves of a dedicated type of content for performing this sort of speech act.

29 For recent appeals to coordination in a pragmatic account of nondescriptive language use, see Yalcin (2011, 2012) and Charlow (2015).

This is an Expressivist theory (for reasons I hope will be at least somewhat apparent to the reader). Notice in particular the following core presupposition of this account: that there is a difference in meaning between the speech act of *expressing* (for example) a particular sortal attitude (the content of which I represent as a constraint on, or property of, degree thresholds) and the speech act of *saying that* one has that same attitude (the content of which I represent with an ordinary proposition, to the effect that one's own threshold for considering an object, say, to be tall has a certain characteristic, say, exceeding 183cm).

The account, however, departs from traditional Expressivist theories like Gibbard's, in one or two significant ways, on which I will end the paper by reflecting. First, and most obviously, Gibbard, perhaps owing to a background commitment to the Humean Theory of Motivation, recognizes two "kinds" of "content": propositional content (content which bears on what the world is like, modeled with sets of possible worlds) and planning content (content that bears on how to plan, modeled with sets – equivalently, properties—of Normative Systems or Hyperplans). For Gibbard (and, we have seen, for MacFarlane as well) non-propositional content is generally theorized as *planning* content. This paper has argued that non-propositional content comes in more varieties than planning content. Although we do express planning attitudes (as well as descriptive beliefs) with language, these are not the only types of states of mind (which are not equivalent to descriptive belief in the truth of a propositional content) that we use sentences of natural language to express. Far from it. Indeed, although I have not argued for it here, I am drawn to the thesis that, *whenever* a sentence's semantic value in context is semantically "parametrized" — whether to a state of information (as with epistemics), an experiencer (as with experiential language, e.g., predicates of personal taste like "tasty"), a degree threshold (as with gradable adjectives), a plan (as with practical language, like imperatives and deontic modals), etc. — an utterance of that sentence can be used by a speaker to express a property of the relevant parameter, and thereby to express an attitude that can be modeled using a set of such parameters. Expressivism's insights can be fruitfully extended and generalized to many different types of language and language use, provided we are willing to entertain the sort of "polymorphism" (in essence, type-heterogeneity) about semantic content at which I am gesturing here. (On content polymorphism for epistemics, see my 2020.)

Second, Gibbard has offered a broadly Gricean account of the function of the speech act *expressing an attitude* (see also Gibbard 2003, 78ff.):

> Suppose Caesar tells Cleopatra, "I was captured by pirates in my youth." Why might he do this? Assume he is simply informing her about his youth; the story, then, will be something like this. He wants her to know about his capture by pirates. He thinks she lacks true belief on the subject, but he believes that she thinks him sincere and that she thinks him an authority on events of his youth. Here to be sincere is to express only beliefs one actually has, and to be an authority on something is to be quite unlikely to be mistaken about it. Caesar thus intends to get Cleopatra to believe that he was captured by pirates in his youth, and to do so in the

> following manner. He utters words that conventionally purport to express, on the part of any speaker, a belief that he was captured by pirates in his youth. He intends her to come to accept that he has that belief, and to do so in virtue of her recognition of this intention. Since she takes him to be sincere, she has reason to accept, upon hearing his words, that he does believe that he was captured by pirates in his youth. Since she thinks him an authority on his youth, she concludes from his believing it that he indeed was captured by pirates in his youth.
>
> (Gibbard 1990, 85).

On Gibbard's broadly Gricean account, a speaker *S* expresses state of mind *M* in order to get her addressee to form the belief that *S* is in *M* (a belief that will, under the right conditions, get the addressee to adopt *M* herself). Coordination in attitude is a natural effect (in a context in which the addressee recognizes the speaker as an authority on the relevant subject-matter) of the addressee forming a *belief about the speaker's state of mind*, on the basis of the speaker making a linguistic performance that indicates her possession of that state of mind.

Here, however, is a difficulty with this explanation.[30] We *have* dedicated linguistic devices (i.e., attitude ascriptions) for telling our addressees what states of mind we are in: instead of expressing his belief that he was captured by pirates in his youth (by asserting that he was), Caesar can self-ascribe the belief (by reporting to Cleopatra that he believes he was captured by pirates in his youth). If the speaker's aim is coordination in attitude, and coordination in attitude is explained by an addressee's belief about the speaker's attitude, speakers could realize the same communicative aim by reporting themselves to be in the relevant state. This, I will argue, presents a threat to the claimed explanatory role of non-propositional (e.g. planning-type) semantic content in Gibbard's broader theory.

To better see the threat, let us consider a "Subjectivist" alternative to an Expressivist account of normative claims. According to the Subjectivist alternative, when someone asserts that *x* is rational, they are semantically expressing a proposition about their own planning attitudes: roughly, the proposition that, according to their plans, *x* is permitted. (That is to say, according to the Subjectivist, they are semantically ascribing a plan permitting *x* to themselves.) Piggybacking on Gibbard's pragmatic theory, the Subjectivist might say the speaker typically does this in order to get their addressee to come to accept that their planning attitudes are this way; insofar as the addressee regards the speaker as an authority on whether to have plans that permit *x*, the addressee will have reason to share/adopt the speaker's planning attitude toward *x*. This Subjectivist theory works much the same as Gibbard's. And so one may begin to wonder what explanatory role the assignment of *non-propositional semantic content* is supposed to fill in Gibbard's semantic and pragmatic theory for normative language.

30 For a related critique, see Schroeder (2008, Section 4).

To forestall an obvious reply, it is true that the Subjectivist theory does not directly account for the (evident) difference in meaning between an attitude ascription in the mold of (1) and a corresponding attitude ascription in the mold of (2).

(1) Beth believes (says/agrees/disagrees) that *x* is rational.
(2) Beth believes (says/agrees/disagrees) that her plans permit *x*.

It does not follow that this difference in meaning is *incompatible* with the Subjectivist theory. Subjectivism, as stated above, is purely a thesis about the semantic content of an utterance of "*x* is rational" (relative to a context of utterance). It incurs no direct commitments regarding the semantic content of such a clause, as *embedded* under an attitude or illocutionary verb. To see this more clearly, consider a version of Subjectivism according to which:

- At a context *c* providing a variable assignment *g*, "*x* is rational$_n$" expresses the proposition that *g*(*n*) permits *x*.
- By default, *g*(*n*) is the planning attitude of *c*'s speaker.

It is a live (indeed quite plausible) possibility in semantic theory that attitude and illocutionary verbs *quantify over* (and thereby shift) variable assignments (in addition to quantifying over more familiar objects like possible worlds) (see e.g. Santorio 2012). Adapting the idea, it is a live possibility, for the Subjectivist, that the truth condition of "Beth believes that *x* is rational" is roughly that, for any planning attitude *n* (such that *n* is compatible with Beth's plans), *n* permits *x*. This truth condition — which we can gloss as "Beth's plans permit *x*" — is evidently distinct from the truth condition of "Beth believes that her plans permit *x*". Note that this form of Subjectivism agrees with Gibbard about the content of Beth's belief that *x* is rational: that its content is best represented, roughly, with a *plan* (rather than a proposition).

Expressivism, then, or Subjectivism? I will try to explain why I still incline toward Expressivism (although Gibbard's theory of semantic interpretation means he will struggle to distinguish his Expressivist theory in similar fashion). The Subjectivist disagrees with the Expressivist about the *semantic content* of Beth's assertion that *x* is rational. Their reasons (as I imagine them) are something like this: on our best theory of assertion (Stalnaker's), the essential effect of an assertion is to update the Common Ground with a propositional content. Now, there is no doubt that a speaker who asserts that *x* is rational *does* generally make it Common Ground that *their* plans permit *x*. Since the proposition that the speaker's plans permit *x* is generally part of the post-assertion Common Ground anyway in such cases, it seems to make good theoretical sense (for the Subjectivist) to say that such a speaker semantically expresses (locutes) the proposition that their plans permit *x* in the course of asserting (illocuting) that *x* is rational. In contexts where we have reason to interpret the speaker as endorsing this feature of their plan as a basis for rational coordination, prescription-type interpretations of claims like "*x* is rational" will arise in broadly the same way as on Gibbard's theory.

This form of Subjectivism shares with Gibbard's theory an apparent aversion to the idea of what I'll call *intrinsically practical content* — content the apprehension and acceptance of which can "directly" constrain a planning state (perhaps, again, owing to background Humean assumptions). For both Gibbard and the Subjectivist, interpreters who accept speaker's claim that *x* is rational must be represented as reasoning their way from beliefs about the speaker's communicative intention to a (self-directed) normative judgment that subsequently constrains the interpreter's planning state, deploying something like the following syllogism.

(1) The speaker intends for me believe that their plans permit *x*.
(2) Given (1), the speaker intends for me to have plans that permit *x*.
(3) Given (2), my plans ought to permit *x*.
(4) So, my plans ought to permit *x*. [Conclusion: revise plans to permit *x*.]

Compare this to the following, rather simpler, account of the reasoning involved in accepting a speaker's claim that *x* is rational.

(1) The speaker expresses a way of planning that permits *x*.
(2) To accept the way of planning the speaker expressed, my plans must permit *x*.
(3) So: revise plans to permit *x*.

Expressing (we might also say "proffering") a way of planning that permits *x*, as I understand the notion, does not imply that one intends their addressee to believe that their plans permit *x*. This is a more flexible, less representationally committed, paradigm for understanding what internal representations an interpreter deploys when they accept a speaker's claim that *x* is rational. Notice, for instance, that the paradigm is easily fitted to cases of "selfless" direction, in which it is Common Ground that a speaker is expressing the view that *x* is permitted (a way of planning that permits *x*), despite having plans that prohibit *x*. The Gibbard-Subjectivist paradigm is not.

Another reason to want this kind of flexibility is the *natural ubiquity* of prescription-type content. Creatures with limited meta-representational capabilities comprehend prescriptions, by interpreting apprehended (non-linguistic) signals as attempts to constrain their behavior. A warning call is interpreted as carrying a prescriptive message — *(you should) tread carefully!* Accepting the message, so interpreted, means that the addressee's internal state is organized in such a way that they are behaviorally disposed to tread carefully. The intentional content of a pain experience can be productively, if partially, theorized as prescriptive — *(you need to) stop this!* The subject of a pain experience is theorized as the recipient of this message; accepting the message means having a plan that requires stopping the pain. In cases like these, must we represent the subject as arriving at a decision to accept the message's instruction as reasoning with representations about the "intentions" of

the message's producer? I think not: the subject simply *apprehends an instruction*, or way to plan, and decides — possibly, but not necessarily, after engaging in higher-level reasoning about the source of the instruction — whether or not to adjust their plans accordingly. Of course, I do not deny that interpreters often (perhaps always) utilize representations about the source of the instruction in trying to identify *which* instruction the source "means" to be transmitting. (Is the pain being sent by the stomach or the heart? To whom is the warning call directed?) Once an interpreter determines the content of the relevant instruction, the interpreter decides to accept it or not. While it is *possible* (depending, of course, on their representational capabilities) for the interpreter at this stage to engage in further practical reasoning, to try to determine whether or not it is a *good idea* to accept the interpreter's instruction, it does not seem to be a prerequisite.

So both Gibbard and the Subjectivist are, I believe, mistaken about the manner in which prescriptive (including planning) content is generally apprehended and accepted. In Gibbard's account (and the Subjectivist's), generating a prescriptive interpretation of a message is a *side-effect* of forming a specific belief about the internal state of the message's source. In our account, speakers express properties of cognitive parameters, and they do so in order to proffer those properties (directly) for cognitive adoption (acceptance) by their addressees. In such cases, a cognitive constraint — a *way of representing* on some question or issue — is inherent in the content that an addressee apprehends, when they apprehend the speaker's message, and the addressee must either accept the message (by representing the question that way) or not (otherwise). Language simply and directly provides speakers the tools for conveying cognitive prescriptions of sundry types and flavors, facilitating wide-ranging coordination in causally significant features of our internal states.

References

Asher, Nicholas, and Alex Lascarides (2001). "Indirect Speech Acts." *Synthese* 128: 183–228. https://doi.org/10.1023/A:1010340508140.

Barker, Chris (2002). "The Dynamics of Vagueness." *Linguistics and Philosophy* 25: 1–36. https://doi.org/10.1023/A:1014346114955.

Braun, David, and Theodore Sider (2007). "Vague, So Untrue." *Noûs* 41: 133–56.

Charlow, Nate (2011). "Practical Language: Its Meaning and Use." www.natecharlow.com/work/dissertation.pdf. Accessed August 2021.

——— (2014). "Logic and Semantics for Imperatives." *Journal of Philosophical Logic* 43: 617–64. https://doi.org/10.1007/s10992-013-9284-4.

——— (2015). "Prospects for an Expressivist Theory of Meaning." *Philosophers' Imprint* 15: 1–43.

——— (2016a). "Decision Theory: Yes! Truth Conditions: No!" In *Deontic Modality*, edited by N. Charlow and M. Chrisman, 47–81. Oxford: Oxford University Press.

——— (2016b). "Triviality for Restrictor Conditionals." *Noûs* 50: 533–64. https://doi.org/10.1111/nous.12111.

——— (2018). "Clause-Type, Force, and Normative Judgment in the Semantics of Imperatives." In *New Work on Speech Acts*, edited by D. Fogal, D. Harris, and M. Moss, 67–98. Oxford: Oxford University Press.

——— (2020). "Grading Modal Judgement." *Mind* 129: 769–807. https://doi.org/10.1093/mind/fzz028.

Dowty, David (1985). "On Recent Analyses of the Semantics of Control." *Linguistics and Philosophy* 8: 291–331. https://doi.org/10.1007/BF00630916.

Fodor, Jerry (1983). *The Modularity of Mind*. Cambridge: MIT Press.

Gibbard, Allan (1981). "Two Recent Theories of Conditionals." In *Ifs*, edited by W. Harper, R. Stalnaker, and G. Pearce, 211–47. Dordrecht: D. Reidel.

——— (1990). *Wise Choices, Apt Feelings*. Cambridge: Harvard University Press.

——— (2003). *Thinking How to Live*. Cambridge: Harvard University Press.

Harris, Daniel W. (2017a). "Imperative Inference and Practical Rationality" (Forthcoming in *Philosophical Studies*) https://doi.org/10.1007/s11098-021-01687-0.

——— (2017b). "Semantics without Semantic Content." (Forthcoming in *Mind and Language*) http://doi.org/10.1111/mila.12290.

Heim, Irene, and Angelika Kratzer (1998). *Semantics in Generative Grammar*. Oxford: Blackwell.

Jacobson, Pauline (1999). "Towards a Variable-Free Semantics." *Linguistics and Philosophy* 22: 117–85. https://doi.org/10.1023/A:1005464228727.

Kaufmann, Magdalena (2012). *Interpreting Imperatives*. Dordrecht: Springer.

Kennedy, Christopher (2007). "Vagueness and Grammar: The Semantics of Relative and Absolute Gradable Adjectives." *Linguistics and Philosophy* 30: 1–45. https://doi.org/10.1007/s10988-006-9008-0.

Khoo, Justin (2017). "Quasi-Indexicals." *Philosophy and Phenomenological Research* 100, 1: 26–53. http://doi.org/10.1111/phpr.12519.

King, Jeffrey C. (2014). "The Metasemantics of Contextual Sensitivity." In *Metasemantics: New Essays on the Foundations of Meaning*, edited by A. Burgess and B. Sherman, 97–118. Oxford: Oxford University Press.

——— (2017). "Felicitous Underspecification and Updates."

——— (2018). "Strong Contextual Felicity and Felicitous Underspecification." *Philosophy and Phenomenological Research* 97: 631–57. https://doi.org/10.1111/phpr.12393.

Kratzer, Angelika (1977). "What 'Must' and 'Can' Must and Can Mean." *Linguistics and Philosophy* 1: 337–55. https://doi.org/10.1007/BF00353453.

——— (1981). "The Notional Category of Modality." In *Words, Worlds, and Contexts*, edited by H. Eikmeyer and H. Rieser, 38–74. Berlin: De Gruyter.

——— (1991). "Modality." In *Semantics: An International Handbook of Contemporary Research*, edited by A. von Stechow and D. Wunderlich, 639–51. Berlin: De Gruyter.

Lewis, David (1969). *Convention*. Cambridge: Harvard University Press.

——— (1979). "A Problem about Permission." In *Essays in Honour of Jaakko Hintikka*, edited by E. Saarinen, R. Hilpinen, I. Niiniluoto, and M. B. Provence Hintikka, 163–79. Dordrecht: D. Reidel.

——— (1980). "Index, Context, and Content." In *Philosophy and Grammar*, edited by S. Kanger and S. Öhman, 79–100. Dordrecht: D. Reidel.

MacFarlane, John (2011). "Epistemic Modals Are Assessment-Sensitive." In *Epistemic Modality*, edited by B. Weatherson and A. Egan. Oxford: Oxford University Press.

——— (2014). *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Oxford University Press.

——— (2016). "Vagueness as Indecision." *Aristotelian Society Supplementary Volume* XC: 255–83. https://doi.org/10.1093/arisup/akw013.

Moss, Sarah (2013). "Epistemology Formalized." *The Philosophical Review* 122: 1–43. https://doi.org/10.1215/00318108-1728705.

——— (2015). "On the Semantics and Pragmatics of Epistemic Modals." *Semantics and Pragmatics* 8: 1–81.

Ninan, Dilip (2005). "Two Puzzles about Deontic Necessity." In *New Work on Modality, Vol. 51 of MIT Working Papers in Linguistics*, edited by J. Gajewski, V. Hacquard, B. Nickel, and S. Yalcin. Cambridge: MITWPL. http://semanticsarchive.net/Archive/WZlNTg0Y/.

Plunkett, David, and Timothy Sundell (2013). "Disagreement and the Semantics of Normative and Evaluative Terms." *Philosophers' Imprint* 13: 1–37. http://hdl.handle.net/2027/spo.3521354.0013.023.

Portner, Paul (2004). "The Semantics of Imperatives within a Theory of Clause Types." In *Proceedings of SALT 14*, edited by Robert B. Young, 235–52. Ithaca: CLC Publications. http://semanticsarchive.net/Archive/mJlZGQ4N/.

——— (2007). "Imperatives and Modals." *Natural Language Semantics* 15: 351–83. https://doi.org/10.1007/s11050-007-9022-y.

——— (2018). "Commitment to Priorities." In *New Work on Speech Acts*, edited by D. Fogal, D. Harris, and M. Moss, 296–316. Oxford: Oxford University Press.

Roberts, Craige (2018). "Speech Acts in Discourse Context." In *New Work on Speech Acts*, edited by D. Fogal, D. Harris, and M. Moss, 317–59. Oxford: Oxford University Press.

——— (2015). "Conditional Plans and Imperatives: A Semantics and Pragmatics for Imperative Mood." In *Proceedings of the 20th Amsterdam Colloquium*, edited by T. Brochhagen, F. Roelofsen, and N. Theiler, 353–62. Amsterdam: University of Amsterdam: Institute for Logic, Language, and Computation.

Russell, Jeffrey Sanford, and John Hawthorne (2016). "General Dynamic Triviality Theorems." *The Philosophical Review* 125: 307–39.

Schroeder, Mark (2008). "Expression for Expressivists." *Philosophy and Phenomenological Research* 76: 86–116. https://doi.org/10.1111/j.1933-1592.2007.00116.x.

Searle, John R. (1975). "Indirect Speech Acts." In *Syntax and Semantics, Vol. 3: Speech Acts*, edited by P. Cole and J. Morgan, 59–82. New York: Academic Press.

Stalnaker, Robert (1978). "Assertion." In *Syntax and Semantics, Vol. 9: Pragmatics*, edited by P. Cole, 315–32. New York: Academic Press.

——— (1984). *Inquiry*. Cambridge: MIT Press.

Starr, William (2020). "A Preference Semantics for Imperatives." *Semantics and Pragmatics*, forthcoming. http://dx.doi.org/10.3765/sp.13.6.

Stevenson, C. L. (1937). "The Emotive Meaning of Ethical Terms." *Mind* 46: 14–31.

Swanson, Eric (2006). "Interactions with Context." http://www-personal.umich.edu/~ericsw/Swanson,%20Interactions%20with%20Context.pdf. Accessed August 2021.

——— (2016). "The Application of Constraint Semantics to the Language of Subjective Uncertainty." *Journal of Philosophical Logic* 45: 121–46. https://doi.org/10.1007/s10992-015-9367-5.

Veltman, Frank (1996). "Defaults in Update Semantics." *Journal of Philosophical Logic* 25: 221–61. https://doi.org/10.1007/BF00248150.

Yalcin, Seth (2011). "Nonfactualism about Epistemic Modality." In *Epistemic Modality*, edited by A. Egan and B. Weatherson, 295–332. Oxford: Oxford University Press.

——— (2012). "Bayesian Expressivism." *Proceedings of the Aristotelian Society* CXII, Part 2: 123–60. https://doi.org/10.1111/j.1467-9264.2012.00329.x.

10

WEAK AND STRONG NECESSITY MODALS:

On Linguistic Means of Expressing "A Primitive Concept OUGHT"[1]

Alex Silk

'Ought' and 'Must'—they are contemptible auxiliaries.

George Eliot[2]

1. Introduction

A notion of 'ought' is central in many areas of philosophical discourse. "A primitive ought," Gibbard tells us, "is the basic conceptual atom that gives normative concepts their special character" (2006, 738). Yet, historically speaking, comparatively little attention has been paid among philosophers to the distinctive features of the meaning and use of 'ought'. 'Ought' is often treated as relevantly equivalent to a range of expressions of obligation and necessity. The eponym of Gibbard's Professorial Chair, Richard Brandt, observed as much half a century ago: "Philosophers often use the following expressions as approximate equivalents: 'It is X's duty to do *A*'; 'It is obligatory for X to do *A*'; 'It would be wrong for X not to do *A*'; and 'X ought to do *A*'" (1964, 374). Here is Åqvist:

1 Thanks especially to Eric Swanson for extensive discussion and comments, and to Billy Dunaway for feedback on reining in the bloated beast that was the penultimate draft. Thanks also to Matthew Chrisman, Brendan Dill, Jan Dowell, Allan Gibbard, Irene Heim, Ezra Keshet, Dan Lassiter, David Plunkett, Paul Portner, Bernhard Salow, Robert Shanklin, Bob Stalnaker, and audiences at SALT 22, MIT, Northwestern, and USC. Preliminary versions of the paper were published in Silk 2012, 2013, ch. 2. My 2016b draws on material from earlier drafts. Portions of §7 are drawn from Silk 2015b.

2 Mary Garth, in *Middlemarch*, Bk. 2, Ch. 14. Shamelessly modified from the original.

> [D]eontic logic . . . is the logical study of the normative use of language and . . . its subject matter is a variety of normative concepts, notably those of *obligation* (prescription), *prohibition* (forbiddance), *permission* and *commitment*. The first one among these concepts is often expressed by such words as 'shall', 'ought' and 'must', the second by 'shall not', 'ought not' and 'must not'.
>
> (Åqvist 2002, 148)

There has been extensive work in descriptive linguistics on discourse differences among Åqvist's "normative-concept-expressing" words. For instance, it's common to distinguish categories of so-called "weak" necessity modals such as 'ought' ('should', 'be supposed to') and "strong" necessity modals such as 'must' ('have to', '(have) got to', 'be required to').[3] Holding the reading of the modals fixed, (1a) is consistent in a way that (1b) is not. 'Ought ϕ' can be followed by 'Must ϕ', but not vice versa, as reflected in (2).

(1) a. I should help the poor, but I don't have to.
    b. #I must help the poor, but it's not as if I should.

(2) a. I ought to help the poor. In fact, I must.
    b. I must help the poor. #In fact, I ought to.

There are also conversational differences. Informally:

> To say that one ought to take a certain option is merely to provide a nudge in that direction. Its typical uses are to offer guidance, a word to the wise . . . , to recommend, advise . . . In contrast, to say that one must take a certain option is to be quite forceful. Its typical uses are to command, decree, enact, exhort, entreat, require, regulate, legislate, delegate, or warn.
>
> (McNamara 1990, 156)

Gibbard hedges his bets on what natural language expression best approximates his favored normative notion:

> Ewing's point is not that the English word 'ought' exactly captures the notion he has in mind. The word often suggests merely the weight of one set of considerations among others . . . The word 'must' might be better for his purposes and mine [?!?!], but it has its own flaws: it suggests greater urgency than Ewing's ought would have when factors in a decision nearly balance out.
>
> (Gibbard 2012, 14–15; more on '?!?!' later)

---

3 See, e.g., Sloman 1970, Leech 1971, Horn 1972, Wertheimer 1972, Harman 1975, Lyons 1977, Woisetschlaeger 1977, Williams 1981, Coates 1983, McNamara 1990, Palmer 1990, 2001, Bybee et al. 1994, Myhill 1995, Myhill and Smith 1995, Huddleston and Pullum 2002, von Fintel and Iatridou 2008, Rubinstein 2012; additional references to theoretical accounts will follow in due course. I focus on modal verbs; see Van Linden 2012 on weak/strong modal adjectives.

An interesting question for a paper—not this one (cf. Silk 2015b[4])—is to what extent insensitivity to linguistic differences among 'ought', 'should', 'must', 'have to', etc. may be a source of philosophical malaise. On the flip side, perhaps investigating the meaning and discourse function of such words can clarify the contours of Gibbard's basic normative "conceptual atom" and improve philosophical theorizing. In this paper I wish to set the stage for such a project by attending to the matters narrowly linguistic: what 'ought'—the eponymous expression of Gibbard's "primitive concept OUGHT" (2012, 204)—itself means. I focus on the distinction in "strength" between 'ought' (and its weak-necessity-modal kin) and strong necessity modals such as 'must'.

Although there has been growing interest in the semantics of 'ought', accounts are often developed in ways that bracket differences among necessity modals. Formal accounts of the weak/strong necessity modal distinction are typically developed piecemeal with an eye toward a narrow range of data. Adjudicating among theories can be difficult, if not premature. The aim of this paper is a more comprehensive theoretical investigation into weak and strong necessity modals. Building on previous work (Silk 2012, 2013), I develop an account of the meaning of 'ought' and the distinction between weak and strong necessity modals. The account systematizes a wide range of semantic and pragmatic phenomena: it generalizes across flavors of modality; it elucidates a special role that weak necessity modals play in discourse and planning; it captures contrasting logical, expressive, and illocutionary properties of weak and strong necessity modals; and it sheds light on how a notion of 'ought' is expressed in other languages. These phenomena have resisted systematic explanation.

Roadmap: §2 presents core data on the effects of standing contextual assumptions on the relative felicity of weak vs. strong necessity modals. The §2-examples highlight what I regard as the fundamental difference between the class of weak necessity modals and the class of strong necessity modals.

§3 presents the basic account of the weak/strong necessity modal distinction. No innovations are introduced in the semantics and pragmatics of strong necessity modals; uses of 'Must ϕ' predicate the (deontic, epistemic) necessity of the prejacent ϕ of the actual world.[5] The apparent "weakness" of weak necessity modals derives from their bracketing whether the necessity of the prejacent is verified in the actual world. 'Ought ϕ' can be accepted without accepting that ϕ is necessary (deontically, epistemically, etc.). Weak necessity modals afford a means of entertaining and planning for hypothetical extensions of the context in which certain considerations (norms, values, etc.) apply, without needing to commit that the considerations aren't defeated.

§4 examines how weak necessity is expressed crosslinguistically to motivate several formal implementations of the informal §3-account. For reasons that will become clear, I call the family of analyses a *modal-past approach* to 'ought' and the weak/strong necessity

4 Here and throughout: For shame.

5 I treat 'ϕ', 'ψ', etc. as schematic letters to be replaced with declarative sentences. For convenience I sometimes refer to the possible-worlds proposition expressed by 'ϕ' by dropping the single quotes, e.g. using 'ϕ is necessary' for '$[\![\phi]\!]^c$ is necessary', where $[\![\phi]\!]^c = \{w: [\![\phi]\!]^{c,w} = 1\}$; I slide between equivalent set-/function-talk.

modal distinction. I argue that the proposed treatment of the crosslinguistic data improves on the treatment by von Fintel and Iatridou (2008). The account gives precise expression to the informal idea that 'ought' is weaker than 'must', and captures ways in which 'ought', unlike 'must', patterns with past-marked modal forms.

§5 applies the formal semantics from §4 to several puzzles of entailingness and performativity with 'ought' and 'must'. Examining these puzzles highlights a second dimension along which modals differ, regarding their tendencies to be used in (what I call) an "endorsing" vs. "nonendorsing" way.

§6 recaps distinctive features of the account and contrasts it with several prominent alternatives, in particular the "collective commitment" analysis developed in Rubinstein 2012.

§7 concludes and raises directions for future research. Potential implications of the linguistic work on modals for broader philosophical theorizing are briefly considered.

For familiarity I follow the literature in labeling modals such as 'ought' and 'should' as "weak necessity" modals, and labeling modals such as 'must', 'have to', '(have) got to' as "strong necessity" modals. The terminology of "weak necessity" and "strong necessity" shouldn't mislead. I am not assuming that uses of the modals in the 'ought'-family invariably convey a weaker felt conversational force, that the modals express different "kinds" of necessity, or even that they comprise a scale of logical/quantificational strength. A claim such as that strong necessity modals truth-conditionally entail weak necessity modals, which truth-conditionally entail possibility modals, constitutes a substantive empirical hypothesis on my terminology. We will see reasons for questioning each of the above claims. (Hereafter I typically use 'ought' as my representative of the 'ought'/'should'/etc. family and 'must' as my representative of the 'must'/'have to'/etc. family.)

2. 'Ought' and 'Must' in Context

Descriptive and theoretical research on modals highlights various conversational differences among necessity modals. This section focuses on one such difference concerning the effects of contextual assumptions on the relative felicity of weak vs. strong necessity modals.[6]

6 See Woisetschlaeger 1977, ch. 5 and McNamara 1990, ch. 3 for prescient early discussion of contextual differences between 'ought' and 'must'. For extensive discussions in descriptive linguistics, see the references in nn. 3, 39, 41. See Rubinstein 2012 and Silk 2012 for recent theoretical emphasis. Rubinstein doesn't consider epistemic examples; I examine her alternative take on the data in §6. For reasons discussed in §§5–6, it's important not to substitute other strong necessity modals (e.g. 'have to') for 'must' in examples unless indicated otherwise; speakers who find 'should' more natural than 'ought' may substitute 'should' for 'ought' throughout. Judgments concerning some of the examples may be vague for some speakers and may vary given subtle changes in context. This is part of what needs to be explained. The positive account developed in the paper will crystalize the informal reactions described in this section. I use '?' to indicate that using the item is dispreferred; '?' marks a weaker infelicity than '#'.

A central purpose of conversation is to share and coordinate our expectations, values, and plans. Sometimes we assert propositions outright. We commit to settling on their truth for the remainder of the conversation. But sometimes we don't wish to impose such a strong restriction on the future course of the conversation. We may want to propose that someone is obligated to do something but be unsure about whether there might be competing norms that could outweigh or cancel her obligation. Or we may want to proceed as if some proposition is true while remaining open to the possibility that our apparent evidence for it is misleading. I suggest that the role of weak necessity modals is to afford a means of making such proposals and expressing such states of mind.

Start with an epistemic case. Suppose we're working on an art project and I ask you where the colored pencils are. Normally you put them in the drawer with the crayons but sometimes you accidentally put them on the shelf. In this scenario it's more appropriate for you to use 'ought' in responding to my question:

(3) *Me*: Do you know where the colored pencils are?
You: They ought to (/should/?must/?have to) be in the drawer with the crayons.

Suppose, alternatively, that we're looking for the colored pencils together, and you saw something that leads you to conclude that they are in the drawer. Perhaps you noticed that they weren't on the shelf, and this is the only other place you think they could be. In this scenario it's more natural for you to use 'must':

(4) *Me*: Do you know where the colored pencils are?
You: They must (/have to/?ought to/?should) be in the drawer with the crayons.

Its following from our evidence that the colored pencils are in the drawer depends on today not being one of the days when you accidentally put them on the shelf. Using 'must' is preferred if, and only if, you know that conditions are normal in this way. What is illuminating is that you can use 'ought' even if you aren't in a position to judge that they are. Accepting your 'ought'-claim doesn't require us to assume that your evidence is undefeated.

Consider a deontic case (n. 6). Suppose I am considering whether to fight in the Resistance or take care of my ailing mother. I mention that the value of family, which supports my helping my mother, is important, and you agree. But the issue is admittedly complex, and we haven't settled whether there might be more important competing values. Sensitive to this, you may find it more appropriate to express your advice that I help my mother by using 'ought' than by using 'must':

(5) *Me*: Family is very important.
You: I agree. You ought to/should (/?must/?have to) tend to your mother.

But if we settle that family is of primary importance, it can become more natural to use 'must' and for us to accept that I have to help my mother:

(6) *Me*: Family is most important—more important than country.
You: I agree. You must/have to (/?ought to/?should) tend to your mother.

My having an obligation to help my mother depends on the value of family being more important[7] in my situation than any competing value. Parallel to the epistemic case, what is illuminating is that you can felicitously use 'ought' to express your advice that I help my mother without assuming that this precondition for my having a genuine obligation is satisfied. Accepting your 'ought'-claim needn't require us to presuppose that the value of family is more important than other potentially competing values.

Cases such as (3)–(6) highlight what I regard as the fundamental difference between the class of weak necessity modals and the class of strong necessity modals. It's common to gloss epistemic notions of necessity as concerning what follows from a body of evidence (knowledge, information), and deontic notions of necessity as concerning what is obligatory.[8] Yet we can accept your epistemic 'ought'-claim in (3) without settling that conditions are relevantly normal and thus without settling that our evidence implies that the colored pencils are in the drawer; and we can accept your deontic 'ought'-claim in (5) without settling that family is the most important relevant value and thus without settling that I have a genuine obligation to help my mother. Accepting 'Ought ϕ' needn't commit one to accepting that ϕ is necessary (epistemically, deontically, etc.).

Whether 'ought' or 'must' is preferred depends on context in the sense of depending on whether certain preconditions for the prejacent to be necessary (in the above sense) are accepted. In (3)–(4), how you express your attitude toward the proposition that the colored pencils are in the drawer depends on your views about the (in)defeasibility of the relevant evidence; in (5)–(6), how you express your advice that I help my mother depends on the status in the context of the value of family vis-à-vis other potentially relevant values. This

7 Or at least not less important; I will bracket complications from incomparabilities and irresolvable dilemmas. For discussion of dilemmas and the 'ought'/'must' distinction, see Swanson 2011, Silk 2012, 2015b, and references therein.

8 See, e.g., Lyons 1977, Coates 1983, Palmer 1990, 2001, Sweetser 1990, Bybee et al. 1994, van der Auwera and Plungian 1998, Nuyts 2001, Huddleston and Pullum 2002. In saying that the uses of 'ought' and 'must' have the same type of reading, I am not assuming that the modals have the same interpretation or stand in entailment relations (§1). In calling uses "epistemic" I don't assume that they are factive/entailing, or even that they convey the same kind of doxastic attitude toward the prejacent (contrast Yalcin 2016). What is important about the uses in (3)–(4) is that they both directly address the question of where the colored pencils are and convey an assessment of the truth or likelihood of the prejacent given a body of information. In calling uses "deontic" I don't assume that they are performative or have the same directive force. What is important about the uses in (5)–(6) is that they both directly address the practical question of what I am to do.

effect of contextual assumptions on the relative felicity of 'ought' and 'must' has been generally underappreciated in theoretical accounts (see n. 6 for notable exceptions).[9]

Two clarificatory remarks: First, I said that 'ought' is preferred in contexts such as in (5) where it isn't settled that the precondition for me to have an obligation is satisfied; yet it is worth observing that 'must' may be appropriate in certain contexts. If you can be presumed a normative authority on the issue and use 'must', I may accommodate by accepting that the value of family takes precedence. This isn't an isolated phenomenon. Suppose that Alice, a young teen, is considering with her mother, Martha, whether to take the A-train or C-train to a concert. The A is quicker, but the C is safer. Martha regards Alice's safety as paramount. Although the primacy of safety isn't common ground, Martha can felicitously say:

(7) You must take the C-train, not the A-train.

The (teleological) necessity of Alice's taking the C-train depends on the goal of traveling safely taking priority over the goal of traveling quickly. Given Martha's authority in the context, she expects Alice to accommodate her assumption that this condition is satisfied. Such contexts notwithstanding—contexts in which the speaker doesn't have or doesn't wish to exercise the relevant authority—'ought' will be preferred.

Second, saying that accepting 'Ought ϕ' doesn't conventionally commit one to accepting that ϕ is necessary doesn't amount to the trivial claim that accepting 'Ought ϕ' needn't commit one to accepting 'Must ϕ'. One might analyze the examples by positing concepts of distinctive kinds of necessity, and explain accepting 'Ought ϕ' as accepting that the weaker kind of necessity holds of ϕ. For instance, one might posit and formalize a concept of weak epistemic necessity such that accepting that it's a weak epistemic necessity that *d* the colored pencils are in the drawer doesn't require accepting that today is relevantly normal and the evidence implies *d*; and one might posit and formalize a concept of weak deontic necessity (weak obligation) such that accepting that I have a weak obligation to help my mother doesn't require accepting that the value of family isn't defeated. Yet such a move isn't forced upon us. An alternative is to stick with the single familiar notions of necessity—e.g., understanding epistemic necessity as following from a body of evidence, and deontic necessity as being obligatory and following from a body of norms (n. 8)—and try saying that 'Ought ϕ' can be accepted without accepting that ϕ is necessary, period. The following sections investigate the prospects for this latter approach. (More on comparisons with the former approach in due course.)

9 The semantics in Finlay 2009, 2010, Lassiter 2011, Swanson 2011 have no obvious mechanism for capturing this effect of context on uses of 'ought' vs. 'must'. von Fintel and Iatridou (2008, 139–40) mention in passing a possibly relevant role for context, but the issue isn't investigated. See Rubinstein 2012, §2.2 for extensive critical discussion of previous comparative approaches and domain restriction approaches to weak necessity modals. (More on von Fintel and Iatridou's and Rubinstein's accounts in §§4, 6.)

3. The Analysis: Preliminary

The core of the approach to the weak/strong necessity modal distinction developed in the following sections is as follows. There is nothing specially "strong" about the necessity expressed by strong necessity modals. Strong necessity modals are given their usual semantics and pragmatics. 'Must ϕ' is true iff ϕ is necessary (in the relevant sense, i.e. epistemically, deontically, etc.; §2),[10] and uses of 'Must ϕ' predicate the necessity of ϕ of the actual world—just as 'May ϕ' is true iff ϕ is possible, and uses of 'May ϕ' predicate the possibility of ϕ of the actual world. The apparent weakness of weak necessity modals, we can try saying, derives from their bracketing the assumption that the necessity of ϕ need be verified in the actual world. Accepting 'Ought ϕ' needn't commit one to accepting that the actual circumstances verify the necessity of ϕ. Weak necessity modals afford a means of coordinating on the implications of our values, norms, etc. without having to settle precisely how they weigh against one another in particular circumstances, and while remaining open to new evidence about how they apply. This section begins developing these ideas within a standard premise-semantic framework for modals. The next section examines how the account may be implemented more precisely in the formal semantics and pragmatics.

As is common, I follow Kratzer (1977, 1981, 1991) in treating modals as semantically associated with a parameter determining a set of premises (propositions). Since modals can occur in intensional contexts, premise sets are indexed to a world of evaluation. What context supplies that determines the reading of a modal is the family of world-indexed premise sets $(P_w)_{w \in W}$, or *premise frame*: a function P from worlds w to premise sets $P(w)$.[11]

It is nontrivial how the considerations which seem intuitively relevant in interpreting modals are to be represented in the formal objects in the compositional semantics. First, the expressive, practical, and discourse-managing roles of modals such as 'ought' have been central in expressivist theories such as Gibbard's. However, the account of the weak/strong necessity modal distinction in what follows will be neutral on matters of expressivism (contextualism, relativism, invariantism)—e.g., on what type of psychological state of mind is conventionally expressed by certain uses of 'ought'/'must' sentences; whether notions of content or truth play a fundamental explanatory role in explaining semantic properties of sentences

10 I will often omit this parenthetical, but it should be understood.

11 See also van Fraassen 1973, Lewis 1973, 1981, Veltman 1976. Kratzer calls premise frames 'conversational backgrounds'. Kratzer's (1981, 1991) semantics uses two premise sets: a "modal base" $F(w)$ that represents a set of relevant background facts in w, and a (possibly inconsistent) "ordering source" $G(w)$ that represents the content of some ideal in w. Making the limit assumption (Lewis 1973, 19–20), Kratzer's semantics treats 'Must ϕ' as true at w iff ϕ follows from every maximally consistent subset of $F(w) \cup G(w)$ that includes $F(w)$ (equivalently (Lewis 1981), iff every $\lesssim_{G(w)}$-minimal world in $\bigcap F(w)$ is a ϕ-world).) These complications won't be relevant here; I simplify by treating modals as evaluated with respect to a single finite, consistent premise set. Nothing will turn on views about the limit assumption, or debates about so-called "weak vs. strong" semantics for epistemic 'must' (whether epistemic 'must' takes a nonempty ordering source).

or the dynamics of discourse; and whether particular premise frames figure in the compositional semantic value, and, if so, whether they are supplied by the context of utterance or a posited context of assessment. I will speak simply of "context," and I relativize parameters such as premise frames simply to worlds. The implementations to follow may be adapted along alternative contextualist/relativist/expressivist/invariantist lines.[12] (I revisit motivations for expressivism in §7. More on the practical roles of 'ought'/'must' below and in §5.)

Premise frames afford a natural way of encoding the contents of bodies of norms, preferences, etc.[13] Call a conditional norm, preference, expectation, etc. a *consideration*. A contextually supplied premise frame P encodes the content of a body of considerations. The premises in a premise set $P(w)$ represent what follows from a body of considerations given the relevant circumstances in w. For instance, suppose you want to go for a run given that it's sunny, you didn't just eat a burrito, and so on. The content of your preference can be encoded in a premise frame that assigns a premise set including the proposition that you go for a run to worlds where it's sunny, etc. Similarly the normative import of a value of charity might be encoded in a deontic premise frame that assigns a premise set including the proposition d that you donate to charity to worlds where you have means of supporting your family, etc.—as with P_d in (8), for relevant worlds characterized with respect to whether or not you have a Job, there are Reputable charitable organizations, and there is a local Soup kitchen, and where u is the proposition that you undermine local aid organizations, and h is the proposition that you help at a local soup kitchen.

(8) $P_d(\text{JRS}) = \{d, h, \ldots\}$
$P_d(\text{J}\overline{\text{R}}\text{S}) = \{u, h, \ldots\}$
$P_d(\text{JR}\overline{\text{S}}) = \{d, \ldots\}$
$P_d(\overline{\text{J}}\text{RS}) = \{h, \ldots\}$

The normative importance of helping those in need is reflected in P_d's assigning a premise set that includes h to worlds where there is a local soup kitchen. The normative importance of charitable giving is reflected in P_d's assigning a premise set that includes d to certain worlds where available aid organizations are reputable. However, the latter norm isn't unconditional; it applies only in worlds where you can support your family, and where the organizations will put the donations to good use. The relative importance of supporting your family over helping others is reflected in P_d's assigning a premise set that doesn't include d to worlds such as $\overline{\text{J}}\text{RS}$ where the aid organizations are trustworthy but you don't have a job.

12 Cf., e.g., Gibbard 1990, 2003, Stephenson 2007, Dreier 2009, Yalcin 2012, MacFarlane 2014, Silk 2015a, 2016a, Swanson 2016a.

13 See Silk 2015a, 2016a, §5.6, 2017 for additional applications of the following approach to interpreting the formal premise-/ordering-semantic apparatus.

With this way of understanding premise frames at hand, let's return to the semantics and diagnoses of the §2-examples. I give strong necessity modals their usual semantics of necessity. 'Must ϕ' is true at w, given a contextually supplied premise frame P, iff the prejacent proposition ϕ follows from $P(w)$, as in (9) (nn. 5, 11). The truth of 'Must ϕ' depends on the value of P at the evaluation world. Asserting 'Must ϕ' commits one to accepting that ϕ follows from what the relevant considerations enjoin given the facts, $P(w)$.[14]

(9) $[\![\text{Must } \phi]\!]^{c,w} = 1$ iff $\bigcap P_c(w) \subseteq [\![\phi]\!]^c$ (*preliminary*)

(10) A sentence S is *accepted* in c iff for every $w \in c$, $[\![S]\!]^{c,w} = 1$

What distinguishes weak necessity modals, I have said, is that they bracket whether the necessity claim is verified in the actual world. We can adopt the following constraint on a semantics and pragmatics for 'ought', for some relevant body of considerations P:

(11) It's not the case that: 'Ought ϕ' is accepted in c only if for every $w \in c$, $\bigcap P(w) \subseteq [\![\phi]\!]^c$

In a manner to be made precise, uses of 'Ought ϕ' present the possibility that ϕ follows from the relevant considerations given certain circumstances, but without committing that such circumstances obtain or that the considerations actually apply. (We will examine what accepting 'Ought ϕ' does commit one to shortly.)

Let's apply the preliminary analyses thus far to our §2-examples. Recall (3)–(4), reproduced in (12)–(13).

(12) *Me*: Where are the colored pencils?
You: They ought to be in the drawer with the crayons.

(13) *Me*: Where are the colored pencils?
You: They must be in the drawer with the crayons.

As with norms and preferences, one's expectations given a body of evidence can be conditional—e.g., conditional on things being normal in the relevant respects. Let w_N be a world in which your routine proceeds as normal, and let $w_{\overline{N}}$ be a world where something abnormal happens to disrupt your routine.[15] The conditional expectations concerning the colored pencils' location can be encoded in a premise frame P_e which (inter alia) assigns

14 Subscripts on premise frames are used simply to indicate the intended contextually determined assignment. The familiar definition of acceptance in (10) blurs the distinction between contexts and the context sets they determine. The context set is the set of live possibilities, the set of worlds compatible with what is accepted for purposes of conversation (Stalnaker 1978).

15 As previously, w_N and $w_{\overline{N}}$ may be understood as representatives of relevant equivalence classes of worlds.

to w_N a premise set $P_e(w_N)$ including the proposition d that the colored pencils are in the drawer, and which assigns to $w_{\bar{N}}$ a premise set $P_e(w_{\bar{N}})$ including $\neg d$. Given P_e, d is an epistemic necessity at w_N and not at $w_{\bar{N}}$. So in order for d to be accepted as epistemically necessary, per (9)–(10), the context set must be restricted to worlds like w_N—worlds where you didn't get distracted before putting the colored pencils away, no one played a trick on us and moved them, etc. If I'm unsure whether you are in a position to assume that nothing unusual led you to place the colored pencils somewhere else, I may challenge your assumption and raise a possibility that is incompatible with the epistemic necessity of d, as in (14).

(14) *You*: The colored pencils must be in the drawer with the crayons.
Me: Really? I see that they aren't on the shelf. But don't you sometimes accidentally put them in the cabinet with the glue sticks?
You: No, I never put them there. (/Oh, I forgot about that.)

But if you use 'ought', you aren't committing to conditions being normal, as reflected in (15); hence my mentioning such alternative possibilities may be beside the point, as in (16).

(15) *You*: The colored pencils ought to be in the drawer with the crayons.
Me: I checked and they aren't there.
You: Oh, then I'm not sure where they are. I would have expected them to be there.

(16) *You*: The colored pencils ought to be in the drawer with the crayons.
Me: #Really? I see that they aren't on the shelf. But don't you sometimes accidentally put them in the cabinet with the glue sticks?
You: I know; that's why I said *ought*!

Epistemic 'Ought ϕ' can be accepted even if it isn't settled that certain conditions relevant to the epistemic necessity of ϕ are satisfied.

Turn to our deontic modal examples in (5)–(6). My having an obligation to take care of my mother depends on the value of family being more important in my situation than other potentially competing values. Hence in order for the proposition m that I tend to my mother to follow from $P(w)$—what the normative considerations enjoin given the circumstances—it must be the case that the value of family takes precedence in my situation in w. In (5), unlike (6), after my assertion is accepted it still isn't settled whether this condition is satisfied. So, were you to use 'must' you would imply that you are foreclosing certain possibilities that I have left open. Unless you are in a position to do so (cf. (7)), your using 'must' is dispreferred. By accepting your claim with 'ought', we can provisionally proceed as if my helping my mother is required without needing to settle that the value of family is more important than other competing values we accept or may come to accept.

These examples highlight a critical role for weak necessity modals in discourse and deliberation. Take the deontic case. There are typically a range of interests, values, norms potentially relevant for determining what to do. How the relevant factors, whatever they are, interact is often highly complex. (One needs only a foray into deontic logic or normative ethics to convince oneself of this.) There may be uncertainty about the facts that would determine which considerations apply. For instance, in (8) one might not know the details about someone's financial or family situation. Other empirical factors—how donations are used, what the short- and long-term impacts are on those in need, etc.—can be even more difficult to assess. Hence one might not be in a position to commit to being in a world like JRS/JR$\overline{\text{S}}$ where norms of charity aren't outweighed or defeated. Scenarios such as those in (5)–(6) compound such challenges for normative and empirical evaluation. Using deontic 'Must ϕ' may thus be inapt. One might not be in a position to commit to being in a world where ϕ follows from what the relevant norms enjoin given the facts. Deontic 'ought' affords a means of guiding our deliberations and plans—indeed our *conditional* plans, as emphasized throughout Gibbard's work on the psychology of normative judgment—while remaining open to new evidence about what values are at stake and how they interact with one another and the relevant facts.

I have proposed that what makes weak necessity modals "weak" is that they bracket whether the necessity of the prejacent is verified in the actual world. One can accept 'Ought ϕ' without presupposing that ϕ follows from what the relevant considerations P enjoin given the facts—e.g., without committing that all preconditions for ϕ to be a genuine obligation are satisfied, that one's evidence for ϕ isn't misleading, and so on. This feature of weak necessity modals isn't the only dimension along which necessity modals differ (more on which in §§5–6). However, I claim that it does distinguish the *class* of weak necessity modals from the *class* of strong necessity modals. Previous accounts of weak necessity modals have often been developed by considering a limited range of modal flavors in a limited range of contexts; extensions to other readings, to the extent that they are discussed at all, are often strained (e.g., Copley 2006, Swanson 2011, Rubinstein 2012, Charlow 2013, Ridge 2014, Portner and Rubinstein 2016, Yalcin 2016). The account in this paper generalizes across flavors of modality, and it captures a precise sense in which the relative felicity of 'ought' and 'must' depends on standing assumptions. Weak necessity modals afford a means of coordinating on the implications of our values, expectations, etc. without needing to settle precisely how they apply and weigh against one another in particular circumstances.

The preliminary account thus far raises many questions. Yet even at the present level of abstraction, we can see that the approach to the weak/strong necessity modal distinction in this paper differs crucially from other main approaches in the literature, such as probabilistic and comparative possibility approaches (Finlay 2009, 2010, 2014, Lassiter 2011) and domain restriction approaches (Copley 2006, von Fintel and Iatridou 2008, Swanson 2011, Rubinstein 2012, Charlow 2013). For instance, domain restriction accounts maintain that accepting 'Ought ϕ' requires accepting that ϕ is a necessity, and that the truth of 'Ought ϕ' at w requires that ϕ is a necessity at w; what distinguishes 'ought' from 'must' is the logical strength of the

necessity (roughly put, implication of ϕ by a superset of premises; equivalently, truth of ϕ throughout a subdomain of worlds). Weak necessity modals are treated as expressing a logically weaker *kind* of necessity. The present approach rejects these claims (§2).

4. Weak Necessity and the Modal Past

Our project is to develop an account to the weak/strong necessity modal distinction that systematizes a broader range of linguistic phenomena (§1). Taking on this more demanding goal requires sustained investigation into diverse domains. This section examines how notions of "ought" are expressed crosslinguistically in order to motivate several ways of formally implementing the proposed approach to the weak/strong necessity modal distinction and the semantics of 'ought'. §5 shows how the account helps explain various seemingly unrelated semantic and pragmatic properties of 'ought' and 'must'.

4.1. Data

Past forms of modals—in English, 'would' for 'will', 'could' for 'can', 'might' for 'may'—are often used not to indicate past time reference, but to express tentativeness or politeness and weaken the apparent force, as in (17)–(19). These forms are also the forms that appear in the consequents of subjunctive conditionals, as in (20). Palmer (2001) dubs such uses of past tense the "modal past."[16]

(17) a. I will add one point to this discussion.
b. I would add one point to this discussion.

(18) a. Alice will/may/can't be at home now.
b. Alice would/might/couldn't be at home now.

(19) a. May/Can I comment on your proposal?
b. Might/Could I comment on your proposal?

(20) If you took the flight tomorrow, you would/could/might get there in time.

---

16 See also James 1982, Coates 1983, Fleischman 1989, Bybee 1995, Iatridou 2000, Huddleston and Pullum 2002. Terminology varies among authors. Weakness interpretations of past tense aren't limited to uses of modal verbs, as reflected in the present time readings of (i)–(iii).

(i) I wanted to ask you a question. (Bybee 1995, ex. 21)

(ii) I thought/was thinking about asking you to dinner. (Fleischman 1989, 8)

(iii) *A*: How old is John?
B: He'd be about sixty. (cf. Huddleston and Pullum 2002, 200–201)

I don't assume that all modal-past forms necessarily convey weakness/tentativeness relative to their nonpast counterparts (cf. epistemic 'might' in contemporary English; Coates 1983, Palmer 1990, Huddleston and Pullum 2002, Collins 2007). My descriptive use of Palmer's label makes no theoretical assumptions about how the modal/weakness/tentativeness interpretations arise or are related to temporal interpretations of past morphology (n. 21).

Strikingly, 'ought' patterns with the past-marked modal forms. First, 'ought' weakens the apparent force of 'must'—hence the common label "weak necessity modal." Second, 'ought', unlike 'must', can appear in subjunctive conditionals:

(21) a. If Alice came to the party tomorrow, Bert ought to leave.
b. #If Alice came to the party tomorrow, Bert must(ed) leave.

Third, 'ought' is nonentailing. For simple clauses 'ϕ', 'Ought ϕ' contrasts with 'Must ϕ' in being compatible with '$\neg\phi$':

(22) I could give to Oxfam, but I won't.

(23) a. Alice ought to be here by now, but she isn't.
b. #Alice must be here by now, but she isn't.

Indeed, when used with the perfect, 'ought' implicates the negation of the prejacent.

(24) We could have given to Oxfam. (*Implicates*: we didn't)

(25) a. We ought to have given to Oxfam. (*Implicates*: we didn't)
b. #We must have given to Oxfam (but we didn't).

'Must' cannot even receive a deontic reading when used with past time reference. 'Ought', unlike 'must', can be used to communicate that an obligation held in the past. That 'ought' can scope under the perfect in (25a) is a fourth respect in which 'ought' patterns with past-marked modal forms (Condoravdi 2002).

In sum, although 'must' doesn't have a past form, 'ought', we can try saying, functions notionally as its modal past (cf. Palmer 1990, 2001). This is surprising. But it becomes less surprising when we examine other languages. Let's use 'OUGHT' for the notion which in English is expressed with weak necessity modals such as 'ought' ('should', etc.), and let's use 'MUST' for the notion which in English is expressed with strong necessity modals such as 'must' ('have to', etc.). As emphasized in von Fintel and Iatridou's (2008) seminal discussion of weak necessity modals, it's crosslinguistically common to mark the semantic distinction between OUGHT and MUST morphologically rather than lexically (see also Palmer 2001, McGregor and Wagner 2006, Van Linden and Verstraete 2008, Matthewson 2010). A notion of OUGHT is often expressed not by using a different word—like 'ought' in English—but by using the modal-past form of a strong necessity modal, i.e. the form of a strong necessity modal that is used in counterfactuals.[17] von Fintel and Iatridou

17 As von Fintel and Iatridou note (2008, 126n.22), 'ought' fits the crosslinguistic pattern historically; it was formerly the past subjunctive of the verb 'owe'. Modal-past forms in other languages may be derived from various elements, not simply past tense (e.g., Iatridou 2000).

(2008) approach the crosslinguistic data in a domain restriction account of weak necessity modals. They treat weak necessity modals as quantifying over "the best of the best" worlds—the relevant $P(w)$-compatible worlds that are compatible with an additional premise set representing a secondary ideal (in Kratzer's terminology, a secondary ordering source).[18] In passing, von Fintel and Iatridou speculate that "the counterfactual marking is co-opted here in a somewhat meta-linguistic kind of way: 'if we were in a context in which the secondary ordering source was promoted [to primary status], then it would be a strong necessity that . . .'" (2008, 139). The tentativeness associated with counterfactual marking is attributed to the fact that the premises in the secondary premise set needn't apply: "The choice of whether to really promote the secondary ordering source is left open" (2008, 139).

I find these suggestions about the crosslinguistic data unsatisfying (see Rubinstein 2012 for further critiques). First, von Fintel and Iatridou briefly suggest treating the secondary ordering source for epistemic readings as representing what is normally the case, and the secondary ordering source for deontic readings as representing "less coercive sets of rules and principles" (2008, 119); yet no general story is given about what primary vs. secondary ordering sources represent per se, or how an ordering source is determined as primary or secondary across contexts. Absent an independent understanding of what makes it the case about a speaker that she is counterfactually promoting a secondary ordering source, the proposed story about the role of the counterfactual marking seems ad hoc. A worry is that the explanation of the tentativeness associated with counterfactual morphology redescribes what needs to be explained. The tentativeness is "explained" by introducing a parameter representing "shakier assumptions" (2008, 119n.9) that may not apply—uncharitably put, a parameter representing considerations one may only be tentatively committed to. Finally, it's unclear how von Fintel and Iatridou's explanation would generalize to tentativeness effects of counterfactual morphology on other lexical items—e.g., possibility modals like 'might', or desire verbs like 'wish' in other languages (as von Fintel and Iatridou note, many languages express a notion of WISH via counterfactual morphology on the word for 'want'; see also n. 16).[19] For instance, interpreting modal-past forms of possibility modals with respect

18 More precisely (see n. 11): For a set of worlds W and premise set S, let the S-minimal worlds in W be the worlds $u \in W$ such that no world $v \in W$ satisfies a proper superset of propositions $p \in S$. (For finite consistent premise sets the set of S-minimal worlds is $W \cap \bigcap S$.) Let s be a set of relevant worlds (e.g., as determined by a Kratzerian modal base), $G_1(w)$ be a premise set representing some primary ideal, and $G_2(w)$ be a premise set representing some secondary ideal. On von Fintel and Iatridou's semantics, whereas 'must' universally quantifies over the $G_1(w)$-minimal s-worlds, 'ought' universally quantifies over a subset of these worlds: the $G_2(w)$-minimal worlds among the $G_1(w)$-minimal s-worlds. 'Ought' quantifies over the $G_2(w)$-minimal worlds that 'must' quantifies over. See Sloman 1970, Williams 1981, McNamara 1990 for informal precedents. I return to general worries for domain restriction analyses below.

19 Interestingly, in various Italian dialects OUGHT can be expressed via counterfactual morphology on volitional verbs such as 'want', as in (i) below. Expressions of MUST with the indicative form are also possible, as in (ii). (Thanks to Federico Faroldi for bringing this to my attention.)

to a secondary ordering source incorrectly predicts a strengthening effect. The tentative use of 'might'/'could' in (19) cannot be derived by counterfactually "promoting" a secondary ordering source and then evaluating the possibility of the prejacent.[20]

4.2. Implementations. A Modal–Past Approach

This subsection examines how independent work on the semantics/pragmatics of counterfactual marking may be incorporated into our analyses of 'ought' and 'must'. I want to be clear that I am not assuming that lexicalized weak necessity modals are decomposed into a strong necessity modal and counterfactual features (schematically: *STRONG+CF*). The crosslinguistic data may provide insight into the semantics of lexicalized expressions of OUGHT; but we should be careful not to read off a semantics for 'ought' from a semantics for 'must' and counterfactual morphology. 'Ought' doesn't mean 'would have to'.

It's generally agreed that counterfactual marking signals that the worlds being talked about ("topic worlds") needn't be candidates for actuality. There are various ways of formalizing this signal and deriving it in the grammar. To fix ideas I assume that counterfactual marking cancels a presupposition that the set of topic worlds (e.g., a modal's domain of quantification) is a subset of the context set.[21]

I have said that no innovations are introduced into the semantics/pragmatics of strong necessity modals. 'Must ϕ' is given its familiar semantics of necessity: 'Must ϕ' is true at w iff ϕ follows from $P(w)$; and uses of 'Must ϕ' carry the usual indicative presupposition that the worlds being talked about are in the context set. One way of implementing the general indicative presupposition is as restricting the domain of the interpretation function to proper points of evaluation—contexts c and worlds w such that $w \in c$. Applying this to the case of a necessity modal yields:[22]

Definition 1. $[\![\text{Must}\,\phi]\!]^c = \lambda w: w \in c \,.\, \bigcap P(w) \subseteq [\![\phi]\!]^c$

(i) Su porceddu e' tottu abruxau it a d'essi boffiu arrustiu a fogu pracidu.
pork.meat get.burn.PPART want.PAST.COND.3SG roast.PPAST with low fire
'Pork meat got burned; it should have been roasted very slowly.'

(ii) Cussa femina bollit ascurtada.
DET.F.SG woman.F.SG want.PRES.3SG listen.to.PPART.F.SG
'That woman must be listened to.' (Fanari 2007, 128; Campidanese, Sardinian)

20 For a set of relevant worlds s and primary and secondary ordering sources $G_1(w)$, $G_2(w)$, let i be the set of $G_1(w)$-minimal s-worlds, and j be the set of $G_2(w)$-minimal i-worlds. Since $j \subseteq i$, that there is a ϕ-world in j asymmetrically implies that there is a ϕ-world in i (n. 18).

21 An alternative is to treat counterfactual marking as positively presupposing that the set of topic worlds isn't included in the context set. For discussion and technical implementations, see Stalnaker 1975, von Fintel 1998, Iatridou 2000, Ippolito 2003, Schlenker 2005, Arregui 2009, Bittner 2011.

22 In a framework with object-language world-variables, the presupposition would be that the assignment function maps the given world-variable to a world in the context set (cf. Schlenker 2005).

Uttering 'Must ϕ' predicates the proposition that follows from $P(w)$ of every world w in the context set. The treatment of 'must' vis-à-vis necessity is parallel to the treatment of 'may', 'can', etc. vis-à-vis possibility:

Definition 2. $[\![\text{May } \phi]\!]^c = \lambda w : w \in c \ . \ \bigcap(P(w) \cup \{[\![\phi]\!]^c\}) \neq \varnothing$

Uttering 'Must ϕ' predicates the necessity of throughout the context set just as uttering 'May ϕ' predicates the possibility of ϕ throughout the context set or uttering 'ϕ' predicates ϕ throughout the context set. It's in this sense that there is nothing special or distinctively "strong" in the semantics/pragmatics of (so-called) strong necessity modals.

Turn to weak necessity modals. A natural idea is that the counterfactual marking in the relevant languages cancels an assumption that the relevant worlds at which the prejacent is necessary are in the context set[23]—hence the observation in §2 that accepting 'Ought ϕ' doesn't require accepting that all the preconditions for ϕ to be necessary are satisfied. The apparent weakness of uses of 'Ought ϕ', compared to uses of 'Must ϕ', derives from failing to presuppose that the topic worlds where ϕ is necessary are in the context set. Call this approach a *modal-past approach* to weak necessity modals and the weak/strong necessity modal distinction. The remainder of this section presents several ways of implementing a modal-past approach in the formal semantics and begins investigating their costs and benefits.

One straightforward way of treating 'ought' as the semantic modal past of 'must' would be to treat 'ought' as having an ordinary semantics of necessity, like 'must', but lacking the presupposition that the worlds at which ϕ is necessary are in the context set.

Definition 3. $[\![\text{Ought } \phi]\!]^c = \lambda w . \bigcap P(w) \subseteq [\![\phi]\!]^c$ (v1)

Informally put, uttering 'Ought' places the necessity claim on the "conversational table," but doesn't conventionally commit one to its truth (cf. Silk 2016b). Implementing a modal-past approach as in Definition 3 faces pressing challenges in the discourse dynamics and compositional semantics. On a standard Stalnakerian theory of conversation, assertions propose to restrict the context set to worlds where the asserted content is true (Stalnaker 1978). But for any $w \in c$, 'Ought ϕ' is true at w according to Definition 3 iff 'Must ϕ' is true at w according to Definition 1.[24] An alternative mechanism would be needed to distinguish how 'ought' and 'must' update context, e.g. allowing uses of 'ought' to distinguish among worlds outside

23 On this I disagree with Arregui 2010. Though Arregui associates 'should' with a past morphology feature, she denies that the feature is interpreted with 'should'. Arregui maintains that 'Should ϕ' presupposes that the modal's quantificational domain is included in the context set (for nonstative 'ϕ'). We have seen that this is incorrect. Nonentailing uses '(Should ϕ) $\wedge \neg\phi$' are consistent (more on which in §5).

24 The right-to-left direction is obvious. Left-to-right: if $[\![\text{Ought } \phi]\!]^c(w) = 1$ then $\bigcap P(w) \subseteq [\![\phi]\!]^c$ but then $[\![\text{Must } \phi]\!]^c(w) = 1$ since, by hypothesis, $w \in c$.

the context set or have some non-eliminative effect. Second, it isn't obvious how the relative weakness of 'ought' vis-à-vis 'must' would carry over in embedded environments that shift the evaluation world, such as indicative conditionals or attitude ascriptions.

One way of avoiding these challenges is to build a counterfactual element into the semantics of 'ought'. Consider Definition 4, where h is a contextually supplied selection function.

Definition 4. $[\![\text{Ought}\,\phi]\!]^{c,w} = 1 \text{ iff } \forall w' \in h_c(w) : \bigcap P_c(w') \subseteq [\![\phi]\!]^c$ (v2)

To a first approximation, one can think of h as picking out a set of relevant worlds that are minimal/"preferred" in some contextually relevant sense—most normal, expected, desirable, etc., depending on the context (cf. Grosz 2012; Starr 2010, 167).[25] In (12) $h(w)$ might be the maximally w-normal worlds where you don't get distracted before putting the colored pencils away, no one hides them, etc.—worlds like w_N. In (26) $h(w)$ might be the maximally w-similar worlds where you have a job, the available charities are trustworthy, etc.—worlds like JRS from (8).

(26) You ought to donate to charity.

'Ought ϕ' is true iff *these* worlds verify the necessity of ϕ, i.e. iff ϕ follows from the relevant considerations P at every $w' \in h(w)$. The point from (11) that accepting 'Ought ϕ' doesn't require settling that ϕ is actually necessary is captured via h. The set of worlds $h(w)$ at which the necessity of ϕ is evaluated might include the evaluation world w, but it might not.

Definition 4 uses a simple selection function to determine the worlds at which the necessity of the prejacent is evaluated. The semantics could be complicated by deriving the set of selected worlds from more basic elements, such as independently represented orderings or premise sets of the relevant types (normality, desirability, etc.). The selection of worlds could also be treated as explicitly depending on the prejacent ϕ or a contextually relevant set of circumstances C, i.e. $h(w, \phi, C)$. This would reflect the idea that in interpreting 'Ought ϕ' one evaluates what follows from P at worlds satisfying certain conditions plausibly relevant to whether ϕ is a necessity.[26] In (12) one looks at worlds u satisfying what is normally the case in matters concerning where the colored pencils are, and one checks whether, conditional on such facts, the relevant information implies that the colored pencils are in the drawer, i.e. whether $\bigcap P_e(u) \subseteq \textit{drawer}$; in (26) one looks at worlds v satisfying what is normally or preferably the case in matters concerning whether to donate, and one checks whether, conditional

25 I often use double quotes around 'preferred' as a reminder that what is intended is the generalized notion of minimality, rather than a specifically bouletic or deontic notion.

26 In Silk 2012 I implemented these ideas by analyzing weak necessity modals as expressing a kind of conditional necessity (cf. Wertheimer 1972). I no longer endorse this way of capturing the points in §§2–3. Conditional necessity analyses make pressing the challenge of distinguishing uses of 'ought' from (implicit or explicit) conditional necessity claims (see below and §7.1).

on such facts, the relevant norms enjoin you to donate, i.e. whether $\bigcap P_d(v) \subseteq$ *donate*. Alternatively, the semantics might use a simple world-indexed function h, and such additional factors might be invoked in an extra-semantic account of how h is determined in concrete discourse contexts. For present purposes I assume the latter option.

The approach in Definition 4 raises the question of what distinguishes 'Ought ϕ' from counterfactual necessity sentences 'If χ, it would have to be that ϕ' (more on which shortly). A comparative semantics such as Definition 5 avoids this issue—where $\prec_w$ is a partial order on propositions along a contextually relevant dimension (likelihood, normality, desirability, etc.), and $s(w, p)$ is the set of closest p-worlds to w, understanding the relevant closeness relation as the relation that would figure in interpreting a counterfactual.

Definition 5.

$$[\![\text{Ought } \phi]\!]^{c,w} = 1 \text{ iff } \{u : u \in s(w, \bigcap P_c(u) \subseteq [\![\phi]\!]^c)\} \prec_w \{v : v \in s(w, \bigcap P_c(v) \not\subseteq [\![\phi]\!]^c)\} \quad \text{(v3)}$$

This treats 'Ought ϕ' as saying that it would be better (in a relevant sense) if ϕ was necessary (in a relevant sense).[27] Like Definition 4, Definition 5 avoids treating the truth of 'Ought ϕ' at w as requiring that ϕ be a necessity at w. The closest worlds u at which ϕ is necessary needn't be in the context set. Yet we can still see how uses of 'ought' may bear on interlocutors' views about what is necessary. 'Ought ϕ' introduces the possibility that ϕ is necessary and comments on it. The attitudinal comment is that the (closest) worlds in which ϕ is necessary are $\prec_w$-better—more desirable, normal, expected, etc., depending on the context. A critical issue is why the semantics for 'ought' and 'must' should be so dissimilar. One must provide a precise sense in which 'ought' is weaker than 'must' (logically, conversationally). Further, it isn't obvious how a comparative semantics like Definition 5 might shed light on how OUGHT-interpretations of *STRONG+CF* arise in other languages.

Definitions 3–5 provide several avenues for developing a modal-past implementation of the core ideas from §2. The semantics avoid analyzing accepting 'Ought ϕ' in terms of ϕ being a necessity (in any sense) at every w in the context set, and they offer precise representations of 'ought' as the notional modal-past of 'must'. 'Ought' is interpreted with respect to the same body of considerations P_c as 'must', and "introduces" the proposition $[\lambda w \,.\, \bigcap P_c(w) \subseteq [\![\phi]\!]^c]$ that the prejacent is a necessity relative to P_c; what distinguishes 'ought' is that the attitude taken toward that proposition needn't be acceptance—in Definition 3, by allowing improper points of evaluation; in Definition 4, by including an additional

27 Definition 5 could be refined depending on one's views on the semantics of comparative possibility (Lassiter 2011, Kratzer 2012), but the basic idea should be clear enough. Contrast the comparative probability semantics for 'ought' from Finlay (2009, 2010, 2014), which treats 'Ought ϕ' as saying (roughly) that ϕ is more likely than any relevant alternative to ϕ. In contrast, Definition 5 treats 'Ought ϕ' as making a comparative claim about the necessity of ϕ, rather than a comparative claim about ϕ, and it generalizes the relevant comparative notion via the context-dependent parameter $\prec$. (See Rubinstein 2014, Portner and Rubinstein 2016) for additional discussion of connections between weak necessity modals and comparatives/gradability.)

layer of modality and evaluating the truth of $[\lambda w\,.\,\bigcap P_c(w) \subseteq [\![\phi]\!]^c]$ throughout a set of possibly counterfactual worlds; and in Definition 5 by commenting on the preferredness (in a relevant sense) of $[\lambda w\,.\,\bigcap P_c(w) \subseteq [\![\phi]\!]^c]$. It will be useful in what follows to have a particular analysis at hand. To fix ideas I will generally assume the semantics in Definition 4. It's less clear how a comparative meaning such as Definition 5 might come to be associated with certain uses of *STRONG+CF* in other languages, and I think that the ideas motivating Definition 3 are more perspicuously developed in a dynamic setting (Silk 2016b). I leave more thorough comparisons for future work.

4.3 Entailments?

An initially plausible thought is that strong necessity modals entail weak necessity modals, which entail possibility modals. However, a worry for the modal-past analyses is that they seem to predict the consistency of sentences 'Ought ϕ and must $\neg\phi$' and sentences 'Ought ϕ and $\neg$may ϕ', as in (27).

(27) #I mustn't/can't/may not lie to Alice, but I ought to.

The 'must'/'may' conjunct is true at w iff the necessity of $\neg\phi$ is verified at w; the 'ought' conjunct is true at w, according to Definition 4, iff the necessity of ϕ is verified at every "preferred" (desirable, expected, etc.) world $w' \in h(w)$. Accepting (27) requires restricting the context set to worlds w such that $P(w)$ implies $\neg\phi$ and, for every $w' \in h(w)$, $P(w')$ implies ϕ. Absent further constraints on h, such conditions are consistent.

I think this prediction is actually a feature, not a bug. 'Must', 'ought', and 'may' cannot be ordered by logical strength simply in virtue of their conventional meanings. 'Must ϕ' $\vDash$ 'May ϕ', no doubt about that; but 'Must ϕ' $\nvDash$ 'Ought ϕ', and 'Ought ϕ' $\nvDash$ 'May ϕ'.

Epistemic 'Must ϕ' commits the speaker to high credence in ϕ and epistemic 'May ϕ' commits the speaker to some credence in ϕ. Epistemic 'must' entails epistemic 'may'. But epistemic 'Ought ϕ' doesn't commit the speaker to any credence in ϕ. This invalidates the entailments between epistemic 'ought' and epistemic 'must'/'may' (cf. Copley 2006, Silk 2016b, Swanson 2016a, Yalcin 2016; see n. 8).[28]

(28) #Alice can't be home yet; she hasn't called, and she always calls right away to let us know she got back safely. She must be home already. I hope there wasn't an accident.

(#$\neg$*Can* $\phi \wedge$ *Must* ϕ)

28 I use 'can't' for the external negation ($\neg <$ *can*), since with epistemic 'may not' the negation is internal. With 'mustn't' the negation is internal (*must* $<\neg$). With deontic 'may not' the negation is external ($\neg <$ *can*/*may*).

(29) Alice can't/must not be home yet; she hasn't called, and she always calls right away to let us know she got back safely. She ought to be home already. I hope there wasn't an accident.

(*Must* $\neg\phi \wedge$ *Ought* ϕ)
($\neg$*Can* $\phi \wedge$ *Ought* ϕ)

Although epistemic 'Ought ϕ' doesn't commit the speaker to any unconditional credence in ϕ, it still conveys a doxastic attitude. Informally, it conveys a conditional attitude about the necessity of ϕ given a body of evidence, on the assumption that conditions are relevantly normal (§§2–4; cf. Groefsema 1995, 72–73, Wedgwood 2007, 118–19, Yalcin 2016). One may infer from the denial of 'ϕ' that conditions aren't normal. Yet since 'ought' can be used to talk about the modal status of ϕ at nonactual possibilities, epistemic 'Ought ϕ' can still be true and accepted.

It's hard to come up with coherent deontic examples analogous to (29). Hard, but perhaps not impossible:

(30) I must tell my wife about the affair. I know I shouldn't; it'll only hurt her. But I must.

(31) I know I shouldn't tell my wife about the affair; it'll only hurt her. But I can't lie to her.

(32) #I must tell my wife about the affair. I know I can't; it'll only hurt her. But I must.

It isn't immediately obvious what to say about these examples. It's interesting that entailments from 'must' to 'ought' and from 'ought' to 'may' seem more compelling with nonepistemic readings. Yet insofar as we want unified semantics for modal verbs that generalize across readings, we should prefer an account that avoids treating the entailments as semantically valid.

5. "Weakness" in Weak Necessity Modals

§4 developed the §3-account by treating 'ought' as the notional modal past of 'must' and incorporating general insights about the semantics/pragmatics of counterfactual marking. This section shows how a modal-past account of the "weakness" of weak necessity modals systematizes several seemingly unrelated puzzles of entailingness and performativity with 'ought' and 'must'.

5.1. Diagnosing Weakness

A modal-past analysis gives precise expression to the informal intuition that 'ought' is weaker and more tentative than 'must'. In uttering 'Ought ϕ' the speaker fails to mark the necessity claim as being about worlds that are candidates for actuality. Yet, as Stalnaker notes,

"normally a speaker is concerned only with possible worlds within the context set, since this set is defined as the set of possible worlds among which the speaker wishes to distinguish" (1975, 69). So, uttering 'Ought ϕ' implicates that one isn't in a position to commit to ϕ being a necessity throughout the set of live possibilities. Grice's first quantity maxim—"Make your contribution as informative as is required" (Grice 1989, 26)—can then be exploited to generate a familiar upper-bounding implicature (Horn 1972, Gazdar 1979; n. 29). Using 'ought' implicates that for all one knows—better, for all one is willing to presuppose in the conversation—'Must ϕ' is false. This implicature has the usual properties of implicatures; it is reinforceable, cancellable, and suspendable:

(1a) I ought to help the poor, but I don't have to.
(2a) I ought to help the poor. In fact, I must.
(33) I ought to help the poor. Maybe I have to.

In (2a) the speaker first conveys that the worlds in which my helping the poor is deontically necessary needn't be live possibilities, and then commits that what holds in the former worlds also holds in the actual world. The implicature data with 'ought' can be treated analogously to implicature data with subjunctive conditionals.

(34) a. If you had the flu, you would have exactly the symptoms you have now. (cf. Anderson 1951, 53)
b. If you had the flu, you would have very different symptoms from the symptoms you have now.
c. If you had the flu, you would be sick. Maybe you do have the flu; you *are* pretty congested.

Likewise we can assimilate the tentativeness of 'ought' to the tentativeness of modal-past forms generally, as in non-counterfactual subjunctive conditionals ("future-less-vivid" conditionals) such as (20) above and (35).

(35) If you came to our party tomorrow—and I'm not saying that you will—you would have a great time.

Using the past form highlights the possibility that the marked clause might not ultimately be accepted. The basis of the scale between 'ought' and 'must' isn't fundamentally logical but epistemic strength (§4). 'Ought' and 'must' are ordered, not in terms of subset/superset relations in their domains of quantification, as per domain restriction accounts (§§2–3), but in terms of epistemic attitude toward the proposition that ϕ is necessary.[29]

29 Cf. Verstraete 2005b, 2006, Van Linden and Verstraete 2008.

5.2. Entailingness and Directive Force

Though many authors have claimed that epistemic 'Ought ϕ' expresses that ϕ is probable,[30] we have seen that this isn't quite right (§4). Epistemic 'Ought ϕ' doesn't commit the speaker to any unconditional credence in ϕ, as reflected in (23), reproduced in (36).

(36) a. Alice ought to be here by now, but she isn't.
b. #Alice must be (/may be, /is probably) here by now, but she isn't.

Epistemic '(Must ϕ) $\wedge \neg\phi$' is anomalous in a way that epistemic '(Ought ϕ) $\wedge \neg\phi$' is not. Surprisingly, there is robust evidence that this is the case with deontic readings as well.[31] When one wishes to convey that one thinks a given obligation won't be satisfied, one typically uses 'ought'/'should' rather than 'must'.

(37) a. He should/ought to come tomorrow, but he won't.
b. *He must come tomorrow, but he won't.
(Palmer 1990, 123; judgment Palmer's)

(38) a. You ought to help your mother, but you won't (/I know you won't).
b. ??You must help your mother, but you won't (/I know you won't).

As Eric Campbell insightfully puts it, "The idea that one *must* (not) do something . . . does not generally indicate that one option is simply better than another option, but that the other is out of bounds," "off the table," "closed off"; alternative possibilities become "*unthinkable*" (2014, 463–65). Of course obligations can go unfulfilled. What is interesting is that speakers appear to assume otherwise, at least for purposes of conversation, when expressing obligations with 'must'.[32]

Call data concerning sentences of the form '(*MODAL* ϕ) $\wedge \neg\phi$' *entailingness data* (though see n. 33). Our discussion in §4 suggests a natural way of capturing entailingness data such as (36)–(38). Evaluating 'Ought ϕ' can take us to worlds outside the context set when assessing the necessity of ϕ. There is no requirement that the value of the premise frame P at worlds outside the context set be compatible with the common ground. So, 'Ought ϕ' can be true at a world w in the context set even if every world in the context set is a $\neg\phi$-world; hence the

30 E.g., Sloman 1970, Horn 1972, Wertheimer 1972, Thomson 2008, Finlay 2010, 2014, Lassiter 2011, Wheeler 2013, Ridge 2014.

31 For theoretical discussion, see esp. Werner 2003, 124–37; Ninan 2005; Portner 2009, 103, 189–96; Silk 2016b. See also, e.g., Leech 1971, Wertheimer 1972, Harman 1973, Lyons 1977, Woisetschlaeger 1977, Williams 1981, Coates 1983, Palmer 1990, Sweetser 1990, Myhill 1996, Huddleston and Pullum 2002, Swanson 2008, 2016a, Close and Aarts 2010, Campbell 2014, Goddard 2014.

32 Cf.: "The basic strong obligation component common to ['got to', 'have to', and 'must'] is 'I can't think: *X* will not *V*'; note that this does not mean that the obligated event will inevitably take place but rather that the speaker is operating on this assumption" (Myhill 1996, 348–49).

coherence of '(Ought ϕ) $\wedge$ $\neg\phi$'. Since 'Must ϕ' doesn't have a broadly counterfactual element to its meaning, the context set must include $\bigcap P(w)$, for all worlds w in the context set. So, if 'Must ϕ' is accepted, $\neg\phi$ cannot be satisfied throughout the context set; hence the incoherence of '(Must ϕ) $\wedge$ $\neg\phi$' on any reading (qualifications shortly).[33,34]

How to derive these points about the relation between the modals' premise sets and the context set depends on general issues of presupposition and verbal mood (cf. §4). For instance, with modals that lack a counterfactual meaning component, there might be a context-set presupposition on the embedded clause that it denotes a proposition defined only at worlds in the context set (e.g., due to an indicative presupposition transmitted to the embedded clause, or a presupposition in the modal's lexical semantics). So, for 'must ϕ' to be true at w, $\bigcap P(w)$ must be a subset of $\{u\colon [\lambda v\colon v \in c\ .\ [\![\phi]\!]^{c,v} = 1](u) = 1\}$, hence $\bigcap P(w)$ must be a subset of the context set. Alternatively, there might be a general presupposition on the values of quantifier domain variables that they be compatible with the common ground at worlds in the context set. What is important here is that non-counterfactual modal constructions (modal sentences, indicative conditionals, etc.) presuppose that the domain of quantification is included in the context set. 'Ought ϕ' lacks this presupposition. Parallel to our points in §3, there is nothing distinctively "strong" posited about 'must' vis-à-vis entailingness. The incoherence of accepting '(Must ϕ) $\wedge$ $\neg\phi$' follows from the modal's ordinary semantics of necessity and general context-set presuppositions associated with indicative.

---

33 This explanation leaves open whether '$\neg\phi$' may be true at some worlds in the context set, and thus seems to predict that accepting deontic 'Must ϕ' is compatible with accepting the epistemic possibility of $\neg\phi$. This prediction appears to be borne out by corpus data. Verstraete (2007) cites the naturally occurring example in (i) with an imperative, adapted with 'must' in (ii).

(i) You've got to take a stand Tom. You've got to do it mate. [. . .] Don't stand for it, because if you do you'll just get trampled on. (CB ukspok) (Verstraete 2007, 242)

(ii) You mustn't stand for it, because if you do you'll just get trampled on.

(iii) ?You must go to confession, but maybe you won't (/you might not).

Though (iii) strikes me as somewhat anomalous, this is arguably due to a general norm of cooperative conversation that interlocutors do what they can to make the actual world be among the preferred/best worlds (cf. Portner 2007, 358). On this diagnosis, (iii) would be anomalous to the extent that it's anomalous to commit someone to help see to it that ϕ while expressly admitting the possibility of $\neg\phi$.

34 Alternatively, one might attempt to explain the entailingness data by positing additional features in the conventional meaning of 'must'. For instance, first, one might say that 'must' is only interpreted with respect to a modal base (or the union of several modal bases), not an ordering source (n. 11). (As far as the interpretation of 'must' goes, "all laws are natural laws" (cf. Piaget 1962, 340).) Since modal bases consist of propositions true at the evaluation world, 'Must ϕ' would entail 'ϕ'. Second, following Swanson 2016a one might include in the semantic entry for 'must' but not 'ought' a constraint requiring high credence in the prejacent. Third, one might appeal to more basic performative properties of 'ought' and 'must' (n. 35). However, it would be theoretically preferable if we could explain the data in terms of independent features of the semantics of 'ought' and 'must', as pursued in the main text. I argue below that the analysis in the main text derives the modals' contrasting performative properties without needing to take them as basic. We can explain the entailingness data without ad hoc stipulations about 'must'.

It's the *non*-entailingness of 'ought', and the consistency of '(Ought ϕ) $\wedge \neg\phi$', that is given special explanation.

Deontic 'ought' and 'must' are often thought to differ in illocutionary force. Paul McNamara characterizes the phenomena well:

> To say that one ought to take a certain option is merely to provide a nudge in that direction. Its typical uses are to offer guidance, a word to the wise ("counsel of wisdom"), to recommend, advise or prescribe a course of action . . . In contrast, to say that one must take a certain option is to be quite forceful. Its typical uses are to command, decree, enact, exhort, entreat, require, regulate, legislate, delegate, or warn. Its directive force is quite strong. (McNamara 1990, 156)

Many previous accounts capture this contrast by stipulating an ad hoc performative element in the lexical semantics of 'must'.[35] Our account of entailingness data suggests a strategy for deriving the contrasting speech-act properties of deontic 'ought' and 'must' from their static semantics and general pragmatic considerations. Accepting 'Must ϕ' is incompatible with denying 'ϕ'. So, if the truth of 'ϕ' is assumed to depend on the actions of the addressee, updating with 'Must ϕ' will commit her to seeing to it that ϕ (or, in the general case, commit the interlocutors to presupposing that the subject of the obligation is committed to seeing to it that ϕ) (cf. Bybee et al. 1994). So, it's no surprise that 'must' should often be thought to be conventionally directive. By contrast, since accepting 'Ought ϕ' is compatible with denying 'ϕ', updating with 'Ought ϕ' needn't commit anyone to seeing to it that ϕ. Even if deontic 'ought' can be used to perform a directive speech act in certain contexts, it doesn't do so as a matter of conventional meaning. Yet given our discussion of the conversational role of 'ought' (§§2–3), it's unsurprising that utterances of deontic 'Ought ϕ' should often perform more moderate speech acts of recommending or advising. Uttering 'Ought ϕ' can convey one's preference that 'ϕ' be accepted, but without imposing the truth of 'ϕ' on the common ground. As Gibbard, following Stevenson, writes, the speaker "is making a conversational demand. He is demanding that the audience accept what he says, that it share the state of mind he expresses" (Gibbard 1990, 172), though in a "more subtle, less fully conscious way" than by issuing "an imperative" (Stevenson 1937, 25)—or, we might say, than by using 'must'. Deontic 'ought' generally provides a less face-threatening alternative to deontic 'must', in particular in contexts where the speaker might be construed as imposing on the addressee or relevant subject.[36]

---

35 See, e.g., Ninan 2005; Portner 2007, 363–65, 2009, 103–5, 189–96; Swanson 2008, 1203–4 (cf. Finlay 2014, 172–74; Ridge 2014, 27–36). See Van Linden and Verstraete 2011, 153 and Van Linden 2012, 69 on differences in force among weak and strong deontic adjectives.

36 Cf. "Bradshaw said, he must be taught to rest. Bradshaw said they must be separated. 'Must,' 'must,' why 'must'? What power had Bradshaw over him? 'What right has Bradshaw to say "must" to me?' he demanded" (*Mrs Dalloway*, Virginia Woolf). See also n. 39.

Our treatments of differences in "strength" between 'ought' and 'must' are compatible with the observation that uses of 'must' may carry an intuitively weaker conversational force in certain contexts. Consider (39).

(39) [Context: You're hosting a party. You wish to offer the guests a cake that you baked yourself. You say:]
   a. You must have some of this cake.
   b. You should/ought to have some of this cake.
   c. You may have some of this cake. (cf. Lakoff 1972a, 910)

As Lakoff observes, using 'must' would be most polite, conveying an offer, while using 'should' would be less polite and using 'may' would be flat-out rude. It's an interesting question what conversational factors are responsible for such apparent reversals of the modals' felt forces. Given our focus on the weak/strong necessity modal distinction I put the issue aside, since the phenomenon generalizes to intuitively "weak" uses of imperatives, as in permission/invitation uses such as (40) (von Fintel and Iatridou 2015), and intuitively "strong" uses of possibility modals, as in (39c) and command uses such as (41).

(40) Here, have some of this cake!

(41) [Context: Celebrity to entourage:]
You may/can leave now.

5.3 Qualification: Endorsing and Nonendorsing Use

I have said that '(Must ϕ) $\wedge \neg\phi$' cannot be coherently accepted on any reading. This claim needs to be qualified. Though there is robust data attesting to the anomalousness of deontic '(Must ϕ) $\wedge \neg\phi$' (n. 31), some speakers report being able to hear sincere uses as consistent in certain contexts.[37] However, a key observation is that even speakers who can hear examples such as (42) as consistent agree that it would be more natural to use a strong necessity modal such as 'have to' or 'be required to', as in (43).

(42) I must go to confession; I'm a Catholic. But I'm not going to. I haven't practiced for years.

(43) I have to (/I'm required to) go to confession; I'm a Catholic. But I'm not going to. I haven't practiced for years.

37 I haven't seen this judgment expressed in published work, though I've heard it voiced in personal conversation. Thanks to Jan Dowell for discussion.

In (43) it's consistent for the speaker to dismiss going to confession because she isn't endorsing the norms which imply that she is obligated to do so. She is simply reporting what these norms require.

This observation suggests that one can hear examples such as (42) as felicitous to the extent that one accepts "objective" uses of 'must' in (roughly) the sense of Lyons 1977, 1995. Adapting Lyons's terminology, say that a modal is used *endorsingly* in an utterance of '*MODAL* ϕ' if the utterance presents the speaker as endorsing/accepting the considerations with respect to which the modal is interpreted; and say that the modal is used *nonendorsingly* if it doesn't. Among strong necessity modals, 'be required to' is typically used nonendorsingly; 'have to' and '(have) got to' are more flexible, with 'have to' tending more toward the nonendorsing side of the spectrum and '(have) got to' toward the endorsing side; and 'must' is typically used endorsingly.[38] It's easier to hear sincere utterances with (e.g.) 'have to'/'be required to' as compatible with the speaker's rejecting or being indifferent about the considerations that would verify the modal claim:

(44) [Context: Some friends are deciding whether to go home or stay out late for a party.]
   a. #You must get home by 11, but I don't care whether you do.
   b. #Bert must get home by 11. Aren't his parents stupid? I would stay out if I were him.

(45) a. You have to (/are required to) get home by 11, but I don't care whether you do.
   b. Bert has to (/is required to) get home by 11. Aren't his parents stupid? I would stay out if I were him.

When one wishes to describe relevant norms without necessarily expressing endorsement of them, one typically uses (e.g.) 'have to' rather than 'must'.

So, the promised qualification is this: endorsing uses of '*STRONG* ϕ' are incompatible with a denial of 'ϕ'. It's only with endorsing uses of strong necessity modals that $\bigcap P(w)$ must be included in the context set. What makes the claims about entailingness and performativity particularly compelling in the case of 'must' (versus e.g. 'have to') is that 'must' is typically

38 These generalizations are supported by robust data in descriptive linguistics. In addition to Lyons 1977, 1995, see Leech 1971, 2003, Lakoff 1972a,b, Coates 1983, Perkins 1983, Palmer 1990, 2001, Sweetser 1990, Myhill 1995, 1996, Myhill and Smith 1995, Verstraete 2001, Huddleston and Pullum 2002, Smith 2003, Leech et al. 2009, Close and Aarts 2010, Goddard 2014. For critical discussion see Swanson 2012, 2016b, Silk 2015b, 2016a. The distinction between endorsing/nonendorsing uses has been noted under various labels in diverse disciplines (e.g., Hare 1952, von Wright 1963, Verstraete 2007, Silk 2016a). I adapt Lyons's terminology since use of 'subjective' and 'objective' can be fraught. I use 'endorsement' as a cover term for acceptance attitudes of various kinds.

used endorsingly.[39] But to the extent to which one finds nonendorsing uses of 'must' acceptable, to that same extent one is predicted to find uses of deontic 'Must ϕ' to be nonentailing and lack directive force.

What is distinctive about weak necessity modals is that even when they are used endorsingly, they are nonentailing and may lack imperative force. 'Ought' and 'should' are like 'must' in typically being used endorsingly (n. 38). Parallel to (44)–(45), the claims in (46) with 'ought' would be more naturally expressed with a modal such as 'supposed to', as in (47).

(46) a. #You ought to get home by 11, but I don't care whether you do.
b. #Bert ought to get home by 11. Aren't his parents stupid? I'd stay out if I were him.

(47) a. You're supposed to get home by 11, but I don't care whether you do.
b. Bert is supposed to get home by 11. Aren't his parents stupid? I'd stay out if I were him.

Yet a characteristic use of 'Ought ϕ' is with an explicit or implicated denial of 'ϕ' (§4).

6. Negotiability and Collective Commitment

Along the way we have noted various ways in which the proposed approach to the weak/strong necessity modal distinction differs from other approaches in the literature. For instance, alternative approaches generally agree in treating uses of 'Ought ϕ' as predicating a distinctive kind of necessity, namely weak necessity, of the prejacent ϕ at the actual (evaluation) world. Very roughly: On domain restriction accounts (Copley 2006, von Fintel and Iatridou 2008, Swanson 2011, Rubinstein 2012, Charlow 2013), ϕ is a weak necessity if ϕ is true throughout a certain set S of worlds, where S is a subdomain of the set of worlds quantified over by 'must'. On probabilistic/comparative possibility accounts (Finlay 2009, 2010, 2014, Lassiter 2011), ϕ is a weak necessity if ϕ is sufficiently likely/desirable, or more likely/desirable than any relevant alternative to ϕ. Although Yalcin's (2016) normality-based semantics denies that 'ought' and 'must' are logically related, 'Ought ϕ' is still interpreted by evaluating the truth of ϕ throughout a set of minimal worlds relative to the actual world.

The approach in this paper rejects treating acceptance of 'Ought ϕ' in terms of ϕ being a necessity, in any posited sense of necessity, at every candidate for the actual world. The apparent "weakness" of (so-called) weak necessity modals is diagnosed instead in terms of a failure to presuppose that the relevant worlds at which the prejacent is a necessity are in the context set (§4). For example, the semantics in Definition 4 adds a layer of modality,

39 It isn't implausible that the drastic decline in frequency of deontic 'must' is due in no small part to the above features of its meaning and use (cf. Myhill 1995, 1996, Krug 2000, Smith 2003, Tagliamonte 2004, Leech et al. 2009, Close and Aarts 2010, Goddard 2014).

predicating the ordinary necessity claim [$\lambda w \,.\, \bigcap P_c(w) \subseteq [\![\phi]\!]^c$] of every world in a certain set of possibly counterfactual worlds. The proposed modal-past approach (even if not implemented in precisely this way) captures contrasting discourse properties of 'ought' and 'must' such as differences in conversational force and relations to standing contextual assumptions; it captures logical properties such as differences in entailingness and the lack of entailments between 'ought' and 'must'/'may'; it captures relations between weak necessity modals and broader modal-past phenomena; and it generalizes across flavors of modality.

There are insights from previous accounts that are preserved in the account developed here. The semantics in §4 allow for connections between weak necessity modals and common ground assumptions (n. 6); mood, counterfactuality, and conditionality (nn. 17, 23, 26); and notions of comparison (n. 27), normality (Makinson 1993, Frank 1996, Yalcin 2016), and probability (though not unconditional probability; n. 30). Previous accounts are often developed with an eye toward one or several of these issues to the exclusion of others. Although we haven't examined each of the connections in equal depth, I hope the discussion has illustrated the fruitfulness of a modal-past approach and its potential for systematizing diverse linguistic phenomena.

Before concluding I would like to compare in more detail the account of the weak/strong necessity modal distinction in this paper with the domain restriction account in Rubinstein 2012, which constitutes the most extensively developed alternative from the literature (see also Portner and Rubinstein 2012, 2016, Rubinstein 2014). We have already observed general points of disagreement between the modal-past approach and domain restriction accounts (see also §§3, 5). Here I focus on the distinctive feature of Rubinstein's account: the appeal to *collective commitment* to a body of priorities (norms, goals, ideals).[40]

Rubinstein's main innovation is to supplement von Fintel and Iatridou's domain restriction semantics with a substantive account of the distinction between primary vs. secondary ordering sources (§4): what makes a primary ordering source "primary" is that it includes premises which are presupposed to be collectively committed to by the interlocutors; what makes a secondary ordering source "secondary" is that it includes premises which are presupposed *not* to be collectively committed to by the interlocutors. 'Must' is only interpreted with respect to primary premises; 'ought' is logically weaker in also being interpreted with respect to secondary premises, which are presupposed to be not collectively endorsed:

> Strong necessity modals are only sensitive to prioritizing premises that the conversational participants are presupposed to be collectively committed to . . . [I]f any participant in the

40 Rubinstein doesn't examine epistemic modals. I leave open how the account would be extended to epistemic readings. As discussed previously, there will be general worries regarding the lack of entailments between epistemic 'ought' and 'may/'must', and the fact that epistemic 'ought' doesn't in general quantify over a set of worlds that are regarded as epistemically possible (§§4–5). See Silk 2018 for critical discussion of Portner and Rubinstein's (2012) appeal to collective commitment in an account of mood selection.

> conversation were given the chance to defend these priorities, it is assumed in the context of the conversation that they would do so. Weak necessity modals take into account all these premises plus some more. For these additional premises, lack of collective commitment is presupposed . . . [A] speaker uses a weak modal when he or she believes (perhaps mistakenly) that the secondary priorities it depends on are still up for discussion.
>
> (Rubinstein 2012, 51–52)

In (48) although the speaker is committed to a priority favoring cost-effectiveness, it isn't presupposed that Alice is committed to it (cf. Rubinstein 2012, 55–60). This lack of collective commitment is what calls for using 'ought', on Rubinstein's view.

(48) [Context: Alice is considering whether to take the subway or a cab to a concert. The subway is cheaper; the cab is quicker. You say:]
You ought to (/should, /?have to, /?must) take the subway.

For clarity, use 'commitment$_R$' for the notion of commitment in Rubinstein's analyses. Rubinstein explicitly identifies commitment$_R$ to a priority p with *endorsing that p is desirable* (e.g., 2012, 78; also Portner and Rubinstein 2012, 471, 475, 477–81). What determines whether p is a "primary" or "secondary" priority, according to Rubinstein, is the interlocutors' mutual presuppositions about the desirability of p.

§5 delineated two dimensions along which modals differ: strength and tendencies for endorsing/nonendorsing use, i.e. the extent to which uses of '*MODAL* ϕ' convey the speaker's endorsement of the considerations that would verify the modal claim. Delineating these dimensions brings out problems with diagnosing the weak/strong necessity modal distinction in terms of collective commitment$_R$. As we have seen, the weak/strong necessity modal distinction crosscuts the distinction between necessity modals that express endorsement in the above sense and those that don't, and hence crosscuts the distinction between necessity modals that express collective commitment$_R$ and those that express a lack of commitment$_R$.

First, there are uses of weak necessity modals that express collective endorsement of the relevant priorities. In (49) we are both publicly committed$_R$ to the value of family; my helping my mother is "publicly endorse[d] . . . as desirable" (Rubinstein 2012, 78) by every conversational participant. Yet 'ought' is felicitous, indeed preferred.

(49) *Me*: Family is very important. I think I would rather stay here.
You: I agree. You ought to tend to your mother.

Second, there are uses of strong necessity modals where the interlocutors expressly deny commitment$_R$ to the relevant priorities, as in (43) and (45). The lack of endorsement can even be common ground:

(50) [Context: Our parents are asleep. We've settled on staying out to go to a party.]
You: When is curfew, again? We need to make sure that we tell Mom we got back before then if she asks.
Me: We have to be home by 11. Aren't her rules stupid? This party is going to be great.

Counterexamples such as these aren't atypical. Corpus studies attest to variations among weak necessity modals, and variations among strong necessity modals, vis-à-vis tendencies to express collective commitment$_R$. Collective commitment$_R$ has been invoked as a basis for distinguishing among the modals in each class.[41] Summarizing his corpus analyses of 'ought' and 'should' in contemporary American English, Myhill concludes, "Using *ought* suggests that people have the same feelings about the specific obligation in question and there is agreement about it, while *should* does not suggest the same feelings or agreement" (1997, 8). Far from being exceptional, expressing collective commitment$_R$ with 'ought' is the norm, as in the naturally occurring discourse in (51) (slightly abbreviated).

(51) *A*: I won't tell anyone . . . but the Dean, of course.
B: And Mrs. Reynolds.
A: Yes. She ought to know. (cf. Myhill 1997, 10)

Conversely, though 'must' tends to express collective commitment$_R$, other strong necessity modals such as '(have) got to' do not. '(Have) got to' differs from 'must' in typically being "associated with *conflicts* between the speaker and the listener" (Myhill 1996, 365), as in the naturally occurring example in (52).

(52) Edie, you've got to stop bothering me when I'm working. (Myhill 1996, 369)

Collective commitment$_R$ to the contents of premise sets may affect the distribution of modals in discourse, but it isn't what explains the distinction between weak and strong necessity modals.

A modal-past approach captures intuitions about negotiability and collective commitment which may be motivating Rubinstein's account. Uses of strong necessity modals "presuppose (collective) commitment" in the same ordinary sense as uses of any context-sensitive expression: one commits that the value of the context-dependent item is as one's utterance assumes, and that the world is as one's utterance says it is, given this assumed value. Such commitments are compatible with not endorsing-as-desirable the premises $p \in P(w)$, for any $w \in c$. In (50) we can accept that the house rules require us to be home by 11 while denying that those rules are desirable or that they are to guide our plans. Conversely, uses

41 See, e.g., Joos 1964, Myhill and Smith 1995, Myhill 1996, 1997, Verstraete 2001, Facchinetti et al. 2003, Leech et al. 2009, Close and Aarts 2010, Goddard 2014.

of weak necessity modals express "negotiability" in the sense that they don't conventionally commit one to being in a world where the relevant considerations verify the necessity of the prejacent. Failing to express commitment to the prejacent's being necessary is compatible with (collectively) endorsing the considerations encoded in the given premise frame or the premises from which the prejacent would follow. In (5) we can endorse the value of family even if we prefer not to settle how it interacts with other potentially competing values.

7. Conclusion

This paper developed a *modal-past approach* to 'ought' and the distinction between so-called weak necessity modals ('ought', 'should') and strong necessity modals ('must', 'have to'). There is nothing specially "strong" about the necessity expressed by strong necessity modals. Strong necessity modals are given their ordinary semantics/pragmatics of necessity; uses of 'Must ϕ' predicate the (deontic/epistemic/etc.) necessity of the prejacent ϕ of every candidate for the actual world. The apparent "weakness" of weak necessity modals derives from their bracketing whether the prejacent is necessary in the actual world. Uses of 'Ought ϕ' fail to presuppose that the topic worlds in which ϕ is necessary are included in the context set. 'Ought ϕ' can be accepted without needing to settle that the relevant considerations (norms, goals, etc.) which actually apply verify the necessity of ϕ. This analysis carves out important roles for weak necessity modals in conversation and deliberation. As emphasized throughout Gibbard's (1990, 2003, 2012) developments of expressivism, 'ought' affords a means of coordinating our *conditional* attitudes and plans. Weak necessity modals allow us to entertain and plan for hypothetical continuations or minimal revisions of the current context; they afford conventional devices for coordinating our norms, values, expectations, without having to settle precisely how the relevant considerations apply and compare. The proposed account systematizes a spectrum of semantic and pragmatic data—e.g., concerning the relative felicity of weak and strong necessity modals, relations between uses of weak and strong necessity modals and standing contextual assumptions, morphosyntactic properties of expressions of OUGHT crosslinguistically, and contrasting logical and illocutionary properties of weak and strong necessity modals. The range of linguistic phenomena that are unified under and explained by the account lend it a robust base of support.

The data considered here aren't the only data to be explained by an overall theory of weak and strong necessity modals. For instance, there are additional linguistic contrasts between the classes of weak and strong necessity modals, such as in data with incomparabilities, comparatives, quantifiers, conditionals, modifiers, and neg-raising.[42] Second, we briefly examined one additional dimension of difference among modals in tendencies for endorsing/non-endorsing use. It's worth investigating interactions between weak/strong necessity modals and other such

42 See, e.g., Lassiter 2011, 2012, Swanson 2011, Rubinstein 2012, 2014, Charlow 2013, Iatridou and Zeijlstra 2013, Silk 2013, 2015b, 2016b, Portner and Rubinstein 2016.

differences in modal meanings, e.g. with implicatures (cf. Verstraete 2005a). Third, although I argued against treating 'ought' and 'must'/'may' as being ordered by logical or quantificational strength, more thorough comparisons of inference patterns with different types of readings are needed. Fourth, our discussion highlighted interactions between weak and strong necessity modals and general issues such as context-sensitivity, counterfactuality, attitude expression, and performativity. These interactions afford rich avenues for future research. In closing I would like to raise two potential avenues for future work, one more narrowly linguistic, one more philosophical.

7.1. Counterfactuality?

In §4 we noted that many languages express a notion of OUGHT by using the form of a strong necessity modal used in counterfactuals. Such languages use the same string to express OUGHT and counterfactual necessity (WOULD HAVE TO). Yet as von Fintel and Iatridou (2008, 128–31) observe, 'ought' cannot generally be replaced by 'would have to':

(53) a. I ought to help the poor, but I don't have to.
b. ??I would have to help the poor, but I don't have to.

Open questions include how non-lexicalized expressions of OUGHT are diachronically related to the dedicated lexical items; how certain uses of *STRONG+CF* might come to be conventionally interpreted as expressions of OUGHT; how counterfactual interpretations of *STRONG+CF* are related to meanings of grammaticalized forms like 'ought'; and how meanings of the grammaticalized forms compare across languages. The modal-past analysis in Definition 4 avoids conflating 'ought' with 'would have to' and giving 'Ought ϕ' the semantics of an implicit counterfactual necessity claim. The necessity of ϕ isn't evaluated at the closest ψ-worlds, for any implicit condition ψ, but at the relevant minimal/"preferred" (desirable, normal, etc.) worlds determined by h_c. Although the focus of this paper has been on lexicalized English weak necessity modals, it's worth considering how such an interpretation might become associated with certain uses of *STRONG+CF* in other languages. Channeling my inner Gibbard, "What I can suggest will have to be quite speculative, but the speculations, I hope, may prompt more solid investigation" (1990, 61).

Using a clause that lacks the usual indicative presupposition places a burden on the interpreter's task of inferring which possibilities are being talked about. A condition specifying which possibilities are relevant typically must be salient, either implicitly or in the linguistic context, as in (54). (cf. Schueler 2011).

(54) [Context: We're trying out guitars in a music store. Looking at an expensive vintage Les Paul, I say:]
a. (If I bought it,) my partner would kill me.
b. (If I wanted to buy it,) I would have to check with my partner first.

So, if no condition is retrievable, using 'would have to' is anomalous, as in (55a). Uses of 'ought' lack this salience/retrievability requirement (cf. (55b)).

(55) [Context: We're strangers in a hotel lobby. I notice you fumbling with your bags. I say:]
a. ??Here, I would have to help you.
b. Here, I ought to help you.

A hypothesis, then, is that distinctive discourse effects of weak necessity modals are products of using an expression that is neither marked as being about the context set (as via indicative mood) nor indicated as being about some other topical possibility (via an implicit or explicit condition).

Suppose we are inquiring about the necessity of ϕ. Uttering '*STRONG* ϕ' requires settling which considerations apply and aren't defeated. We might not be prepared to restrict the future course of the conversation in this way. Nevertheless each of us takes some ways of extending the conversation and addressing the question under discussion to be better or more likely than others. Using the modal-past form allows us to consider the necessity of ϕ as holding not necessarily in the current context, as with a strong necessity modal, but in a preferred (likely/normal/desirable) continuation or minimal revision of the current context, whatever that might turn out to be (§§2–3). One's judgments about replacing 'would have to' with 'ought' should thus improve to the extent that the antecedent for 'would have to' describes what one regards as "preferred" conditions relevant for evaluating questions about the necessity of ϕ. Indeed the discourse effects of 'would have to' and 'ought' are strikingly closer in (57) than in (56).

(56) a. (If I was a mobster, which I'm not,) I would have to kill you.
b. ??I ought to kill you.

(57) [Context: Alice is considering with her mother whether to take the A- or C-train. The A is quicker; the C is safer. Her mother says:]
a. If safety was most important, you would have to take the C-train. In fact, safety is more important, as we can agree. So you must/have to take the C.
b. You ought to take the C-train. In fact, safety is most important, as we can agree. So you must/have to take the C.

As often occurs in grammaticalization and pragmaticalization,[43] the original (counterfactual) meaning may be semantically bleached, and the reinterpreted (weakness) meaning becomes more abstract; the attitude expressed about the necessity claim can thus appear

43 See, e.g., Horn 1984, Heine et al. 1991, Lehmann 1995, Traugott 1995, Diewald 2011, Narrog and Heine 2011, Davis and Gutzmann 2015.

vague or nonspecific. Such an interpretation is essentially what is delivered by the lexically unspecified generalized notion of minimality associated with *h* in Definition 4.[44] This grammaticalization path predicts the observed variations in how OUGHT is expressed crosslinguistically—i.e., with some languages using only a conventionalized weakness interpretation of *STRONG+CF* (French); some languages using only a grammaticalized form (English); and some languages using both (German).

To my knowledge, no general account has been given of how weakness interpretations of past forms are derived, or how weakness vs. implicit-counterfactual interpretations of past forms are determined across contexts (cf. (17)–(19), n. 16). The above hypothesis may provide a basis for weakness (tentativeness, politeness) effects associated with modal-past forms generally. Thorough synchronic/diachronic crosslinguistic work on the interpretation of modal-past forms, and how OUGHT-interpretations of *STRONG+CF* are conventionalized and in some cases grammaticalized, is called for.

7.2. Philosophical Therapy?

I would be remiss if I concluded without asking how, if at all, our linguistic inquiry might inform theorizing about a "primitive concept OUGHT" (Gibbard 2012, 204). How might investigating a word 'ought' teach us something about "concepts fraught with" a "primitive ought" (Gibbard 2003, x, 21, 179; 2012, 14)?[45] Recall our discussions of the practicality of deontic 'ought' and 'must' judgments. It's common in discourse and deliberation to investigate what we are actually required to do. We wish to guide our planning and influence one

44 The hypothesis in the main text considers weakness uses of *STRONG+CF* on the model of a conditional consequent (i.e., lacking a particular retrievable antecedent). An alternative is to understand the uses on the model of the antecedent. Interestingly, there is a precedent for associating an interpretation like the one given to *h* in Definition 4 with a certain type of 'if'-clause. Grosz (2012) argues that optatives and exclamatives should be treated on the model of nonlogical 'if'-clauses under a covert exclamation operator *EX* ("nonlogical" in the sense that the embedded clause provides the subject matter of the attitude), as reflected in (i)–(ii).

(i) Oh, that it would snow!
  a. *EX* [it snows]
  b. ≈ "It would be good if it snowed"

(ii) Oh, that it snowed!
  a. *EX* [it snowed]
  b. ≈ "It's surprising that it snowed"

Grosz's *EX* operator is essentially a scalar analogue of *h*, also generalized to express a lexically unspecified attitude toward the embedded proposition. A Grosz-style scalar version of Definition 4—that 'Ought ϕ' is true iff the proposition that ϕ is necessary is at least as *h*-"preferred" (desirable/normal/etc.) as a contextually relevant standard—might help capture connections between weak necessity modals and gradability/comparatives (Finlay 2014, Rubinstein 2014, Portner and Rubinstein 2016). Crosslinguistic comparisons among morphological and lexical expressions of OUGHT, nonlogical 'if'-clauses, and optatives and exclamatives may provide fruitful avenues to explore.

45 See Silk 2015b (esp. §§6–7) for additional discussion and applications.

another's behavior in light of the norms we accept. An endorsing strong necessity modal like 'must' is well suited to the task. However, using 'must' is often awkward. We may want to talk about obligations which held in the past, or which may go unfulfilled or be overridden in the future. Or we may want to communicate information about a body of norms without necessarily registering commitment to them or enjoining others to share in such commitment. A modal like 'ought' or 'be required to' may thus be more suitable. There is a range of expressive resources at our disposal for coordinating our actions and attitudes. This is for the better given the variety of our purposes. But it also raises a philosophical risk. Bracketing differences among necessity modals might turn out to be harmless for the purposes of (meta)normative inquiry. But it might not.

A central innovation of Gibbard's expressivism is the thesis that normative concepts are essentially *plan-laden*:

(58) "To believe that one ought to do *X* is to plan to do *X*."
"*Thinking what I ought to do is thinking what to do*."
"*Ought* claims . . . are claims about what to do."
(Gibbard 2003, ix–x, 10; 2012, 204)

Generalizing: "The clear distinctive feature of normative concepts, I now think, lies in their conceptual ties. Oughts of action tie in conceptually with acting" (Gibbard 2011, 36). Many take it as obvious that some form of judgment internalism is true. After all, normative judgments are constitutive of deliberation, and deliberation is essentially practical; its aim is action. But many find clear counterexamples. What about the psychopath, or someone who is tired or depressed (such as perhaps the poor reader who has made it thus far)?

Attending to various dimensions of difference among necessity modals may shed light on conflicting intuitions about the ostensible "plan-ladenness" of "concepts fraught with ought." Gibbard hedges: the concepts "fraught with ought" are so-fraught "not for every sense of the term ['ought']," but for the "crucial sense" explained by the pattern in (58) (2003, x). It is revealing that although Gibbard couches the thesis about the practicality of the normative using 'ought', he pumps the intuition using 'must':

> Oughts of action tie in conceptually with acting. Take, for example, the belief that the building is on fire and the one and only way to keep from being burned to a crisp is to leave forthwith. If that's the case, we'd better leave forthwith, but it isn't strictly incoherent, conceptually, to have this belief and not to leave. Contrast this with the normative belief that one *must* leave forthwith. It is, I maintain, conceptually incoherent to hold this belief and not leave, if one can.
> (Gibbard 2011, 36)

As we have seen, it's hard to hear a sincere utterance of 'Must ϕ' as consistent with the speaker's being indifferent about whether ϕ. Such judgments aren't nearly as anomalous

when expressed with weak necessity modals or modals that are more naturally used non-endorsingly (§§4–6).

(59) #I *must* (/I've *got* to) get home by 10, but forget that; I'm not going to.
(60) I {ought to, should, am supposed to, have to} get home by 10, but forget that; I'm not going to.

We should be wary of general claims about normative language and judgment. A thesis such as (58) may seem compelling when considering examples using terms that are paradigmatically entailing and endorsing; but when we consider cases using other terms, counterexamples can appear in the offing. If only "fraught with MUST" were catchier.

The skeptically inclined might be apt to wonder what is at-issue in debates about judgment internalism—i.e., if not linguistic issues about directive/entailing/endorsing uses of language, or empirical psychological issues about conditions under which different attitudes and judgments guide and motivate us. One might wonder how probative it is to couch metaethical inquiry in terms of a class of "normative" language/concepts/judgments at all. More fruitful, perhaps, to leave talk of the "normative" to the side, and ask directly about motivational states of mind and directive uses of language, and what type of person to be and how to live.

References

Anderson, Alan Ross (1951). "A Note on Subjunctive and Counterfactual Conditionals." *Analysis* 11: 35–38.

Åqvist, Lennart (2002). "Deontic Logic." In *Handbook of Philosophical Logic, Vol. 8*, 2nd ed., edited by Dov M. Gabbay and Franz Guenthner, 147–64. Dordrecht: Kluwer Academic Publishers.

Arregui, Ana (2009). "On Similarity in Counterfactuals." *Linguistics and Philosophy* 32: 245–78.

——— (2010). "Detaching *If*-Clauses from *Should*." *Natural Language Semantics* 18: 241–93.

van der Auwera, Johan, and Vladimir A. Plungian (1998). "Modality's Semantic Map." *Linguistic Typology* 2: 79–124.

Bittner, Maria (2011). "Time and Modality without Tenses or Modals." In *Tense across Languages*, edited by Monika Rathert and Renate Musan, 147–88. Tübingen: Niemeyer.

Brandt, R. B. (1964). "The Concepts of Obligation and Duty." *Mind* 73: 374–93.

Bybee, Joan (1995). "The Semantic Development of Past Tense Modals in English." In *Modality in Grammar and Discourse*, edited by Joan Bybee and Suzanne Fleischman, 503–17. Amsterdam: Benjamins.

Bybee, Joan, Revere Perkins, and William Pagliuca (1994). *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Campbell, Eric (2014). "Breakdown of Moral Judgment." *Ethics* 124: 447–80.

Charlow, Nate (2013). "What We Know and What to Do." *Synthese* 190: 2291–323.

Close, Joanne, and Bas Aarts (2010). "Current Change in the Modal System of English: A Case Study of *Must*, *Have To* and *Have Got To*." In *The History of English Verbal and Nominal Constructions*, edited by Ursula Lenker, Judith Huber, and Robert Mailhammer, 165–81. Amsterdam: John Benjamins.

Coates, Jennifer (1983). *The Semantics of the Modal Auxiliaries*. London: Croom Helm.

Collins, Peter (2007). "*Can*/*Could* and *May*/*Might* in British, American and Australian English: A Corpus-Based Account." *World Englishes* 26: 474–91.

Condoravdi, Cleo (2002). "Temporal Interpretation of Modals: Modals for the Present and for the Past." In *The Construction of Meaning*, edited by David I. Beaver, Luis D. Casillas Martínez, Brady Z. Clark, and Stefan Kaufmann, 59–88. Stanford: CSLI Publications.

Copley, Bridget (2006). "What Should *Should* Mean?" MS, CNRS/Université Paris 8.

Davis, Christopher, and Daniel Gutzmann (2015). "Use-Conditional Meaning and the Semantics of Pragmaticalization." In *Proceedings of Sinn und Bedeutung 19*, edited by Eva Csipak and Hedde Zeijlstra, 197–213. Göttingen: University of Göttingen.

Diewald, Gabriele (2011). "Pragmaticalization (Defined) as Grammaticalization of Discourse Functions." *Linguistics* 49: 365–90.

Dreier, James (2009). "Relativism (and Expressivism) and the Problem of Disagreement." *Philosophical Perspectives* 23: 79–110.

Fanari, Martina (2007). *Expressing Deontic Modality with Volitional Verbs: The Case of Sardinian*. PhD thesis, University of Pavia.

Finlay, Stephen (2009). "Oughts and Ends." *Philosophical Studies* 143: 315–40.

——— (2010). "What *Ought* Probably Means, and Why You Can't Detach It." *Synthese* 177: 67–89.

——— (2014). *Confusion of Tongues: A Theory of Normative Language*. New York: Oxford University Press.

von Fintel, Kai (1998). "The Presupposition of Subjunctive Conditionals." In *MIT Working Papers in Linguistics 25: The Interpretive Tract*, edited by Uli Sauerland and Orin Percus, 29–44. Cambridge: MIT Press.

von Fintel, Kai, and Sabine Iatridou (2008). "How to Say *Ought* in Foreign: The Composition of Weak Necessity Modals." In *Time and Modality*, edited by Jacqueline Guéron and Jacqueline Lecarme, 115–41. Springer.

von Fintel, Kai, and Sabine Iatridou (2015). "A Modest Proposal for the Meaning of Imperatives." MS, MIT.

Fleischman, Suzanne (1989). "Temporal Distance: A Basic Linguistic Metaphor." *Studies in Language* 13: 1–50.

van Fraassen, Bas C. (1973). "Values and the Heart's Command." *Journal of Philosophy* 70: 5–19.

Frank, Anette (1996). *Context Dependence in Modal Constructions*. PhD thesis, University of Stuttgart.

Gazdar, Gerald (1979). *Pragmatics: Implicatures, Presupposition, and Logical Form*. New York: Academic Press.

Gibbard, Allan (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge: Harvard University Press.

——— (2003). *Thinking How to Live*. Cambridge: Harvard University Press.

——— (2006). "Reply to Critics." *Philosophy and Phenomenological Research* 72: 729–44.

——— (2011). "How Much Realism? Evolved Thinkers and Normative Concepts." In *Oxford Studies in Metaethics*, vol. 6, edited by Russ Shafer-Landau, 33–51. New York: Oxford University Press.

——— (2012). *Meaning and Normativity*. New York: Oxford University Press.

Goddard, Cliff (2014). "*Have To*, *Have Got To*, and *Must*: NSM Analyses of English Modal Verbs of 'Necessity.'" In *Nonveridicality and Evaluation: Theoretical, Computational and Corpus Approaches*, edited by Maite Taboada and Radoslava Trnavac, 50–74. Leiden: Brill.

Grice, Paul (1989). *Studies in the Ways of Words*. Cambridge: Harvard University Press.

Groefsema, Marjolein (1995). "*Can*, *May*, *Must*, and *Should*: A Relevance Theoretic Account." *Journal of Linguistics* 31: 53–79.

Grosz, Patrick Georg (2012). *On the Grammar of Optative Constructions*. Amsterdam: Benjamins.

Hare, R. M. (1952). *The Language of Morals*. Oxford: Oxford University Press.

Harman, Gilbert (1973). "Review of Roger Wertheimer, *The Significance of Sense: Meaning, Modality, and Morality*." *The Philosophical Review* 82: 235–39.

——— (1975). "Reasons." *Crítica: Revista Hispanoamericana de Filosofía* 7: 3–17.

Heine, Bernd, Ulrike Claudi, and Friederike Hünnemeyer (1991). *Grammaticalization: A Conceptual Framework*. Chicago: University of Chicago Press.

Horn, Laurence R. (1972). *On the Semantic Properties of Logical Operators in English*. PhD thesis, University of California, Los Angeles.

——— (1984). "Toward a New Taxonomy for Pragmatic Inference: Q-Based and R-Based Implicature." In *Meaning, Form, and Use in Context: Linguistic Applications*, edited by Deborah Schiffrin, 11–42. Washington, DC: Georgetown University Press.

Huddleston, Rodney, and Geoffrey K. Pullum, eds. (2002). *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Iatridou, Sabine (2000). "The Grammatical Ingredients of Counterfactuality." *Linguistic Inquiry* 31: 231–70.

Iatridou, Sabine, and Hedde Zeijlstra (2013). "Negation, Polarity, and Deontic Modals." *Linguistic Inquiry* 44: 529–68.

Ippolito, Michela (2003). "Presuppositions and Implicatures in Counterfactuals." *Natural Language Semantics* 11: 145–86.

James, Deborah (1982). "Past Tense and the Hypothetical: A Cross-Linguistic Study." *Studies in Language* 6: 375–403.

Joos, Martin (1964). *The English Verb: Form and Meanings*. Madison: University of Wisconsin Press.

Kratzer, Angelika (1977). "What 'Must' and 'Can' Must and Can Mean." *Linguistics and Philosophy* 1: 337–55.

——— (1981). "The Notional Category of Modality." In *Words, Worlds, and Contexts: New Approaches in Word Semantics*, edited by Hans-Jürgen Eikmeyer and Hannes Rieser, 38–74. Berlin: de Gruyter.

——— (1991). "Modality/Conditionals." In *Semantics: An International Handbook of Contemporary Research*, edited by Arnim von Stechow and Dieter Wunderlich, 639–56. New York: de Gruyter.

——— (2012). *Modals and Conditionals: New and Revised Perspectives*. New York: Oxford University Press.

Krug, Manfred (2000). *Emerging English Modals: A Corpus-Based Study of Grammaticalization*. Berlin: Mouton de Gruyter.

Lakoff, Robin (1972a). "Language in Context." *Language* 48: 907–27.

——— (1972b). "The Pragmatics of Modality." In *Proceedings of the Chicago Linguistic Society 8*, 229–246. Chicago: CLS.

Lassiter, Daniel (2011). *Measurement and Modality: The Scalar Basis of Modal Semantics*. PhD thesis, New York University.

——— (2012). "Quantificational and Modal Interveners in Degree Constructions." In *Proceedings of SALT 22*, edited by Anca Chereches, 565–83. Ithaca, NY: CLC Publications.

Leech, Geoffrey (1971). *Meaning and the English Verb*. London: Longman.

——— (2003). "Modality on the Move: The English Modal Auxiliaries 1961–1992." In *Modality in Contemporary English*, edited by Roberta Facchinetti, Manfred Kung, and Frank Palmer, 223–40. New York: Mouton de Gruyter.

Leech, Geoffrey, Marianne Hundt, Christian Mair, and Nicholas Smith (2009). *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Lehmann, Christian (1995). *Thoughts on Grammaticalization*. München: Lincom.

Lewis, David (1973). *Counterfactuals*. Cambridge: Harvard University Press.

——— (1981). "Ordering Semantics and Premise Semantics for Counterfactuals." *Journal of Philosophical Logic* 10: 217–34.

Lyons, John (1977). *Semantics*, vol. 2. Cambridge: Cambridge University Press.

——— (1995). *Linguistic Semantics: An Introduction*. Cambridge: Cambridge University Press.

MacFarlane, John (2014). *Assessment Sensitivity: Relative Truth and Its Applications*. Oxford: Clarendon Press.

Makinson, David (1993). "Five Faces of Minimality." *Studia Logica* 52: 339–79.

Matthewson, Lisa (2010). "Cross-Linguistic Variation in Modality Systems: The Role of Mood." *Semantics and Pragmatics* 3: 1–74.

McGregor, William B., and Tamsin Wagner (2006). "The Semantics and Pragmatics of Irrealis Mood in Nyulnyulan Languages." *Oceanic Linguistics* 45: 339–79.

McNamara, Paul (1990). *The Deontic Quadecagon*. PhD thesis, University of Massachusetts.

Myhill, John (1995). "Change and Continuity in the Functions of the American English Modals." *Linguistics* 33: 157–211.

——— (1996). "The Development of the Strong Obligation System in American English." *American Speech* 71: 339–88.

——— (1997). "*Should* and *Ought*: The Rise of Individually-Oriented Modality in American English." *English Language and Linguistics* 1: 3–23.

Myhill, John, and Lauren A. Smith (1995). "The Discourse and Interactive Functions of Obligation Expressions." In *Modality in Grammar and Discourse*, edited by Joan Bybee and Suzanne Fleischman, 239–92. Amsterdam: Benjamins.

Narrog, Heiko, and Bernd Heine, eds. (2011). *The Oxford Handbook of Grammaticalization*. Oxford: Oxford University Press.

Ninan, Dilip (2005). "Two Puzzles about Deontic Necessity." In *New Work on Modality*, edited by Jon Gajewski, Valentine Hacquard, Bernard Nickel, and Seth Yalcin, 149–78. Cambridge: MITWPL.

Nuyts, Jan (2001). *Epistemic Modality, Language, and Conceptualization*. Amsterdam: John Benjamins.

Palmer, F. R. (1990). *Modality and the English Modals*, 2nd ed. New York: Longman.

——— (2001). *Mood and Modality*, 2nd ed. Cambridge: Cambridge University Press.

Perkins, Michael R. (1983). *Modal Expressions in English*. Norwood: Ablex.

Piaget, Jean (1962). *The Moral Judgment of the Child*. New York: Collier Books.

Portner, Paul (2007). Imperatives and Modals. *Natural Language Semantics* 15: 351–83.

——— (2009). *Modality*. Oxford: Oxford University Press.

Portner, Paul, and Aynat Rubinstein (2012). "Mood and Contextual Commitment." In *Proceedings of SALT 22*, edited by Anca Chereches, 461–87. Ithaca, NY: CLC Publications.

——— (2016). "Extreme and Non-Extreme Deontic Modals." In *Deontic Modality*, edited by Nate Charlow and Matthew Chrisman, 256–82. New York: Oxford University Press.

Ridge, Michael (2014). *Impassioned Belief*. Oxford: Oxford University Press.

Rubinstein, Aynat (2012). *Roots of Modality*. PhD thesis, University of Massachusetts Amherst.

——— (2014). "On Necessity and Comparison." *Pacific Philosophical Quarterly* 95: 512–54.

Schlenker, Philippe (2005). "The Lazy Frenchman's Approach to the Subjunctive (Speculations on Reference to Worlds and Semantic Defaults in the Analysis of Mood)." In *Romance Languages and Linguistic Theory 2003*, edited by Twan Geerts, Ivo van Ginneken, and Haike Jacobs, 269–309. Amsterdam: John Benjamins.

Schueler, David (2011). *The Syntax and Semantics of Implicit Conditionals: Filling in the Antecedent*. PhD thesis, University of California, Los Angeles.

Silk, Alex (2012). "Modality, Weights, and Inconsistent Premise Sets." In *Proceedings of SALT 22*, edited by Anca Chereches, 43–64. Ithaca, NY: CLC Publications.

——— (2013). *What Normative Terms Mean and Why It Matters for Ethical Theory*. PhD thesis, University of Michigan, Ann Arbor.

——— (2015a). "How to Be an Ethical Expressivist." *Philosophy and Phenomenological Research* 91: 47–81.

——— (2015b). "What Normative Terms Mean and Why It Matters for Ethical Theory." In *Oxford Studies in Normative Ethics*, vol. 5, edited by Mark Timmons, 296–325. Oxford: Oxford University Press.

——— (2016a). *Discourse Contextualism: A Framework for Contextualist Semantics and Pragmatics*. Oxford: Oxford University Press.

——— (2016b). "Update Semantics for Weak Necessity Modals." In *Deontic Logic and Normative Systems*, edited by Olivier Roy, Allard Tamminga, and Malte Willer, 237–55. Milton Keynes: College Publications.

——— (2017). "Modality, Weights, and Inconsistent Premise Sets." *Journal of Semantics* 34: 683–707.

——— (2018). "Commitment and States of Mind with Mood and Modality." *Natural Language Semantics* 26: 125–66.

Sloman, Aaron (1970). "'Ought' and 'Better'." *Mind* 79: 385–94.

Smith, Nicholas (2003). "Changes in Modals and Semi-Modals of Strong Obligation and Epistemic Necessity in Recent British English." In *Modality in Contemporary English*, edited by Roberta Facchinetti, Manfred Kung, and Frank Palmer, 241–66. New York: Mouton de Gruyter.

Stalnaker, Robert (1975). "Indicative Conditionals." In Stalnaker (1999), 63–77.

——— (1978). "Assertion." In Stalnaker (1999), 78–95.

——— (1999). *Context and Content: Essays on Intentionality in Speech and Thought*. Oxford: Oxford University Press.

Starr, William (2010). *Conditionals, Meaning, and Mood*. PhD thesis, Rutgers University.

Stephenson, Tamina (2007). "Judge Dependence, Epistemic Modals, and Predicates of Personal Taste." *Linguistics and Philosophy* 30: 487–525.

Stevenson, Charles L. (1937). "The Emotive Meaning of Ethical Terms." *Mind* 46: 14–31.

Swanson, Eric (2008). "Modality in Language." *Philosophy Compass* 3: 1193–207.

——— (2011). "On the Treatment of Incomparability in Ordering Semantics and Premise Semantics." *Journal of Philosophical Logic* 40: 693–713.

——— (2012). "Imperative Force in the English Modal System." Handout, Philosophy-Linguistics Workshop, University of Michigan.

——— (2016a). "The Application of Constraint Semantics to the Language of Subjective Uncertainty." *Journal of Philosophical Logic* 45: 121–46.

——— (2016b). "Metaethics without Semantics." MS, University of Michigan.

Sweetser, Eve (1990). *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. Cambridge: Cambridge University Press.

Tagliamonte, Sal (2004). "*Have To*, *Gotta*, *Must*: Grammaticalisation, Variation and Specialization in English Deontic Modality." In *Corpus Approaches to Grammaticalization in English*, edited by Hans Lindquist and Christian Mair, 33–55. Philadelphia: John Benjamins.

Thomson, Judith Jarvis (2008). *Normativity*. Chicago: Open Court.

Traugott, Elizabeth Closs (1995). "Subjectification in Grammaticalisation." In *Subjectivity and Subjectivisation. Linguistic Perspectives*, edited by Dieter Stein and Susan Wright, 31–54. Cambridge: Cambridge University Press.

Van Linden, An (2012). *Modal Adjectives: English Deontic and Evaluative Constructions in Synchrony and Diachrony*. Berlin: Mouton de Gruyter.

Van Linden, An, and Jean-Christophe Verstraete (2008). "The Nature and Origins of Counterfactuality in Simple Clauses: Cross-Linguistic Evidence." *Journal of Pragmatics* 40: 1865–95.

——— (2011). "Revisiting Deontic Modality and Related Categories: A Conceptual Map Based on the Study of English Modal Adjectives." *Journal of Pragmatics* 43: 150–63.

Veltman, Frank (1976). "Prejudices, Presuppositions, and the Theory of Conditionals." In *Amsterdam Papers in Formal Grammar*, vol. 1, edited by Jeroen Groenendijk and Martin Stokhof, 248–81. Central Interfaculteit, University of Amsterdam.

Verstraete, Jean-Christophe (2001). "Subjective and Objective Modality: Interpersonal and Ideational Functions in the English Modal Auxiliary System." *Journal of Pragmatics* 33: 1505–28.

——— (2005a). "Scalar Quantity Implicatures and the Interpretation of Modality: Problems in the Deontic Domain." *Journal of Pragmatics* 37: 1401–18.

——— (2005b). "The Semantics and Pragmatics of Composite Mood Marking: The Non-Pama-Nyungan Languages of Northern Australia." *Linguistic Typology* 9: 223–68.

——— (2006). "The Nature of Irreality in the Past Domain: Evidence from Past Intentional Constructions in Australian Languages." *Australian Journal of Linguistics* 26: 59–79.

——— (2007). *Rethinking the Coordinate-Subordinate Dichotomy: Interpersonal Grammar and the Analysis of Adverbial Clauses in English*. Berlin: Mouton de Gruyter.

Wedgwood, Ralph (2007). *The Nature of Normativity*. Oxford: Oxford University Press.

Werner, Tom (2003). *Deducing the Future and Distinguishing the Past: Temporal Interpretation in Modal Sentences in English*. PhD thesis, Rutgers University.

Wertheimer, Roger (1972). *The Significance of Sense: Meaning, Modality, and Morality*. Ithaca, NY: Cornell University Press.

Wheeler, Samuel C. III (2013). *Neo-Davidsonian Metaphysics: From the True to the Good*. New York: Routledge.

Williams, Bernard (1981). "Practical Necessity." In *Moral Luck*, 124–31. Cambridge: Cambridge University Press.

Woisetschlaeger, Erich Friedrich (1977). *A Semantic Theory of the English Auxiliary System*. PhD thesis, MIT.

von Wright, Georg Henrik (1963). *Norm and Action: A Logical Inquiry*. London: Routledge and Kegan Paul.

Yalcin, Seth (2012). "Bayesian Expressivism." *Proceedings of the Aristotelian Society* 112: 123–60.

——— (2016). "Modalities of Normality." In *Deontic Modality*, edited by Nate Charlow and Matthew Chrisman, 230–55. New York: Oxford University Press.

11

HOW TO OUTFOX SLY PETE:

A Picture of the Pragmatics of Indicatives[1]

Caleb Perl

Expressivists hold that the use of a sentence expresses the speaker's mental state. And they deny that it expresses a representational condition on what the world is like. Expressivism about indicative conditionals, for example, is the view that the use of an indicative just expresses the speaker's conditional credence.

Indicative conditionals are particularly important for expressivists, because they provide some of the most powerful linguistic evidence in favor of an expressivist approach, and against more traditional alternatives. And this evidence is important, because there are important problems about expressivism in general. The expressivist struggles to explain *embedded* judgments, like embeddings of the vocabulary under adverbs like "probably"—the classic Frege-Geach problem.[2]

Expressivist accounts of indicatives are important for assessing the force of those challenges. If the best account of indicatives is expressivist, we have a powerful *license for optimism* in favor of the expressivist approach in general, a license for optimism that the approach

1 For comments on an earlier version of this paper, I'm grateful to audiences at the 2011 Society for Exact Philosophy and the 2013 Pacific APA, and to Andrew Bacon, Paddy Blanchette, Aaron Bronfman, Marian David, Janice Dowell, Billy Dunaway, James Higginbotham, Daniel Immerman, Ben Lennertz, Gillman Payette, Daniel Rothschild, Barry Schein, Johannes Schmitt, Jeff Speaks, Gabriel Uzquiano, and Jon Wright. I'm especially grateful to my commentators at the 2013 Pacific, Alexi Burgess and John MacFarlane, and to Mark Schroeder and Scott Soames for many rounds of extremely helpful comments and conversations.

2 Important critical discussions include those by Peter Geach (1965), Bob Hale (1993), Mark Schroeder (2008, 2015b), John Searle (1962), and Nicolas Unwin (1999, 2001).

can somehow be made to work. And expressivism has very interesting upshots in several domains, particularly in metaethics. A license for optimism on behalf of the expressivist about indicatives is also a license for optimism for the metaethical expressivist.[3]

I attempt a piece of philosophical jujutsu. I leverage a phenomenon that appears to be evidence *for* expressivism about indicatives *against* it. That is, I try to show that that phenomenon eliminates *all* expressivist accounts of indicatives. If I'm right, indicatives don't give the expressivist any license for optimism. I then introduce my own constructive account of the phenomenon. My account offers a systematic explanation of the data that eliminates expressivist accounts. This account advances our understanding of what an adequate account of indicatives would have to look like.

The Gibbard Phenomenon

This section introduces the phenomenon that has traditionally taken to be evidence for expressivism, and against traditional accounts.

I'll take the semantics from Angelika Kratzer (1986) as my example of a traditional account.[4] Her account starts with her account of modals like *must*, where those modals contribute quantifiers over the highest ranked worlds in a contextually supplied set. Kratzer proposes that conditionals restrict the quantifier that ranges over the contextually supplied set. If ⌜must(p)⌝ quantifies over all the highest ranked worlds in that set, ⌜if q, must(p)⌝ quantifies over all the highest-ranked worlds in the set.

On Kratzer's proposal, indicatives express a representational condition that the world needs to satisfy, about what's true of a particular set of worlds. Allan Gibbard has given a challenge to *any* account that takes indicatives to express some such representational condition. His original challenge was about a character called Sly Pete. (That's why this paper is about outfoxing Sly Pete—it's about outfoxing the puzzle he articulated.[5]) But his original case had some confusing features. Jonathan Bennett gave a better example:

---

3 Simon Blackburn (2016) has recently made a spirited version of this kind of point. Allan Gibbard describes some differences between our use of indicatives and our moral discourse, while still favoring an expressivist account of both (Gibbard 2012, 70–72). Mark Schroeder (2015a) has a detailed discussion of the similarities and differences between the problems for the metaethical expressivist and for the expressivist about indicatives.

4 I set aside Robert Stalnaker's (1975) similar account. His account is similar enough to Kratzer's for the following discussion to smoothly generalize. My presentation of Kratzer suppresses lots of important details.

5 Here's Gibbard original case:

> Sly Pete and Mr. Stone are playing poker on a Mississippi riverboat. It is now up to Pete to call or fold. My henchman Zack sees Stone's hand, which is quite good, and signals its content to Pete. My henchman Jack sees both hands, and sees that Pete's hand is rather low, so that Stone's is the winning hand. At this point, the room is cleared. A few minutes later, Zack slips me a note which says? If Pete called, he won,? and Jack slips me a note which says? If Pete called, he lost.? I know that these notes both come from my trusted henchmen, but do not know which of them sent which note. I conclude that Pete folded. (Gibbard 1981, 231)

Top Gate holds back water in a lake behind a dam; a channel running down from it splits into two distributaries, one (blockable by East Gate) running eastwards and the other (blockable by West Gate) running westwards.[6] (Bennett 2003, 85) Crucially, it's not possible for all three gates to be open at the same time, given the construction of the system. Wesla saw that West Gate was open, and Esther saw that the East Gate was open. Wesla tells us (W), and Esther tells us (E): (W) If Top Gate opened, all the water ran westwards.

(E) If Top Gate opened, all the water ran eastwards.[7]

Gibbard thinks that expressivists can correctly explain this kind of case. For the expressivist, "indicative conditionals are . . . to be understood through their conditions of acceptance or assertability, and where *a* and *b* are propositions, one accepts the indicative conditional 'If *a*, then *b*' iff one's conditional credence in *b* given *a* is sufficiently high" (Gibbard 1981, 212).[8] For him, then, both utterances are appropriate because the two agents have the right conditional credences. This account doesn't identify a representational condition that the world needs to satisfy for the two sentences to be appropriate. It instead identifies a mental state that makes the utterances appropriate. That's why it's an expressivist account.

If your semantics *does* take (W) and (E) to express representational conditions that the world needs to satisfy, I'll say that you're offering a *traditionalist* account. Gibbard thinks that traditionalists struggle to capture this kind of case. The symmetry between Wesla and Esther forces the traditionalist to take the propositions (W) and (E) express to have the same truth-value.[9] And traditionalists struggle to explain why they would.

One option for the traditionalist is to take the context to make some substantive contribution: (W) and (E) express propositions that differ more than just in the direction the water goes. For example, maybe Wesla's use of (W) expresses a proposition about the information that Wesla has. Given this traditionalist approach, the hearer's ability to learn from what Wesla said can't always rely on her learning the propositions expressed. (The hearer won't always know what information Wesla has.) These traditionalists must offer a metalinguistic account of the conversational dynamics. And when they do, they appeal to the mental states that the expressivist takes to be semantically expressed.[10] Then the traditionalist's semantics starts to look like an idle wheel that doesn't do genuine explanatory work.

---

6 To make it plausible that all three gates can't be open at the same time, Bennett includes much more detail. I eliminate the extra detail, because it makes an already complicated discussion even more complicated.

7 Kratzer supposes that bare conditionals like (W) and (E)—conditionals without an explicit modal like *must*—have an unpronounced modal at logical form that the antecedent restricts.

8 For broadly similar views about the language of subjective uncertainty, see Adams (1975), Bennett (2003), Edgington (1995), Moss (2015, 2018) Schneider (2010), Swanson (2006), and Yalcin (2007).

9 To deny that they have the same truth-value, the traditionalist must find some asymmetry between the two that makes one false and the other true. And Bennett can just stipulate away any asymmetry.

10 Angelika Kratzer's account of this puzzle is one example (Kratzer 2012, 121).

Another kind of traditionalist might deny that the context makes any substantive contribution to what (W) and (E) express. For this kind of traditionalist, the propositions expressed differ only in the direction the water goes. This sort of traditionalist seems forced to accept a material conditional semantics for indicatives. The propositions expressed by (W) and (E) can have the same truth-value only if their consequents have the same truth-value at that nearest world. But those consequents can't have the same truth-value at that world. The water can't all flow both east and west.

In general, Gibbard's puzzle pushes traditionalists to either use the resources the expressivist uses or to give a material conditional semantics for the indicative. And there are powerful reasons to reject a material conditional semantics.[11] So the traditionalist seems forced to use the resources that the expressivist uses—which seems like good inductive evidence that the expressivist captures what's really going on with indicatives.[12]

Expressivism Can't Work

This section attempts philosophical jujutsu. It argues that, *contra* initial appearances, the Gibbard phenomenon decisively eliminates expressivist accounts of indicatives.

What Do Wesla and Esther Know?

I claim that (1) is true.

> (1) Wesla was reasonable in taking generous bets that either the water ran westwards or Top Gate didn't open, because she knows that if Top Gate opened, all the water ran westwards. And Esther was reasonable in taking generous bets that either the water ran eastwards or Top Gate didn't, because she knows that if Top Gate opened, all the water ran eastwards.

(1) is definitely appropriate—and its appropriateness is powerful evidence that it's true.

There are also more theoretical arguments that (1) is true. It has to be true as long as conditional proof can extend knowledge. Wesla can know how the gates work, and can know that West Gate is open. If she supposes that Top Gate opened, she can then prove that all the water ran westward. So she'd be in a position to know her conditional, if conditional proof can extend knowledge. Ditto for Esther. So (1) has to be true if conditional proof can extend knowledge. And we should allow that conditional proof can extend knowledge. Supposing otherwise leads to implausibly wide-ranging skepticism about our knowledge of indicatives.

Another argument that (1) is true is from the knowledge norm of assertion. Take Williamson's formulation that "one must: assert *p* only if one knows *p*" (Williamson 2000, 243).

---

11 See, for example, pp. 90ff. of Gibbard (2012).

12 This kind of reasoning is particularly clear in Gibbard (1981) and Bennett (2003).

Everybody agrees that Wesla and Esther both make appropriate assertions. Indeed, that agreement is why expressivists take the Gibbard phenomenon to be evidence against more traditional accounts. (No expressivist who feels the pull of the argument in 1 for expressivism can disagree! In feeling the pull of that argument, you're taking both utterances to be appropriate.) But given the knowledge norm, the assertions are appropriate only if they know what they assert. Since their assertions *are* appropriate, they must know what they assert.

Against Expressivism

Expressivists cannot explain why (1) is appropriately assertable.

Let 'ϵ' designate the expressivist's threshold for the agent's credence being sufficiently high. The expressivist must hold that $\epsilon > .5$. There are otherwise plenty of counterexamples to expressivism.

(2) If I flip this fair coin, it will land heads. And if I flip this (same) fair coin, it will land tails.

(My conditional credences are both .5.) But the unassertability of (2) is a datum. So far, so unproblematic, as long as $\epsilon > .5$.

The expressivist also needs to give a theory of factive verbs like *know*. (She is giving an account of the use of indicative conditionals, which can be embedded under factive verbs—so she needs to predict when we'll attribute knowledge of an indicative.) I'll take the expressivist to hold that attributions of knowledge commit the speaker to the complement—that, as Sarah Moss puts the idea, "the inference from 'S knows that p' to p is valid" (Moss 2013, 12). Follow Moss in calling this idea Factivity$_2$. Moss considers this idea in the course of considering expressivist accounts of the language of subjective uncertainty in general—of indicative conditionals, but also of adverbs like *probably*. As Moss rightly emphasizes, the expressivist needs some account very much like this. Moss notes that (3) is infelicitous.

(3) * John knows that it is probably raining, and Bob knows that it probably isn't.

Factivity$_2$ gives a natural explanation of this infelicity. Given Factivity$_2$, there is a valid inference from (3) to *it probably is raining and it probably isn't raining*. And the semantic value of that sentence imposes incompatible constraints on any mental state. *That* is the reason that (3) is so odd. We see the same pattern with indicatives. We won't be willing to take any two people to know the two coin-tossing indicatives in (2), for the same reason. But this explanation of (3)'s infelicity crucially assumes Factivity$_2$.[13]

13 There is another reason to accept Factivity2. We need to somehow distinguish nonfactive verbs like *believe* from factive verbs like *recognize* or *realize*. ⌜Jane thinks that if p, q⌝ differs systematically from ⌜Jane recognizes that if p, q⌝. And it seems like the second sentence differs systematically in just the way that (a

Our two observations force the expressivist to predict that (1) is unassertable. Given FACTIVITY$_2$, (1) is assertable only if both embedded indicatives are assertable. But the expressivist can't allow that both indicatives are assertable. Their consequents can't both hold—if all the water flows west, all the water can't run east. So the sum of the two conditional credences has to be 1 or lower: Pr(all the water runs east | Top Gate opened) + Pr(all the water runs west | Top Gate opened) ≤ 1. So at least one of the conditional credences has to be .5 or less. Since ϵ (the constraint on the assertability of conditionals) has to be greater than .5, the expressivist has to hold that one of the indicatives isn't assertable. So, she mistakenly predicts that (1) is unassertable.[14]

In fact, the situation for the expressivist is even more dire. It's possible to *believe* that Wesla and Esther both know their conditionals, without being *certain* that Top Gate didn't

---

generalization of) FACTIVITY2 captures: you can validly infer the semantic value of the complement from the semantic value of the second sentence, but not from the semantic value of the first. The expressivist who rejects FACTIVITY2 owes us her own constructive account of (3), and of the difference between *believe* and *recognize*. It's reasonable to assume FACTIVITY2, because it's so hard to see another constructive account of this difference.

14 Now you might wonder whether the contrast between Bennett's case and examples with *probably* are as sharp as I've been suggesting. Here are the bare examples, with the *probably* example slightly changed.

(1) Wesla knows that if Top Gate opened, all the water ran westwards, and Esther knows that if Top Gate opened, all the water ran eastwards.

(3′) * Wesla knows that all the water probably ran westwards, and Esther knows that all the water probably ran eastwards.

I've suggested that (1) but not (3′) is perfectly assertable in the right context. Now it can take some work to make (1) sound appropriate—to hear it, you have to keep firmly in mind that only two gates can open at once. And you might think that enough work can make (3′) appropriate, too. Suppose that Wesla and Esther each have evidence that makes it 80 percent likely that Top Gate opened. Then it seems like (3′) should be appropriate if (1) is appropriate; they're just drawing the same kinds of inferences. (In particular, we can connect (3′) with rational betting behavior in the way I did with (1). We can say that Wesla is rational in betting that all the water went westwards, because she knows that all the water probably went westwards.)

There still is a genuine contrast between (1) and (3′). Factive verbs like *knows* can be used without committing the speaker to the complement; we can say things like *what everyone knows about Nixon isn't true*. (Robert Stalnaker (1974), Scott Soames (2009), and Dorit Abusch (2010) all have particularly illuminating discussions of this possibility.) We should understand appropriate uses of (3′) as involving exactly this possibility: those uses are appropriate because the speaker is not committed to the truth of both complements. As soon as this possibility is salient, though, we should ask whether (1) admits of a similar treatment—that it's appropriate, but only because the speaker is not committed to both complements. (1) may have such uses. Crucially, though, (1) also has uses where this treatment is inappropriate. We can see that those uses exist by noting how much we can *infer* from certain uses of (1). In particular, we can infer from certain uses of (1) that the speaker believes that Top Gate didn't open. (Gibbard made a closely related point in his original discussion of these cases (Gibbard 1981, 231).) *That* inference is intelligible only if the use of (1) does commit the speaker to both complements. By contrast, the speaker can't be committed to both of (3′)'s complements; she can't think that it's probable that it'll rain and probable that it won't rain. In other words, we should allow that (1) but not (3′) can be true even on the use where the speaker is committed to both complements. That possibility is what poses the basic challenge to the expressivist.

open. Suppose that Jane is reliable 95 percent of the time. I don't know anything about what's going on with the gate system, other than that only two gates can open at once. Jane says:

(1) Wesla knows that if Top Gate opened, all the water ran westwards, and Esther knows that if Top Gate opened, all the water ran eastwards.

I come to believe (1) only because Jane told me. Crucially, though, I still have very low but nonzero credence that Top Gate opened. My only evidence about what happened is from Jane, and she's only reliable 95 percent of the time.[15] So even though I believe that (1) is true, it's not because my credence that Top Gate opened was 0. It's hard to see how the expressivist can capture this example.[16]

Similar problems also arise for some other accounts of indicatives. For example, Frank Jackson (1987) holds that indicatives semantically express only the corresponding material conditional and conventionally implicate a high conditional credence in the consequent given the antecedent. And this further claim is essential to his proposal. It's what allows him to avoid the "paradoxes of material implication," where *if p, q* is true if p is false. According to Jackson, the corresponding indicative conditionals are true, but a speaker would never assert them unless she also had the right conditional credences.

This further (but essential!) claim prevents Jackson from explaining the data that interest us. Jackson faces a choice: do factive embeddings of indicatives also implicate high conditional credences? If they don't, he faces immediate counterexamples. The assertability conditions of factive embeddings would then be the same as the truth-conditions of factive embeddings; *I know that this coin will land heads if flipped, and I know that this coin will land tails if flipped* will express a truth whenever I know the coin won't be flipped.[17] Given the other choice, where factive embeddings of indicatives *do* conventionally implicate high

15 See Hawthorne et al. (2016) for some arguments that belief is possible in this sort of case.

16 Some expressivists want to deny that (W) and (E) have truth values. (Gibbard reaffirms his commitment to this denial in *Meaning and Normativity* (Gibbard 2012, 70).) Such expressivists might deny that (1) has a truth-value, on the grounds that neither (W) nor (E) have a truth-value. So they might be unconcerned to explain how (1) is assertable.

But the expressivist is on much shakier ground in denying that (1) is assertable. Once we know all the facts of the case, we're unwilling to use (W), and we're unwilling to use (E). That's why it's initially plausible for the expressivist to deny that they have truth-values: her denial captures one aspect of our competence with indicatives. But the situation is quite different with (1). Another aspect of our competence with indicatives is that we are willing to use (1). (I supplied some other arguments for (1)'s felicity, as a way of bolstering the claim that our willingness to accept it is part of our competence with indicatives.) And our willingness to use (1) should make us doubt that (W) and (E) do lack truth-values in this context. We should doubt that the missing truth-value explanation is the right diagnosis of why we're unwilling to use the indicatives. We should look for an account that predicts that there's else wrong with using (W) or (E) when we know all the facts.

17 If the coin won't be flipped, ⌜the coin is flipped → p⌝ is true for all p. And on the present choice, those factive embeddings do *not* conventionally implicate a high conditional credence.

conditional credences, the 2.2 problem arises. Utterances of the factive embedding in (1) will conventionally implicate a pair of credences that are jointly inconsistent: a high credence in q conditional on p and a high credence in ¬ q conditional on p. So it's hard to see how Jackson's approach can be made to work, either.

Can Anyone Capture This Data?

I've just issued a challenge to expressivist accounts of indicatives, and indeed to any account that links indicatives with conditional credences. An expressivist might respond aggressively. They might claim that *nobody* has an adequate account of data that I've just described. In a slogan: "my problem, but also your problem: so not my problem!" This aggressive response has a lot going for it. It *is* a mistake to dismiss a theory for not explaining data that nobody explains.

And in fact the Gibbard phenomenon is even more complicated than we've seen so far. We can change Bennett's case so that it's impossible for Wesla and Esther to both know their conditionals. Just suppose that all three gates can be up at once. There are four cases to consider. In the case where all gates were open, neither knows her conditional; if Top Gate opened, some of the water ran each way. The same is true if both gates were closed: then if Top Gate opened, the water didn't run either east or west. Now consider the case where the East Gate was shut and the West open. Then Esther can't know her conditional; if the Top Gate opened, all the water ran *west*. Something similar is true if the East Gate is open and the West shut—Wesla then can't know her conditional. So even before we know anything about the positions of the gates in this second context, we know that they both can't know.[18] This range of data is really puzzling! It's not clear how we should explain it. So maybe it's a mistake to reject expressivist accounts for not capturing it.

In fact, this response on behalf of the expressivist is even more compelling when we see how hard it is for traditionalists to capture the full range of data. I'll give two examples. J. R. G. Williams (2008) has suggested that Bennett's indicatives are vacuously true, because indicatives are vacuously true whenever what's taken for granted entails that the indicative's antecedent is false.[19]

To evaluate the idea, we need some way of modeling our ability to update from what Wesla and Esther have said. As a first pass, I will explore what happens if the update proceeds as if Wesla and Esther had asserted propositions about the body of information shared

18 There may be ways of recreating the Gibbard phenomenon: other stipulations that allow Wesla and Esther's conditionals to be compatible again. That fact doesn't change the present point. The point is that there are some cases where the indicatives are intuitively *incompatible*, and I can defend this point by stipulating away whatever features you try to add.

19 His account builds from Stalnaker's account of indicatives, but I simplify the discussion by pretending that Williams is discussing a version of Kratzer's account, rather than Stalnaker's account, for continuity with the rest of the paper.

between me the hearer, Wesla, and Esther. I'll model the effect of updating with a set of possible worlds—the worlds compatible with what is taken for granted in the conversation. For example,  will be the set of worlds where all three gates are closed. So the following context represents what's common ground in Bennett's original case, where only two gates can open at once, before we hear from anyone.

$$\text{CInitial} = \{(\text{  }), (\text{}), (\text{}), (\text{}), (\text{}), (\text{}), (\text{})\}$$

The central problem for Williams is to explain how we use indicatives to update from this context set. The natural idea would be that the most similar worlds to some arbitrary world *w* are the worlds compatible with what's accepted where the lower gates are in as close to the same positions as they are in *w* as possible. Updating with Wesla's uttering (W) then produces this context set.

$$\text{CAfter(W)} = \{(\text{}), (\text{}), (\text{})\}$$

(Importantly, narrowing down the context set doesn't require certainty that the worlds eliminated don't obtain. In conversation, we often accept that some possibilities don't obtain because we think it's sufficiently unlikely that they obtain; for discussion, see Stalnaker (2002).)

Now suppose that Esther assertively utters (E). The proposition expressed is *false* at every world in the context set. In particular, it's false even at  worlds. There are still worlds compatible with what is accepted where Top Gate opened:  worlds. So, given the natural idea from the last paragraph, (E) cannot be vacuously true, because there are worlds in the context set where Top Gate opened. In that case, (E) would be false, because all the worlds where Top Gate does open are worlds where all the water went westward.[20]

It is tempting to try to revisit the natural idea just sketched. But the natural idea is itself highly attractive, because it's a highly principled way of delivering exactly the right result about the other case, where all three gates can open at once. To represent this possibility, we extend the initial context set to include worlds where all three gates open at once.

$$\{(\text{  }), (\text{}), (\text{}), (\text{}), (\text{}), (\text{}), (\text{}), (\text{})\}$$

When Wesla assertively utters (W), we eliminate all the worlds we eliminated in the first kind of context. But we also eliminate  and worlds, because some water goes both ways at those worlds if Top Gate opened.

20 It's standard to take indicatives to carry a presupposition—that an assertive utterance of "if p, q" in a context C presupposes that p is compatible with what's taken for granted'. Karttunen and Peters offer an early statement of this kind of view (Karttunen and Peters 1979, 10); Kai von Fintel (1996) gives a more modern treatment. Williams can't appeal to this presupposition to help himself out of the present problem. The presupposition is satisfied at every point before we update with the two indicatives.

$C_{After(W)}$= {(), ()}

It's again true that (E) is false at every world in the context set. But that's the right result, here: the indicatives are intuitively *in*compatible. There is a tension between getting the original case right and getting the second case right. Mere appeal to vacuous truth by itself isn't enough.

Here's another way to make this point. Williams suggests that "the conversational effects of asserting an indicative conditional with the truth-conditions suggested earlier are the same as those described for the material conditional" (Williams 2008, 215n10). This suggestion correctly predicts that (W) and (E) are compatible in the original context, but it mistakenly predicts that they are compatible in the modified context, too.

ORIGINAL CONTEXT, ONLY TWO GATES CAN OPEN

{(), (), (), (), (), (), ()}

After Asserting (W): { (), (), (), (), ()}

After Asserting (E): { (), (), (), ()}

MODIFIED CONTEXT, ALL THREE GATES CAN OPEN AT ONCE

{(), (), (), (), (), (), (), ()}

After Asserting (W): { (), (), (), (), ()}

After Asserting (E): { (), (), (), () }

Williams wrongly predicts that (W) and (E) are compatible in the modified context.

More generally, we want to explain how (W) and (E) can be *compatible* in the first context but *incompatible* in the second. Williams' appeal to vacuous truth doesn't do that. So the expressivist can continue to insist that Williams hasn't given us an adequate account of the Gibbard phenomenon. As a result, she can continue insisting that her own difficulties with the data aren't evidence against her. Let's now turn to another traditionalist idea. The idea is that the updating runs through metalinguistic reasoning about the assertability conditions of the two sentences. The two conditionals are assertable just in case the evidence available to Wesla and Esther warrants each of their assertions. And the person who hears their assertions can reason about the sort of evidence that they have, presumably that both gates were open. As a result, the hearer can infer that both gates were open, and that Top Gate didn't open. Angelika Kratzer has suggested just that, about this puzzle.[21]

This suggestion won't work. No inference about Wesla and Esther's evidence can explain the difference between the two contexts. Consider cases where Wesla and Esther's only evidence is testimonial:

21 She proposes that the Gibbard phenomenon is a case where "discourse participants can extract the information they are after under the presumption that assertability conditions are satisfied" (Kratzer 2012, 121).

Original Context: only two gates can open at once;

- Jane has told Wesla (W)
- James has told Esther (E)

Modified Context: all three gates can open at once;

- Jane has told Wesla (W)
- James has told Esther (E)

In this case, there isn't any difference in Wesla or Esther's evidence between the original and modified contexts: in both, the only evidence is testimonial. But since you and I aren't aware of differences in Wesla's evidence, inferences about her evidence can't explain why we think (W) and (E) compatible in one case and incompatible in the other.[22]

Similar problems seem to arise for other contextualist suggestions. Another contextualist might appeal to Stalnakerian diagonalization, suggesting that (W) and (E) communicate different diagonalized propositions in the original context than in the modified context. The diagonalized propositions need to be different between the two contexts, because the propositions believed are incompatible in the modified context but compatible in the original one. But we need a constructive account of what those propositions are, and how they come to be communicated.

Up to this point, the expressivist can reasonably remain unperturbed. Nobody seems to have an adequate account of this data.

A Presuppositional Proposal

I now introduce my own constructive account of this data. This constructive account answers the challenge from the last section: it explains how to capture all the facts about the Gibbard phenomenon.

I'll develop this explanation in a contextualist framework, to show that the contextualist does have a perfectly adequate account of the Gibbard phenomenon. But my explanation is in principle available to noncontextualists as well. Importantly, though, it's not available to

22 Now Kratzer may insist that the two bits of evidence are different in the two cases: in the original case, the bit of evidence concerns Jane's information about the original setup; in the modified case, it concerns Jane's information about the modified setup. The indirect speech reports express different propositions, and so constitute different evidence. That suggestion doesn't change the point I'm making. You and I aren't in a position to know the difference between Jane's information—her evidence could be testimonial, too! So, this Kratzerian suggestion entails that the only evidence that any of us have in this case is metalinguistic evidence: that Jane's evidence warrants Jane's assenting to the sentence *all the water ran westward if Top Gate opened*. But that bit of metalinguistic evidence is exactly the same in the two cases.

expressivists—or at least to expressivists who take conditional credences to be expressed. If it's right, that sort of expressivism about indicatives has to be wrong.

The Basic Idea

I propose that utterances of indicatives carry more *presuppositions* in the original context than they do in the modified case. In particular, I propose that an utterance of (W) in the original context carries a new conditional presupposition about the East Gate—something similar for (E).

(W) If Top Gate opened, all the water ran westwards.
presupposes that
Top Gate opened → the *East* Gate was closed

(E) If Top Gate opened, all the water ran eastwards.
presupposes that
Top Gate opened → the *West* Gate was closed

I'll present my idea in two parts. The first part will show that we can explain the difference between the original context and the modified context *if* utterances of (W) and (E) carry these presuppositions in the original context but not in the modified one. The second part will show that the utterances do plausibly carry different presuppositions in the two contexts.

Let's start with the difference that these presuppositions would make. In the original context, where only two gates can open at once, the following context set represents ignorance of the positions of the gates.

$\{$(w), (w), (w), (w), (w), (w), (w)$\}$

Suppose that Wesla assertively utters (W). The first thing we do is accommodate the presupposition of the utterance, which means eliminating (w) -worlds.[23]

$C_{\text{AfterAccom}} = \{$(w), (w), (w), (w), (w), (w)$\}$

Then we eliminate worlds where the proposition (W) asserts is false.[24]

$C_{\text{After(W)}} = \{$(w), (w), (w), (w)$\}$

23 Accommodation is in fact controversial. (For helpful discussion, see Mandy Simons et al. (2011), Kai von Fintel (2008), and Christopher Gauker (2008).) My explanation does not crucially depend on presuppositions being accommodated. What matters essentially is that we update with the presupposed content before we update with the asserted content, and alternative frameworks vindicate that essential point.

24 I'm again assuming that the position of the lower gates is what matters for similarity—that the most similar worlds to *w* where Top Gate opened are worlds where the lower gates are in the same position they are in *w*.

Now suppose that Wesla assertively utters us (E). The first thing we do is accommodate the presupposition of the utterance, which means eliminating -  worlds.

$$\text{CAfterAccom} = \{(\text{  }), (\text{}), (\text{})\}$$

Updating with the proposition that (E) asserts then has no effect, because that proposition is vacuously true. That's why we hear these indicatives as intuitively compatible.[25]

It's worth being clear about the reason why this proposal works for contextualists. Contextualists take indicatives to communicate propositions about bodies of information. My idea is that the presuppositions are shifting what propositions we hearers interpret the utterance as communicating. To illustrate this effect, think of the body of information that's been updated with Wesla's information. We can take Esther's utterance to communicate a proposition about what's common ground for us, or we can take it to communicate something about what's common ground, *plus* the presupposition of her utterance. So, the proposition we interpret the utterance as communicating can be about two different modal bases:

Body of information *not* updated with the presupposition:

$$\{(\text{  }), (\text{}), (\text{}), (\text{})\}$$

Body of information *updated* with the presupposition:

$$\{(\text{  }), (\text{}), (\text{})\}$$

If (E) communicates a proposition about the second body of information, it is vacuously true. That is how the presupposition makes a difference: it shifts what proposition we interpret the utterance as communicating so that the proposition communicated can be vacuously true. And if this presupposition had been absent, the proposition (E) communicates couldn't be vacuously true.

Now let's shift to the modified context, where all three gates can open at once. I claim that my new presuppositions are *absent* in that context. If they are, the indicatives are *incompatible*. Updating with both indicatives goes as follows.

$$\{(\text{  }), (\text{}), (\text{}), (\text{}), (\text{}), (\text{}), (\text{}), (\text{})\}$$

$$\text{CAfterAccepting(W)} = \{(\text{  }), (\text{})\}$$

$$\text{CAfterAccepting(W)and(E)} = \{\ \}$$

Neither (E) nor (W) is vacuously true at any point. The context always includes some world where Top Gate opened. So (E) isn't vacuously true after we've updated with (W). It's

25 Here and throughout the rest of the paper, I ignore complications about time. It's possible, for example, that the position of the gates changes after the eyewitness reports. I take the indicates to express propositions about something like intervals of time, with (W) expressing that all the water ran westwards if Top Gate opened in the relevant interval. If Wesla doesn't *know* that the West Gate stays open for the entire interval, she wouldn't know the indicative she asserts, and would violate the knowledge norm of assertion.

false at every world where Top Gate opened. That's the reason why we hear (W) and (E) as inconsistent in the second context. There's no consistent way to update with both of them.

We've just seen that *if* the presupposition differs between these two cases, then we can also predict a difference in the compatibility of the conditionals in the two cases. The next section will argue that the two utterances *do* plausibly carry different presuppositions in the two contexts.

Before making that argument, though, I want to anticipate an important objection. The objection is that my proposal makes it irrational for either Wesla or Esther to assertively utter their indicatives. I say that Wesla's uttering (W) presupposes something about the position of the East Gate. But she can appropriately utter (W) without any evidence about the position of the East Gate—and in that case, it doesn't seem like she has evidence for what she's presupposing.

This objection makes two kinds of mistakes. First and less importantly, the presupposition is a material conditional: that Top Gate opened → the *East* Gate was closed. And Wesla can know that material conditional in virtue of knowing that the West Gate is open while knowing the construction of the gate system. Second and more importantly, the norms for presupposition are different and less demanding than the norms on assertion. The benchmark account that Irene Heim (1992) developed illustrates this important point. As a very rough approximation, her account predicts that ⌜A asserts that S⌝ is true iff the propositions that A asserts, when combined with the presuppositions of S, together entail S.[26] So even if knowledge is the norm of assertion, Wesla's utterance is guaranteed to satisfy the norm. The propositions that Wesla knows (that the West Gate was open), plus the presupposition of her utterance, do together entail (W). So her assertion is guaranteed to satisfy the knowledge norm, whatever attitude she has to the presupposition of her utterance.[27]

Explaining the Presupposition

To show that some utterance *u* presupposes *p*, we need to first explain why the utterance is *associated* with *p*. That is, we need to explain why we interpret someone who makes that utterance as accepting p. I will call this the Association question. The answer to the Association question is sometimes semantic. *p* might be part of the content that *u* semantically

26 In fact, she also requires that A *believes* the presuppositions of S—but see Stalnaker (2002) for considerations for requiring acceptance rather than belief.

27 For a systematic argument that the norms of presupposition are different and less demanding than the norms on assertion, see Caleb Perl (2020).

As a further complication, consider the case where Wesla saw at 9 am that the West Gate is open, and asserts that if Top Gate opened before 10 am, all the water ran westwards. Unfortunately for her, the West Gate was shut at 9:30 am. In this case, the propositions that she knows plus the presuppositions of her utterance do *not* entail (W), because she doesn't *know* that the West Gate stayed open the whole time. More generally, whether Wesla knows (W) and whether it was appropriate for her to assert (W) swing free of whether she's justified in believing the presupposition.

expresses.[28] In other cases, the answer is pragmatic. *p* might be a conversational implicature of the assertive utterance *u*.[29]

However, answering the Association question isn't enough to show that an utterance *u* presupposes *p*. Presuppositions are distinctive because speaker-hearers interpret them as not the main point of the utterance, as backgrounded and not-at-issue. So, to show that *u* presupposes *p*, we also need to explain why p is interpreted as backgrounded, not-at-issue content. I'll call this the Backgrounding question. Foundational theories of presupposition aim at answering this question. For example, Stalnaker has developed an approach to presupposition that answers the Backgrounding question by appeal to the way that rational hearers would interpret what the speaker is doing.[30] I'm proposing that utterances of (W) carry a presupposition about the position of the East Gate, in the original context.

(W) If Top Gate opened, all the water ran westwards.

presupposes that

(P) Top Gate opened → the East Gate was closed

Let's start with the Association question: why is (W) associated with (P)? The answer is that (P) is an obvious a priori consequence of (W), given the propositions that are common ground. Given what's common ground, only two gates can be open at once. If the water ran westward, Top Gate and West Gate were both open. So if Top Gate opened, the East Gate was closed. That's why (P) is associated with (W).[31]

This answer to the Association question should be uncontroversial. Compare a Russellian account of definite descriptions, where "the F is G" semantically expresses that there is a unique F, which is also G. Utterances of "the F is G" presuppose and don't just assert that

28 Think of the difference in presupposition between an utterance of "John started dancing" and an utterance of "John continued dancing."

29 Utterances with universally quantified expressions like "every F is G" presuppose that there are some Fs. The answer to the Association question in this case is plausibly pragmatic: that the utterance is not cooperative unless the speaker accepts that there are some Fs.

30 Stalnaker (1973, 1974, 1998, 2002).

31 My answer to the Association question makes a plausible if substantive assumption. It assumes that presuppositions don't need to be met in order for the sentence to have a truth-value. (Since the presupposition is an a priori consequence of the proposition asserted, the proposition asserted is false if the presupposition is.) Though my assumption is substantive, it is highly plausible. It's the assumption to make when you recognize that presuppositions can have a wide variety of sources, and that presuppositions can have other sources than the need to avoid truth-value gaps. See p. 452 of Stalnaker (1973) and pp. 86–91 of Soames (2009) for particularly clear discussions.

there is a unique F.[32] The Russellian answer to the Association question is that the proposition that there is a unique F is an obvious a priori consequence of the proposition expressed. My answer to the Association question has the same structure as the Russellian's answer. I'll now argue that (P) (that Top Gate opened → the East Gate was closed) is a presupposition of uses of (W)—a backgrounded, not-at-issue commitment.[33] My argument starts by observing that the point of using (W) is to communicate that there's a connection between Top Gate's opening and all the water running westward. The connection holds because every way of extending the state of affairs where the West Gate is closed to also be a state of affairs where Top Gate opened is a state of affairs where all the water runs westward. And, crucially, the *East* Gate's being open doesn't in itself threaten that connection. The connection is totally grounded in the construction of the gate system and the West Gate's being open. Now the East Gate's being open does trivialize the connection. (It guarantees that Top Gate won't open.) But trivial connections are still connections.

I then claim that no cooperative speaker can use (W) intending to thereby assert something about the **East** Gate. A cooperative speaker who asserts (W) is intending to assert a connection between Top Gate's opening and all the water running westward. And the position of the East Gate doesn't bear on that connection. That connection can hold if the East Gate is open, and it can hold if the East Gate is closed. Since (P) is a proposition about the position of the East Gate, no cooperative speaker can intend to assert it by using (W). That's my answer to the Backgrounding question. (P) is a presupposition of uses of (W) because no cooperative speaker could use (W) to assert (P). Commitments of an utterance that aren't asserted are presupposed—and (P) is a commitment of uses of (W).

Go back to the comparison with the Russellian account of definites. Those Russellians hold that uses of ⌜the F⌝ are associated with the commitment that something is F. And they explain why that commitment isn't asserted, thereby explaining why it is presupposed. In doing that, they're making the same inference that I'm making, and are assuming that the speech act of assertion contrasts with the speech act of presupposing.[34] That's what I've just done, too.

---

32 David Beaver's taxonomy of presupposition triggers starts with definite descriptions and a reference to the relevant literature (Beaver 2001, 10), and John Hawthorne and David Manley have a helpful discussion of the pressures in favor of treating the existential claim as a presupposition—see §5.6 of their (2012); see especially 193n97.

33 You might think that it's odd to *argue* for a claim about what's presupposed, if you think that intuitions of language users are what settles what is presupposed. There are some examples where ordinary language users are in a good position to tell if a particular commitment is presupposed; for example, they might be in a position to tell whether "John stopped dancing" presupposes that John used to be dancing. However, there are other cases where ordinary language users are not in a position to tell what's presupposed; see Craige Roberts (ms) for an extended discussion of one such example. So, I don't think there's anything odd about needing to argue for a claim about what's presupposed.

34 This assumption is common ground in an otherwise heterogenous range of frameworks, like those of Barbara Abbott (2000), Dorit Abusch (2010), Márta Abrusán (2011), David Beaver (2001), Bart Geurts (1999), Irene Heim (1982, 1983), Daniel Rothschild (2011), Philippe Schlenker (2010), Mandy Simons (2001), Mandy

Now turn to the modified context, where all three gates can open at once. I claim that (P) is *not* a presupposition of uses of (W) in that context. The difference is that in the modified context, the East Gate's being open *does* threaten the connection that (W) expresses. The connection can't be totally grounded in the construction of the gate system and the West Gate being open. In that context, the West Gate being open doesn't combine with the Top Gate being open to guarantee that the East Gate is shut—the construction of the gate system allows all three to be open. So in the modified context, the East Gate's being open doesn't trivialize the connection. It undermines it. As a result, a cooperative speaker who uses (W) in the modified context *does* intend to assert something about the position of the East Gate. She intends to eliminate possibilities where it's open. Since presupposing contrasts with asserting, (P) is not a presupposition of uses of (W) in the modified context.

This result is very encouraging. We've already seen that (W) and (E) would be compatible in one context and incompatible in another if they carry different presuppositions in the two contexts. And we've just seen why they would carry different presuppositions.

Generalizations

I do not offer a material-conditional semantics for the indicative conditional. In fact, it's essential for my explanatory ambitions that I do not. If I did, I would predict compatibility between the indicatives in the context where all three gates are open. I'm instead assuming a semantics like that offered by Kratzer (1986) or Stalnaker (1975).

I also hold that the utterance of an indicative usually presupposes that the antecedent might be true.[35] And I think that presupposition is accommodated when it's present, guaranteeing that the indicative is not vacuously true. But I deny that an utterance of an indicative *always* carries the orthodox presupposition. In particular, I deny that the utterance carries that presupposition in the special context where it carries the presupposition introduced in 4.[36] That's why indicatives can only be vacuously true in certain special contexts. This section explores if my account will work more generally.

A General Recipe

Here's a more general recipe for constructing cases like Gibbard's. Start with three possibilities: A, B, and C, and suppose that exactly one of the three will happen. (Imagine, for concreteness, that A is the possibility where Andrea killed somebody, B the possibility where Billy did, and C the possibility where Candice did.) To complete the recipe, suppose that

Simons et al. (2010), Robert Stalnaker (1973, 1974, 1998, 2002), Rob van der Sandt (1992), and Deirdre Wilson and Dan Sperber (1979).

35 Karttunen and Peters offer an early statement of this kind of view (Karttunen and Peters 1979, 10); Kai von Fintel (1996) gives a more modern treatment.

36 I develop my account in more detail elsewhere (Perl (ms)), explaining why my new presupposition would "trump" the orthodox one.

one speaker knows that B didn't happen, and the other knows that C didn't happen.[37] I take it that the first speaker can know that if A didn't happen, C happened, and that the second speaker can know that if A didn't happen, B happened.

In order for my proposal to capture this kind of case, (utterances of) the indicatives would have to carry the following presuppositions.

(4) if A didn't happen, B happened

presupposes that A didn't happen → C didn't happen.

(5) if A didn't happen, C happened

presupposes that A didn't happen → B didn't happen

For example, an utterance of *if Andrea didn't do it, Billy did* would need to presuppose that if Andrea didn't do it, Candice didn't do it either. The indicatives would be vacuously true if they carried these presuppositions.

When does my presupposition arise, in general? My presupposition arises when there are *compatible* states of affairs that would each ground connections between a common antecedent and *incompatible* consequents.

General Proposal: An utterance with the form 'if P_0 is true, then P_1 is true' presupposes that P_0 is true → P_2 isn't true

- if P_1 and P_2 are incompatible,
- if there are states of affairs A_1 and A_2 that can both happen,
- if A_1 would ground a connection between the proposition P_0 and the proposition P_1, and
- if A_2 would ground a connection between the proposition P_0 and the proposition P_2,

This proposal captures Bennett's example.

- P_0 = Top Gate opened
- P_1 = all the water ran westwards
  A_1 = the West Gate was open, and one or neither of the Top or East Gates were open,
- P_2 = all the water ran eastwards
  A_2 = the East Gate was open, and one or neither of the Top or West Gates were open.

A_1 and A_2 could both happen. (Suppose that the East and West Gates were both open.) At the same time, though, they ground incompatible connections, since P_1 and P_2 can't

37 This general recipe comes from a discussion by Dorothy Edgington (1997)—see especially p. 107.

both be true. But A_1 would ground a connection between P_0 and P_1, and A_2 would ground a connection between P_0 and P_2.

Moreover, this Proposal has the right structure to capture the case where exactly one of Andrea, Billy, or Candice did it. In that case, we know that exactly one of them did it. So the state of affairs where *Billy* didn't do it would ground a connection between Andrea not doing it and Candice doing it. And the state of affairs where *Candice* didn't do it would ground a connection between Andrea not doing it and Billy doing it.[38] At the same time, though, both those states of affairs could happen; just imagine that Andrea did it.

My Proposal also makes the right prediction about the modified context, where all three gates can open at once. There are four relevant states of affairs: where no gates were open, where they were all open, where only the East Gate was shut, and where only the West Gate was shut. Only the last two states of affairs can ground a connection between Top Gate opening and all the water running in the same direction. Importantly, though, those last two states of affairs *cannot* both happen. If the first one happens, the East Gate is shut, and if the second one happens, the East Gate is open. And my proposal is only a claim about the cases where *compatible* states of affairs would ground incompatible connections. In this case, then, we would expect my new presupposition to be absent. And that's just what we need for the indicatives to be *incompatible* in this modified context.

Most importantly of all, this Proposal fits the pragmatic reasoning sketched earlier. Suppose that there is a state of affairs A_1 that grounds a connection between P_0 and P_1, and another state of affairs A_2 that grounds a connection between P_0 and P_2, even though P_1 and P_2 are incompatible. Suppose further that A_1 and A_2 could both happen. So A_2's happening doesn't itself threaten the connection between P_0 and P_1. As a result, someone who intends to assert a connection between P_0 and P_1 doesn't intend to assert something about A_2 or about P_2, since that state of affairs doesn't threaten the connection that she's intending to assert. However, the proposition that if P_0 is true, then P_2 isn't true an obvious a priori consequence of what she intends to assert. That's why that proposition will be presupposed: it can't be part of what she intends to assert.

It's crucially for this line of reasoning that the two states of affairs can both happen. That's why A_2's happening doesn't threaten the connection that's grounded in A_1's happening. So the cases where the relevant states of affairs *can't* both happen are cases where the pragmatic reasoning doesn't go through. And that's the fundamental reason why we hear (W) and (E) as incompatible in the modified context, but compatible in the original one.

38 Now talk about what the state of affairs would ground these connections is importantly elliptical. I mean what it would ground *in combination with* the facts that are common ground, like the fact that exactly one of Andrea, Billy, or Candice did it.

The Presuppositional Claim Is More Plausible Than My Particular Account

I've done two things in explaining Gibbard's case. I've suggested that indicatives carry an unappreciated presupposition that allows indicatives to be vacuously true. I've also sketched a constructive explanation of why the indicatives would carry an unappreciated presupposition in these cases. I'm more confident in positing the new presupposition than in my constructive explanation of where it comes from. I intend the constructive explanation more as a proof of concept than as the last word.

After all, there is better evidence for my new presupposition than for my constructive explanation of it. We've seen that (W) and (E) can be compatible in one context but incompatible in another. The best explanation of this difference is that they can be vacuously true in one context but not in the other. To make good on this explanation, though, we need some account of when indicatives can be vacuously true. And the best way to do that is to posit a mechanism that modulates the context set *before* we update with the indicative.[39] And presuppositions are just the right mechanism to modulate the context set in that way.[40]

There is also direct evidence that indicatives do carry just the presupposition that I posit. One hallmark of presuppositions is that they *project* from embeddings. For example, if S presuppose p, then utterances of ⌜Maybe S⌝ also tend to presuppose p. And there are certain facts that be explained only by positing my presupposition and taking it to project.

(6)?? The **East** Gate was open, but maybe if Top Gate opened, all the water ran **westwards**.

The two conjuncts in (6) can be compatible. That was the point of 1: Wesla can know her conditional even while the East Gate was open.[41] So we can't say that (6) is infelicitous because it always expresses something false.

Now this observation isn't enough by itself to show that we need to posit my presupposition. Theorists about indicatives tend to hold that indicatives presuppose that their antecedent might be true.[42] Will that presupposition make the conjuncts in (6) incompatible? No. Even if there's a possibility where East Gate was open and Top Gate open, there's another possibility where Top Gate was shut and East Gate open. And that other possibility is a

39 That was the upshot of the discussion of Williams' proposal, in 3.

40 This point is even more compelling when we recognize that normal assertive utterances of 'if p, q' presuppose that p might be true. (Karttunen and Peters offer an early statement of this kind of view (Karttunen and Peters 1979, 10); Kai von Fintel (1996) gives a more modern treatment.) If we think that a presupposition can sometimes expand the context set to prevent indicatives from being vacuously true, it would be unsurprising if another presupposition can contract the context set.

41 I assume that if ⌜Wesla knows that *s*⌝ and ⌜Wesla believes that maybe *s*⌝ are both true, ⌜Wesla knows that maybe *s*⌝ is true too. This principle may need still further refinement—maybe, for example, it only holds when Wesla has inferred ⌜maybe *s*⌝ from *s*. I don't think those refinements will affect the present point.

42 Lauri Karttunen and Stanley Peters (1979), Kai von Fintel (1996).

possibility where Wesla's conditional is vacuously true. And the second conjunct in (6) only requires there to be *some* such possibility.

Crucially, though, we *can* explain (6)'s infelicity if my presupposition projects. If it projects, an utterance of (6) communicates four propositions: (1) that the East Gate was open, (2) Top Gate might have opened, (3) Top Gate opened → the East Gate was shut, and (4) maybe if Top Gate opened, all the water ran. (2) and (3) are presuppositions, and (1) and (4) are what's semantically expressed. (1), (2), and (3) are jointly inconsistent. If there's a possibility where Top Gate opened, as (2) requires, (3) guarantees that that's a possibility where the East Gate was shut, which contradicts (3).

As a result, we *have* to posit just the sort of presupposition that I posit.[43] So if we find counterexamples to my constructive explanation of the presupposition, we should just go looking for another presuppositional account that does better.

Philosophical Upshots

I've presented my presuppositional proposal schematically, without filling in lots of important details. My overarching goal here is to establish the philosophical significance of the proposal, showing how a range of philosophical questions look different if anything like it is right.

For one thing, my presuppositional proposal is the best way to allow that conditional proof can systematically extend knowledge. The epistemology of indicatives is simpler and cleaner if it's right.

My proposal also is evidence against expressivist accounts of indicatives. So it's helpful to see this paper as offering two different arguments against expressivist accounts of indicatives. One is that we have to acknowledge that Wesla and Esther can both know their indicatives in the original contexts, and that expressivists can't acknowledge that they can. This argument does not rest on any constructive account of the semantics and pragmatics of indicatives. It rests on our pretheoretical conviction that Wesla and Esther do both know their conditionals, or the claim that conditional proof can extend knowledge, or on the knowledge norm of assertion.

4–5 have been tacitly making *another* argument against expressivist accounts of indicatives. I've argued for my presuppositional proposal as the best way to capture the difference

43 Unfortunately, my presupposition doesn't quite display as robust projection behavior as some central presupposition triggers.

The cases where my presupposition doesn't project should be explained by supposing that features of the context can cancel embedded presuppositions—those features can prevent the presupposition from projecting. Robert Stalnaker (1974) influentially unified a range of otherwise difficult data by supposing that the feature can have this effect, and Scott Soames (2009) notes important further points about this phenomenon. Dorit Abusch (2010) has an especially helpful recent discussion of these points. She emphasizes differences between different triggers and notes that it is easier to cancel some kinds of presuppositions than others. She suggests that the explanation of the presupposition is what determines how easy it is to cancel them. Importantly, the broadly Gricean explanation I've given of my new presupposition is exactly the sort of thing that we would expect to be easier to cancel.

between Bennett's original context, where Wesla and Esther can both know their indicatives, and the modified context, where they can't. And we saw back in 1 that expressivists struggle to make sense of vacuously true indicatives. Since my presuppositional proposal makes essential use of vacuously true indicatives, I conclude that it's off-limits for expressivists. Similar points generalize elsewhere. For example, allowing for vacuously true indicatives is also incompatible with any strict connection between indicatives and conditional credences—the sort of strict connection that David Lewis (1976) exploits.

I've also shown how orthodox contextualists can explain the Gibbard phenomenon. In fact, though, dynamic theorists like Malte Willer (2013, 2014) can also accept my presuppositional account.[44] Now those dynamic theorists can also offer the same semantic clauses as expressivists offer; they'll just interpret them differently. (Sarah Moss (2018) is helpfully explicit about this point.[45]) So my problem isn't a problem about the expressivist's semantic clauses. It's with her interpreting them as expressing conditional credences, rather than context change potentials. More generally, vacuously true indicatives are an interesting and novel way of testing different foundational claims about semantics. For example, it would be interesting to explore if relativists like John MacFarlane and Niko Kolodny (2010) can also make sense of vacuously true indicatives.

My proposal also gives us an exciting new tool for exploring other topics, like Vann McGee's purported counterexample to *modus ponens*.

Opinion polls taken just before the 1980 election showed the Republican Ronald Reagan decisively ahead of the Democrat Jimmy Carter, with the other Republican in the race, John Anderson, a distant third. Those apprised of the poll results believed, with good reason:

(7) If a Republican wins the election, then if it's not Reagan who wins it will be Anderson.
(8) A Republican will win the election.

Yet they did not have reason to believe

(9) If it's not Reagan who wins, it will be Anderson. (McGee 1985, 462)

We tend to reject (9), and McGee interprets that rejection as showing that (9) is *false*.[46] But given my presuppositional proposal, (9) might be vacuously true, but odd for other reasons. After all, we can attribute knowledge of (9). If you don't know the amount of support

44 In fact, my approach may look even more natural in a dynamic framework than in a contextualist framework.

45 See Mark Schroeder (2015c) for further discussion of the differences between these interpretations.

46 MacFarlane and Kolodny (2010) go even further, and argued that this example is evidence in favor of their relativist approach.

for the different candidates, but only that some Republican will win, it seems like you can know that (9) is true. And if you do know it, McGee wouldn't have found a counterexample to *modus ponens*, after all.

References

Abbott, Barbara (2000). "Presuppositions as Nonassertions." *Journal of Pragmatics* 32: 1419–37.

Abrusán, Márta (2011). "Predicting the Presuppositions of Soft Triggers." *Linguistics and Philosophy* 34, no. 6: 491–535.

Abusch, Dorit (2010). "Presupposition Triggering from Alternatives." *Journal of Semantics* 27, no. 1: 37–80.

Adams, Ernest (1975). *The Logic of Conditionals*. Dordrecht: D. Reidel.

Beaver, David (2001). *Presupposition and Assertion in Dynamic Semantics*. Stanford, CA: CSLI Publications.

Bennett, Jonathan (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Blackburn, Simon (2016). "All Souls Night." In *Does Anything Really Matter? Essays on Parfit on Objectivity*, edited by Peter Singer. Oxford: Oxford University Press.

Edgington, Dorothy (1995). "On Conditionals." *Mind* 104: 235–329.

——— (1997). "Truth, Objectivity, Counterfactuals, and Gibbard." *Mind* 106, no. 421: 107–16.

Gauker, Christopher (2008). "Against Accommodation: Heim, van der Sandt and the Presupposition Projection Problem." *Philosophical Perspectives* 22, no. 1: 129–63.

Geach, Peter (1965). "Assertion." *Philosophical Review* 74: 449–65.

Geurts, Bart (1999). *Presuppositions and Pronouns*. Oxford: Elsevier.

Gibbard, Allan (1981). "Two Recent Theories of Conditionals." In *Ifs*, edited by R Stalnaker W. L. Harper, and G. Pearce, 211–47. Dordrecht: D. Reidel.

——— (2012). *Meaning and Normativity*. Oxford: Oxford University Press.

Hale, Bob (1993). "Can There Be a Logic of Attitudes?" In *Reality, Representation, and Projection*, edited by Crispin Wright and John Haldane, 337–63. Oxford: Oxford University Press.

Hawthorne, John, Daniel Rothschild, and Levi Spectre (2016). "Belief Is Weak." *Philosophical Studies* 173: 1393–404.

Hawthorne, John, and David Manley (2012). *The Reference Book*. New York: Oxford University Press.

Heim, Irene (1982). *The Semantics of Definite and Indefinite Noun Phrases*. PhD thesis, University of Massachusetts.

——— (1983). "On the Projection Problem for Presuppositions." In *Second Annual West Coast Conference on Formal Linguistics*, edited by Daniel Flickinger, Michael Barlow, and Michael Westcoat, 114–26. Stanford, CA: CSLI Publications.

——— (1992). "Presupposition Projection and the Semantics of Attitude Verbs." *Journal of Semantics* 9, no. 3: 183–221.

Jackson, Frank (1987). *Conditionals*. Oxford: Blackwell.

Karttunen, Lauri, and Stanley Peters (1979). "Conventional Implicature." In *Syntax and Semantics, Vol. 11: Presupposition*, edited by Choon-Kyu Oh and David Dinneen, 1–56. New York: Academic Press.

Kolodny, Niko, and John MacFarlane (2010). "Ifs and Oughts." *Journal of Philosophy* 107, no. 3:115–43.

Kratzer, Angelika (1986). "Conditionals." *Chicago Linguistics Society* 22, no. 2: 1–15.

——— (2012). *Modals and Conditionals: New and Revised Perspectives*. Oxford: Oxford University Press.

Lewis, David (1976). "Probabilities of Conditionals and Conditional Probabilities." *The Philosophical Review* 85: 297–315.

McGee, Vann (1985). "A Counterexample to Modus Ponens." *Journal of Philosophy* 82, no. 9: 462–71.

Moss, Sarah (2013). "Epistemology Formalized." *Philosophical Review* 122, no. 1: 1–43.

——— (2015). "On the Semantics and Pragmatics of Epistemic Vocabulary." *Semantics and Pragmatics* 8: 1–81.

——— (2018). *Probabilistic Knowledge*. Oxford: Oxford University Press.

Perl, Caleb (ms). "Kinematics of Indicatives." University of Colorado, Boulder.

——— (2020). "Presuppositions, Attitudes, and Why They Matter." *Australasian Journal of Philosophy* 98, no. 2: 363–81.

Roberts, Craige (ms). "Only, Presupposition and Implicature." Unpublished manuscript, Department of Linguistics, The Ohio State University, Columbus, OH.

Rothschild, Daniel (2011). "Explaining Presupposition Projection." *Semantics and Pragmatics* 4, no. 3: 1–42.

Schlenker, Philippe (2010). "Presuppositions and Local Contexts." *Mind* 119, no. 474: 377–91.

Schneider, Benjamin (2010). "Expressivism Concerning Epistemic Modals." *Philosophical Quarterly* 60, no. 240: 601–15.

Schroeder, Mark (2008). *Being For: Evaluating the Semantic Program of Expressivism*. Oxford: Oxford University Press.

Schroeder, Mark (2015a). "Attitudes and Epistemics." In *Expressing Our Attitudes*, 225–56. Oxford: Oxford University Press.

Schroeder, Mark (2015b). *Expressing Our Attitudes: Explanation and Expression in Ethics, Volume 2*. Oxford: Oxford University Press.

Schroeder, Mark (2015c). "Is Semantics Formal?" In *Expressing Our Attitudes*, 209–24. Oxford: Oxford University Press.

Searle, John (1962). "Meaning and Speech Acts." *Philosophical Review* 71: 423–32.

Simons, Mandy (2001). "On the Conversational Basis of Some Presuppositions." *Semantics and Linguistic Theory* 11: 431–48.

Simons, Mandy, Judith Tonhauser, David Beaver, and Craige Roberts (2011). "What Projects and Why." *Proceedings of Semantics and Linguistic Theory (SALT)* 22: 309–27.

Soames, Scott (2009). "Presupposition." In *Philosophical Essays, Volume 1: Natural Language: What It Means and How We Use It*, 73–130. Princeton, NJ: Princeton University Press.

Stalnaker, Robert (1973). "Presuppositions." *Journal of Philosophical Logic* 2, no. 4: 447–57.

Stalnaker, Robert (1974). "Pragmatic Presuppositions." In *Semantics and Philosophy*, edited by Milton K Munitz and Peter K Unger, 197–213. New York: New York University Press.

Stalnaker, Robert (1975). "Indicative Conditionals." *Philosophia* 5: 269–86.

Stalnaker, Robert (1998). "On the Representation of Context." *Journal of Logic, Language, and Information* 7: 3–19.

Stalnaker, Robert (2002). "Common Ground." *Linguistics and Philosophy* 25, no. 5–6: 701–21.

Swanson, Eric (2006). *Interactions with Context*. PhD thesis, Department of Linguistics and Philosophy, MIT.

Unwin, Nicholas (1999). "Quasi-Realism, Negation and the Frege-Geach Problem." *The Philosophical Quarterly* 49, no. 196: 337–52.

——— (2001). "Norms and Negation: A Problem for Gibbard's Logic." *The Philosophical Quarterly* 51, no. 202: 60–75.

van der Sandt, Rob (1992). "Presupposition Projection as Anaphora Resolution." *Journal of Semantics* 9, no. 4: 333–77.

von Fintel, Kai (1996). "The Presupposition of Subjunctive Conditionals." In *MIT Working Papers in Linguistics 25*, edited by Orin Percus and Uli Sauerland, 29–44. Cambridge: MIT Working Papers in Linguistics.

——— (2008). "What Is Presupposition Accommodation Again?" *Philosophical Perspectives* 22, no. 1: 137–70.

Willer, Malte (2013). "Dynamics of Epistemic Modality." *Philosophical Review* 122, no. 1: 45–92.

——— (2014). "Dynamic Thoughts on Ifs and Oughts." *Philosophers' Imprint* 14, no. 28: 1–30.

Williams, Robert (2008). "Conversation and Conditionals." *Philosophical Studies* 138: 211–23.

Williamson, Timothy (2000). *Knowledge and Its Limits*. Oxford: Oxford University Press.

Wilson, Deirdre, and Dan Sperber (1979). "Ordered Entailments: An Alternative to Presuppositional Theories." In *Syntax and Semantics, Vol. 11: Presupposition*, edited by Choon-Kyu Oh and David Dinneen, 299–323. New York: Academic Press.

Yalcin, Seth (2007). "Epistemic Modals." *Mind* 116: 983–1026.

12

MODELING WITH HYPERPLANS

Seth Yalcin

A key gizmo in Gibbard's formal model of normative judgment is the thing he used to call a "complete system of norms" (Gibbard 1986, 1990), and which, by around Gibbard (2003), evolved into the *hyperplan*. In this paper, I want to ask what hyperplans are, and ask how best to use them in modeling normative thinking.

One part of this paper is exegetical. There is perhaps less than universal agreement in the literature on how Gibbard's systematizing with hyperplans works exactly. I offer a take.

The other part is exploratory. I think there is a way of theorizing with hyperplans that is not quite Gibbard's way, but which is also expressivistic, and which is worth looking at. I have tried to say so in Yalcin (2012, 2018), but here I offer a more focused development. I will call (my take on) Gibbard's package of views about how to model with hyperplans Plan A. I spend the first section of the paper setting Plan A out. The alternative I will set out, Plan B, is the topic of the sections after that. Even if you don't leave the paper preferring Plan B to Plan A, I hope you'll find the contrast clarifying. In separating these two ways of theorizing with hyperplans, really I'm trying to bring out two rather different ways of conceptualizing expressivism.

1. Plan A: Gibbard's Model

1.1. Easing into the Picture

Hyperplans ultimately figure in an abstract model of what it is for the idealized rational thinker-planner to be normatively opinionated. To get into the model, a helpful starting point is the sort of picture of agents we find in a decision-theoretic perspective, where we suppose that agents have (1) a take on what the world is (probably) like, representable

formally by a probability measure over a space of possible outcomes, and (2) some preferences, representable as a partial ordering of possible outcomes. Or if we are satisfied to ignore distinctions between degrees of confidence, as I will be for now, we can replace the probability measure with a set of possible situations, the situations that might obtain for all the agent believes—the agent's *doxastic alternatives*. A textbook modeler in this vein will use it to say that a rational agent is the sort of being that generally acts in ways that would realize her preferences, at least relative to possible situations compatible with how she takes things to be. That gives an initial skeletal picture about how belief and desire connect to action, one not totally alien from the way ordinary belief and desire talk is invoked to explain behavior.

Saying only this much still leaves a lot of wiggle room, of course, when it comes to the question how exactly to interpret natural language talk of mental states—ascriptions of belief, desire, and the like—using elements of the model. We informally use this language when explaining in the first place what the model is supposed to be representing, but we could be more systematic. So, we could add to the story, stating idealized truth-conditions for ordinary ascriptions of belief and desire using the concepts of this model. The model has a component approximating what the folk call "belief" and a component approximating what the folk call "preference." We could invoke these components in stating truth-conditions for belief and desire talk. Take for instance:

(1) Holmes thinks it is not too late to catch the train.
(2) Holmes wants to get on the train.

The usual move is to say that when (1) is true, that's to do with the agent's doxastic alternatives: it is true just in case none of Holmes's doxastic alternatives are ones where it is too late to catch the train.[1]

For (2), it is initially tempting to say that its truth is to do just with the agent's preference ordering, as the truth of (1) is just to do with the agent's doxastically open possibilities. But on reflection, wanting is a more subtle thing than that: what one wants depends partly on what one thinks the world is like, and not just on which possible states one prefers to others. I want to sue the man whose dog bit me, but this want of mine cannot necessarily be read off my preference ordering alone. It's not that I prefer situations where I sue the man to all others—after all, in the possible situations I most prefer, I am never bitten by a dog at all. It seems like wanting a possible outcome is more like "preferring it to certain relevant alternatives, the relevant alternatives being those possibilities that the agent believes will be realized if he does not get what he wants" (Stalnaker 1984, 89). If that is right, then whether an agent wants something is a function of both their preference ordering and their doxastic state. Wanting cannot be modeled with a preference ordering alone, but neither does it map

1 So it goes in the familiar tradition of possible worlds modeling descending from Hintikka (1962) and running prominently through theorists like Stalnaker (1976, 1984) and Lewis (1979).

to its own special feature of the model, as belief basically does. When you say somebody wants something, you're saying that their preference ordering and their doxastic state, taken as a pair, satisfy a certain complex property—one in the direction of what Stalnaker suggests, I'd claim, though the details don't matter right now.[2]

In the last paragraph, I mean to draw out the point that although states of preference are basic to our model, some of the key pieces of ordinary language we use to describe an agent's preferences—desire ascriptions, ascription of want—might well have a more subtle connection to the model than one might offhand suppose. We have to distinguish concepts from the model (like preference orderings) from ordinary language notions (like wanting), and we have to allow for the possibility that they may be connected by something other than a straight line. This will be good to have in mind as we turn to normative judgment.

1.2. Adding Normative Judgment

Now suppose we want to extend our model, adding a formalization of what it is to have normative views. I want to approach this via a concrete example, like:

(3) Holmes thinks he ought to pack.

Given the sort of model we are assuming, we can ask: what does a model of Holmes's state have to be like to render this true? What features of our model, if any, correspond to what this sentence says?

An unfussy normative realism would favor the view that there is nothing fundamentally new to say here. Holmes's thinking he ought to pack is not deeply different than his thinking it's not too late to catch the train. Both are about the way he takes the world to be, about what he thinks the facts are.[3] On this view, to see whether our model of Holmes reveals that he thinks he ought to pack, we just check whether his doxastic alternatives are all possibilities where he ought to pack. Possible states of the world fix facts such as Holmes's being required to pack, on this view—just as much as they fix facts about train schedules.

Alternatively, we can say that's wrong. Holmes's thinking he ought to pack is not just another one of his views about how the world is. We cannot discover whether he is in this state just by looking at the doxastic alternatives he keeps open. There is something like an official list of classic motivations for rejecting realist-style truth-conditions for (3), much of it descending from Hume (usually it includes: the alleged metaphysical queerness of the realist's normative properties, the direction of fit of normative judgment, its alleged internal link to motivation, the open question argument, and the supposed intractability of normative disagreement), but our interest is just in getting some interesting alternative analyses of (3)

2 Heim (1992), Levinson (2003), and Lassiter (2011) contain more on the semantics of 'wants'.

3 Perhaps *modulo* whatever *de se* or self-locating elements we might think infect these thoughts.

on the table, Gibbard's especially. Gibbard wants to say that the truth of (3) turns on features of Holmes's mental economy beyond his doxastic alternatives. What features?

Well, so far our model of an agent includes only a set of doxastic alternatives and a preference ordering, so either ought-thoughts are tied up somehow with preference, or else we need to expand our model of Holmes's mental economy. Consider briefly the first option. An idea in this vein would be to basically identify ought-thoughts with wants: we could say that states of *thinking it ought to be that p* are tantamount to states of *wanting that p*. Both kinds of ascription serve to place constraints on the agent's preference ordering (and the agent's doxastic state, too, if such states really are equivalent to states of wanting). Thus *ought*-thoughts are not purely doxastic—we could say they are "preference-laden." Carnap and Russell seemed to favor this kind of analysis—they suggested that normative statements express wishes (see Carnap 1935, 24; Russell 1935, 236–7). This approach recommends the idea that (3) serves to tell us about what Holmes wishes or wants.

The evident trouble is that it seems like (3) can be true even when it's also true that Holmes prefers not to pack. What I think ought to be the case and what I want, it seems obvious, can come apart. Maybe desiring something is a way of valuing it, but there clearly are ways of valuing, or evaluating, things that are not just ways of preferring them; so anyway I'll assume without argument. Should we nevertheless try to say that if (3) is true, then really Holmes *does* want to pack in some maybe very attenuated or nuanced sense of *want*? Or, less commitally, that the relevant state of mind is "desire-like"? Such a view could be developed,[4] but we mean to get to Gibbard, and his view seems to me not helpfully put this way, so I will just take it our model must go beyond doxastic states and preference orderings, beyond belief and desire as we have them now modeled, to bring normative attitudes into its ken.

Now eying directions for expansion, one possibility would be to bring in feelings. There is the notorious emotivist strand of expressivism, prominent in Ayer (1936), which ties normative states of mind to emotional states. This view is in the ballpark of saying that (3) reports that Holmes has good vibes about packing. I set this idea aside too in order to focus on a different idea, the idea of bringing in states of mind to do with planning and intention—the idea Gibbard favors and develops in Gibbard (2003) particularly.

1.3. Planning

A good first step into it is to follow Bratman (1987). Against philosophers like Anscombe and Davidson, we don't try to explain states of intention away adverbially, analyzing them in terms of the notion of acting intentionally, the latter then analyzed as a matter of the right fit obtaining between the action and the agent's beliefs and desires. Instead, we think

4 Hare offered a sophisticated version of this kind of approach, holding that moral claims express preferences the agent takes in a certain sense to be universalizable (Hare 1981). One often sees expressivism characterized as a view that holds that normative statements express desire-like states of mind (e.g., Ridge 2014; Schroeder 2009). Often what is intended here is a similarity in "direction of fit."

of intentions as pieces of plans agents have, and we take it planning states of mind are their own thing—they do not admit of reduction to belief, preference, or some admixture. They have respectable standing of their own, on par with those other states. Our formal model of agents needs expanding to make room for them, take it. We could, following Bratman, see the addition of intention to the picture as rendering our model less idealized, bringing in a state of mind whose *raison d'etre* is partly the fact that we don't have unlimited mental resources for reevaluating the pros and cons of the options we face as the facts change.

So we want to expand the model, adding planning. Here is where hyperplans enter the picture. Gibbard's approach is to elaborate our model of doxastic states, interleaving in a planning dimension. Previously, we had it that a doxastic state is a set of possibilities—a set of ways that the world could factually be, for all the agent believes. Now following Gibbard we say that a doxastic state is given by a set of pairs of a possibility and a hyperplan. Let me explain what formally a hyperplan is, and then try to say how by adding these things to our model Gibbard means to clarify planning thinking.

On the formal side:

> A *hyperplan*, we can stipulate, covers any occasion for choice one might conceivably be in, and for each alternative open on such an occasion, to adopt the plan involves either rejecting that alternative or rejecting it. In other words, the plan either forbids an alternative or permits it.
>
> (Gibbard 2003, 56)

For any possible occasion of choice, a hyperplan settles some nonempty subset of the options available on that occasion—intuitively, the ones deemed permissible by the hyperplan. We can take hyperplans to be functions on centered worlds.[5] Fed a centered world c—an "occasion of choice"—where the agent centered at c faces a set of options O_c (the options fixed by the centered world c), the hyperplan h outputs some nonempty subset of O_c—intuitively, those options deemed permissible by h. What formally are options? We could think of an option as a set of centered worlds, the set of centered worlds where the option is realized.

A hyperplan is not the sort of thing tied to any one particular agent—it is not a complete contingency plan for some one particular planner. Rather, it settles what is okay to do for *every* planner, actual or possible, and for any possible situation any planner could find herself in. Hyperplans are defined on arbitrary centered worlds.

Let's say a *fact-plan alternative* is a pair of a centered world and a hyperplan. A Gibbardian plan-laden belief state is a set of fact-plan alternatives. Thus, Gibbard is thinking of plan-laden belief as cutting a strictly richer space of alternatives than we previously had it. It is

5 "A *situation s* is a triple $\langle w,i,t\rangle$ of a world w, an agent i in w, and a time t at which agent i in world w has a choice of what to do. For each such situation s, there is a set $a(s)$ of *alternatives*. These are maximally specific acts open to person i at time t in world w. A *hyperplan p* assigns to each situation s a non-empty subset $p(s)$ of the alternative set $a(s)$" (Gibbard, 2003, 100).

helpful to see him as building on Lewis (1979). First, he is taking on board Lewis's main idea, that to model belief *de se* in full generality we do better to think of doxastic alternatives as centered worlds rather than uncentered worlds. Then going beyond Lewis and considering the yet-richer space of fact-plan alternatives, he proposes that subsets of this space can well represent our states of mind that combine belief and decision.

Since a fact-plan alternative fixes a centered world, a set of fact-plan alternatives can from a modeling perspective do everything a set of centered worlds can do. Therefore, Gibbard's model of belief is at least as robust as Lewis's. Indeed, Gibbard will model ordinary (if ideal) "prosaically factual" belief essentially along Lewis's lines, as a matter of what centered worlds figure in the fact-plan alternatives the agent leaves open. For example, let P be Holmes's plan-laden belief state, a set of fact-plan alternatives. Gibbard will say that (1)—an ascription to Holmes of the prosaically factual (if self-locating) belief that isn't not too late to catch the train—is true just in case:

$$\text{For all } \langle c, h\rangle \in P, \text{ it isn't too late to catch the train at } c$$

You can see that the hyperplan component is an idle wheel. That wheel hits the pavement only when we come to planning and normative judgment. Gibbard says that you have views about how things are and views about what to do. Your views about what to do are not settled by your views about how things are. You might be fully opinionated about how things are and yet be unsettled about what to do. Your views about what to do are reflected in the hyperplans your plan-laden belief state rules in or out.

How exactly? Let us bring in a term of art of Gibbard's, "the thing to do," which he stipulates to be expressive of planning states. Take

(4) Holmes thinks packing is the thing to do.

Following Gibbard, we're understanding this to mean that Holmes is resolved on packing, that he plans to pack.[6] We can get a handle on how hyperplans are put to work by asking: what does Gibbard say are the truth-conditions of (4), stated in terms of his model of plan-laden belief?

It is useful to first mention a wrong answer to this question. One might think Gibbard's idea is that (4) is true iff

$$\text{For all } \langle c, h\rangle \in P\text{: } h \text{ permits only options that entail packing}$$

6 Why then don't we just talk about the ordinary sentence 'Holmes plans to pack' instead of (4)? I take it Gibbard fixates on (4) because he is setting up to eventually talk about something like (3), which resembles (4) more closely. More on the relation between (3) and (4) below.

On this reading, Holmes's thinking packing is the thing to do is a matter only of what hyperplans are left open by the fact-plan alternatives in P. Thus, we have sort of the formal opposite of the case of straight factual belief: here it is the factual, centered worlds dimension that goes idle. Gibbard is sometimes described as a "pure" expressivist, where this is meant to contrast with the idea of a mixed or "hybrid" expressivism, which would hold that to be in a normative frame of mind is to be in a state that mixes the cognitive (doxastic) with something noncognitive (nondoxastic).[7] If one viewed Gibbard through this lens, an interpretation of his formalism like this might seem natural.

But this is not Gibbard's idea. Actually, this idea doesn't really make sense, because it doesn't really make sense to say a hyperplan permits (or requires) something *simpliciter*; rather, it does so only relative to a choice of a centered world. Instead, Gibbard's idea is that (4) is true if

> For all $\langle c, h\rangle \in P$: $h(c)$ permits only the options o in O_c that entail packing by the agent that o is centered on

The idea is to go to each fact-plan possibility compatible with Holmes's state and ask whether, when you evaluate the hyperplan component of that fact-plan possibility at its centered world component, the output of the hyperplan permits only packing outcomes. If so, then (4) is true: Holmes thinks packing is the thing to do. In this way, thinking that packing is the thing to do is a complex property of the individual fact-plan possibilities one's state leaves open.

In one way, there is a similarity between the state of thinking that something is the thing to do and the state of wanting, as we modeled it earlier following Stalnaker. Again, when you say somebody wants something, you are effectively saying that the pair of their doxastic state and their preference ordering satisfies a certain complex property; you are not just saying something about their preference ordering. In a somewhat similar way, for Gibbard, if you say somebody thinks something is the thing to do, you are effectively ascribing a complex property to the combination of their factual beliefs and their plans, not just saying something about the hyperplans they rule in or out. The truth of (4) turns partly on Holmes's prosaically factual beliefs. (To this extent Gibbard is already an expressivist of the hybrid style.[8]) On reflection, this should be obvious: one cannot think that packing is the thing to do unless one takes oneself to have the option to pack, and whether one has that option is a factual matter.

In another way, there is a formal disanalogy between the state of thinking that something is the thing to do and the state of wanting. Whether an agent is in the former kind of state is

7 See, for instance, Schroeder (2009).

8 There is of course still an important difference between Gibbard's expressivism and the hybrid-style realist-expressivism of, for instance, Copp (2001). The latter goes in for a realist metaphysic of normative properties, whereas Gibbard does not.

a matter of whether each of the fact-plan alternatives their state of mind leaves open, taken individually, satisfies the relevant condition. By contrast, there is no set of "desire alternatives" such that what the agent desires is a matter of what holds throughout those alternatives.

Let me stress a key difference between Lewis's kind of enrichment to logical space and Gibbard's. Lewis paired possible worlds with centers to capture belief *de se*. Belief *de se* is still factual belief, in the sense that there is a fact of the matter as to what centered world(s) you actually occupy. Your belief state is doing well when it doesn't exclude centered worlds centered on you. When Gibbard pairs centered worlds with hyperplans, though, the resulting space is not fully factual, in the sense that there is no fact of the matter as to which fact-plan alternative you occupy, owing to the hyperplan component. There is, let's presume, an actual world, and there is a fact of the matter about how you are situated in that world, but there is no "actual hyperplan."[9] There is no fact of what to do, such that your plan-laden state of belief is doing well if it characterizes that plan. This gets at the point that the role of the plan-laden part of your belief state isn't to represent plan facts. Again, your view about what to do is not a view about how things are.

1.4. Planning and Normative Thinking

What has all this modeling of planning to do with normative thinking? Let's put these three sentences again in front of us:

(3) Holmes thinks he ought to pack.
(4) Holmes thinks packing is the thing to do.
(5) Holmes plans to pack.

It seems it should come out that (4) and (5) have basically the same truth-conditions, because of how Gibbard stipulates the meaning of "thing to do"; and we just said what those truth-conditions are. The tour through planning was supposed to result in a story about normative judgment, the sort of state ascribed by (3). Is Gibbard proposing that (3) has essentially the same truth-conditions as (4)/(5)? Is Gibbard's model of this normative state just the state we have identified for (4)/(5) using hyperplans?

Gibbard is certainly angling for the idea that "we can understand normative judgments as an aspect of planning" (Gibbard 2006, 732). He says:

> If I think that something is now the thing to do, then I do it. My hypothesis about ordinary *ought* judgments is that they are judgments of what to do, of what is the thing to do. I don't, then, think that I ought right now to defy the bully unless I do defy him. If I fail to defy him, then as a matter of the very concept of *ought*, I don't believe I ought to.
>
> (Gibbard 2003, 153)

9 Thus, I called the hyperplan component a "nonfactual parameter" in Yalcin (2011).

There are three concerns one might have about tying planning and normative judgment so closely together. Going through these will help to bring the picture out.

1.4.1. Problem: Third-Personal Oughts and Their Connections to Plans

The first concerns third-personal oughts and their connections to plans. Suppose Holmes thinks packing is the thing *for Watson* to do, but not for himself. How does Holmes's judgment here about the thing to do tie into Holmes's own planning, given that he obviously can't literally decide for Watson?

Gibbard replies that in the relevant sense, Holmes can decide for Watson—he can decide what to do when in Watson's position. Roughly I understand Gibbard to model as follows: let's say WATSON is a function that takes a centered world and shifts its center to Watson (so it maps $\langle x,w\rangle$ to $\langle$Watson,$w\rangle$, for any x, at least when Watson can be found at all in w). Then Holmes thinks packing is the thing *for Watson* to do just in case

$$\text{For all } \langle c, h\rangle \in P\text{: } h\,(\textsc{watson}(c)) \text{ permits only the options } o \text{ in } O_{\textsc{watson}(c)} \text{ that entail packing by the agent } o \text{ is centered on}$$[10]

That gets at Holmes's view about what to do when in Watson's shoes (as Holmes takes those shoes to be). We don't routinely speak of deciding and planning for others, and insofar as we do, it doesn't come as freely as our talk of the things others ought to do. I understand Gibbard to grant this. He will say that he means to get at the underlying structure of planning thinking. The link between this structure and our language might be indirect, just as the link between preference orderings and statements of want is indirect.

1.4.2. Problem: Failing to Plan to Do What You Think You Ought

Second problem: can't you think you ought to do something without planning to do it? Gibbard explains this by saying that we are sometimes fragmented in our normative thinking:

> A person often isn't "of one mind" in accepting a plan or not. For a crucial sense of 'ought', I say, the following holds: if you do accept, in every relevant aspect of your mind, that you ought right now to defy the bully, then, you will do it if you can. For if you can do it and don't, then some aspect of your mind accounts for your not doing it—and so you don't now plan with every aspect of your mind to do it right now. Whatever aspect of your motivational system issued in your doing otherwise didn't accept the plan to defy him right now. And so, it seems to me, there's a part of you that doesn't really think you ought to. You are of more than one mind on whether you ought to defy him.
>
> (153)

10 A variant of this would replace WATSON with something like an individual concept reflecting Holmes's mode of presentation of Watson.

This seems to trace the problem to the idealization of the model, rendering it perhaps analogous to the sort of idealization we were already involved in when modeling a doxastic state with a set of possibilities. The problem of logical omniscience afflicting our model of belief might also owe to the fact that realistic agents can be of "more than one mind" about something—fragmented and compartmentalized, as Stalnaker (1984) and Lewis (1988) discuss. Whether this is a big problem for the model depends on how much of it survives when we complicate it enough to reduce or eliminate the idealization. The kind of fix Stalnaker and Lewis envisage on the doxastic side seems to preserve intact the key idea of modeling with "a way things are according to the agent" (though it's certainly a matter of debate); perhaps Gibbard can make a response in a similar shape, representing a divided mind with distinct sets of fact-plan alternatives. For the purposes of this paper, I treat this objection as contained.

1.4.3. Problem: Deciding among Several Permissible Options

The third problem is the one that will take up the most space. Can't you plan to do something without thinking you ought to do it? This worry is to do with the fact that plans frequently involve arbitrary choices. Suppose I plan to sit in the couch over there, by walking to it. I execute this scheme, deciding to take the first step with my left foot. I decided to step left first, but the choice was arbitrary; in my view, Righty was equally up to the task.[11] Had I stepped right instead, the world would not have been lesser to me in any way. My aim was just to get to the couch, and one cannot get to the couch except by some route. The thing is that while I decided left foot first, it's not that I thought I ought to step left first. Where is the space in Gibbard's model for this difference?

To be clear, the objection here isn't that Gibbard can't model the difference between thinking an option required and thinking of it as one permitted option among several. His hyperplan apparatus is perfectly designed to model that difference—remember that a hyperplan may permit more than one option.[12] Rather, the objection, flatfootedly, is that what you actually plan to do might be more specific—might be strictly more resolved—than your view about what you ought to do, and therefore your planning state and your state of *ought-thoughts* are not really the same, and therefore the truth-condition we sketched for (4) won't carry over directly to (3). And therefore we remain in the lurch about (3), when really that was what we were out to analyze in the first place.

We could also put the problem using a piece of terminology from Gibbard (2006). Let's say a *strategy* is a hyperplan that permits exactly one option for each contingency. Then the point is that, at least restricted to the occasion I face, I have in some clear sense both a strategy and a plan: my strategy calls for left foot first, whereas my plan—the part of me that

11 We can say I thought about it for a second, if that makes it easier to describe me as "deciding." The present issue doesn't turn on which decisions or paths of action are selected "subpersonally."

12 His model also has no problem modeling the difference between lacking a view about which of several alternatives is permitted and thinking them all permitted.

reflects what I view as permitted and required—just says to step with some foot or other, permitting either foot. If we grant that I decided to step left, then it seems that the part of me that embraces a plan permitting multiple alternatives is not identical to my decisional state.

This draws the ambiguity in "thinking what to do" out somewhat. Explaining the state of *thinking what ought to be so* in terms of the state of *thinking what to do* can seem like a way of analyzing a state specified using a normative concept in terms of state that is not so specified. But while *thinking what to do* can mean *deciding what action to perform*, it can also just mean *thinking what is to be done*, with the infinitive conveying an implicit normative modality. In other words, "thinking what to do" can just mean "thinking what ought to be done"—in which case the former marks no special progress. ("The house is to be cleaned" can be heard as a prediction, but when said to the housekeeper, it rings as normative. Gibbard's discussion of "to be desired" (Gibbard 2003, 22) displays sensitivity to just this kind of contrast.) There evidently are these two different things "thinking what to do" can mean. Obviously, we want to make sure that any theory that explains *thinking what ought to be the case* in terms of *deciding what action to perform* doesn't rely along the way on this ambiguity.

Gibbard says some things that seem to speak to this issue.[13] He agrees that there is this difference we are talking about:

> Is there a difference, then, between rejecting an alternative—not permitting it to myself—and just not choosing it? Surely there is. The two differ in "valence" or oomph. To think this distinction intelligible, must I already think that permitting myself an alternative consists in attributing some special kind of property to it? No, distinguishing in this way is clearly a part of planning, but there is no need to think, at the start of inquiry, that distinguishing this way is a matter just of factual belief. My claims here concern what one commits oneself to in planning, and the facts I'm allowing at the outset are straightforward and prosaic. We can distinguish preference and indifference without first admitting facts of a kind more ethereal.
>
> (Gibbard 2003, 153)

Put aside the issue of whether a normative realism would work better for explaining the difference we're getting at—our interest is just in understanding Gibbard's alternative to an explanation based on "ethereal" facts. He indicates here that the difference is one tied up with the difference between preference and indifference, a difference he stresses earlier in the book:

> Buridan's ass might have been wiser, and a wiser ass would choose one bale of hay or choose the other. She wouldn't thereby rule out choosing either—or at least there's an important sense in which she wouldn't. She wouldn't be in disagreement with plumping for the other from indifference. It is in the nature of planning, after all, to distinguish rejecting an alternative

13 I am especially indebted here to conversations with Sophie Dandelet, and also to her Dandelet (2017).

> by preference from simply not choosing it in that, from indifference, one chooses another. Rejecting an alternative is something more than just taking a different alternative when there is more than one alternative that one doesn't reject by preference.
>
> (Gibbard 2003, 55)

The idea is that when I stepped left first, I did it out of indifference. In a counterfactual scenario where I acted the same but permitted myself only to step left, Gibbard seems to say that I do this by preference. In these two scenarios, I act the same but I am representable by different sets of fact-plan alternatives, and ultimately this difference is grounded, Gibbard appears to say, in my state of preference.

I pause exegesis mode here to say that this injection of preference into the story surprises me. Initially it seemed the proposal was going to be that planning, understood as a distinct state of mind from preference, is what normative judgment is tied up with. But Gibbard seems to say that what one views as permitted is in some sense explained by what one prefers, since he seems to be saying that an agent's viewing several options in a situation as permissible entails that the agent is preferentially indifferent among the options. That's a very strong tie to hypothesize between planning and preference. (I won't dwell on Gibbard's motivations here, which seem to do with how he hopes to account for when agents with different views about what to do count as disagreeing.)

Be all that as it may, this detour through preference doesn't address the problem we started with. Again, I decided left foot first, but it's not that I thought I ought to step left first. I planned to step left first, but it's not that I thought stepping left was *the* thing to do. Even granting that my stepping was out of indifference, the problem is that we still have a psychological gap between the part of me that has (indifferently) settled how I act—I decided, I picked a strategy—and another part of me that reflects what I think would have been permissible. The problem isn't "how can we make sense of me embracing a plan that permits several options, and distinguish that from embracing just one." The problem is rather that it can be that I both (1) view several options as permitted and also (2) have decided which to select. One set of fact-plan alternatives cannot reflect these two facts about me, then. From a modeling point of view, we need a set of fact-plan alternatives to handle the state of me that corresponds to (1), and another set for the state of me that corresponds to (2). Evidently there's my state of normative judgment, which pronounces on what's permissible, and there's my decisional state, which seems more directly related to what exactly I end up doing. Maybe these two states have a similar logical structure, both to be elucidated with hyperplans, and maybe they tie into each other in intimate ways—for instance, perhaps (ideally) you only decide to do what you take to be permitted—but it is hard to see how they could shake out to be the same state.[14]

14 Discussing Gibbard (1990), Railton sees what is maybe a similar gap: "If there is an element of language that is purely action-guiding, I suspect it is closer to 'the thing to do' than to 'the rational thing to do', or to 'the

Once we grant the gap here, we can flag two ways that decision and normative judgment might intertwine. First, a normative view might itself be the result of decision. (Maybe "deliberation" is the word we'd more naturally use.) We speak of deciding, not just what to do but also what is permitted and required (of us, or of others). The result of a decision might be an intention to act, or it might be a normative judgment to the effect that a certain action is called for. When we talk about "deciding what to do," we perhaps usually mean to be talking about a state that results in an intention to act in some way. But we could also mean "deciding what is to be done," where the result of that is foremost a normative judgment, perhaps by an agent not yet resolved on a particular course of action. Second, if we think that one can only decide to do what one takes to be permitted (a view it seems Gibbard would find attractive, at least for agents in some sense ideal), then although deciding and normative judgment are strictly different things, a decision to act always goes with a normative judgment to the effect that the act is permissible.

Reconsider now:

(3) Holmes thinks he ought to pack.
(4) Holmes thinks packing is the thing to do.
(5) Holmes plans to pack.

Again, take it we have these truth-conditions for (4):

> For all $\langle c, h\rangle \in P$: h (c) permits only the options o in O_c that entail packing by the agent o is centered on

We were asking whether these truth-conditions are basically right, according to Gibbard, for (3).

If what we said in recent paragraphs about the gap between planning and normative judgment is on track, then the state ascribed in (5) is not the same as the state ascribed in (3), though they might be importantly connected. Now when Gibbard introduces "the thing to do" talk, he stipulates that it is to be expressive of decisions.[15] That stipulation

---

thing it makes most sense to do'" (Railton 1992, 966). See also Scanlon (2006). Engaging Railton, Gibbard seems to grant the possibility of a contrast between *being the thing to do* and *being the rational thing to do* (Gibbard 2003, 152), which sounds like the contrast between decision and normative judgment that I am asking about.

15 He writes: "Suppose, let me stipulate, the phrase works like this: to conclude, say, that fleeing the building is *the thing to do* just *is* to conclude what to do, to settle on fleeing the building. By sheer stipulation, then, the meaning of this phrase 'the thing to do' is explained expressivistically: if I assert 'Fleeing is the thing to do', I thereby express a state of mind, deciding to flee. I then proceed to ask how language like this *would* work. In the back of my mind, of course, is the hypothesis that important parts of our actual language do work this way. Mostly, though, I don't argue for this hypothesis; rather I ask whether the hypothesis is coherent

pushes us to see (4) and (5) as equivalent, and therefore it pushes us to see (4) and (3) as inequivalent—which would recommend the position that the truth-conditions of (4) are not the same as the truth-conditions for (3).

What's a charitable interpreter to say here? It certainly seems that Gibbard's proposal is at least that the truth-conditions of (3) *mimic* the ones we have stated for (4). Holmes is in a (hyper)plan-laden state of belief that reflects his normative judgments, his views about what is permitted and required in various situations. Model this with a set of fact-plan pairs N, which we hold in our minds for the moment separately from P. Then, roughly, (3) is true just in case:

> For all $\langle c, h\rangle \in N$: $h(c)$ permits only the options o in O_c that entail packing by the agent o is centered on

Is the normative state modeled with N exactly the planning state modeled with P, the related state of Holmes that we characterize when we say something like (4)? For the reasons reviewed, it seems we shouldn't say that, but as a textual matter, I am not sure how to read Gibbard on the question. Certainly, he wants to say that his notion of planning is not exactly the ordinary notion, for the reasons we reviewed in the previous two sections: it's an ideal model aimed at capturing the abstract structure of contingency planning. The point we're running up against right now, though, grants this, and still observes that there is a gap between this ideally conceived planning and ideally conceived normative judgment. Does Gibbard grant this point? If he does, then I don't know where he makes it with a reassuring level of emphasis. You've seen the quotes above tying planning and *oughts* together, which pull the other way. Consider however this passage:

> Reserve the term 'ought' as a quick way of saying "has most reason." The unqualified dictum I started with was this: to believe that one ought to do a thing—that one has most reason to do it—is to decide to do it. This Scanlon rejects, and rightly; it needs the qualifications I have just been stating, and which I stated in the book. My slogan to a closer and more verbose approximation might be this: to believe that a person ought to do a thing is to require it of oneself for the hypothetical case of forthwith being in that person's precise situation.
>
> (Gibbard 2006, 731–32)

The unqualified dictum explains a state of normative judgment in terms of a decisional state not specified with the help of normative vocabulary (viz., deciding to do a thing), whereas the more precise statement explains the state using normative vocabulary (invoking talk about what one "requires" of oneself). Maybe this favors the reading that Gibbard would

---

and what its upshots would be. Only much later in the book do I turn to our actual everyday thoughts and ask if the shoe fits" (Gibbard 2003, 8).

say that (3) doesn't shake out to be quite the same as (5), though there is a deep affinity in underlying structure.[16]

To put the question again: is the normative state modeled with *N* the same as the planning state modeled with *P*, the related state of Holmes that we characterize when we say something like (4)? Since I want Plan A to align with Gibbard, and since I seem to find conflicting tendencies in what he says, my official stipulation will be that Plan A is agnostic on this crucial question.

Anyway, the thrust of it is this. Belief states are plan-laden in at least the sense that they have a structure to be articulated with hyperplans as above: they are states that cut the space of fact-plan alternatives. One's normative judgments are explained essentially in terms of this plan-laden structure—specifically, along the lines of the truth-conditions recently set down for (3). The not-fully factual character of these judgments traces to the conditions they place on the fact-plan alternatives they rule out, conditions partly a function of the (nonfactual) hyperplan dimension of those alternatives. We have here an alternative to the realist idea that states of normative judgment are states that represent normative properties or facts.

So we have, finally, a view on the table about what Holmes's state of mind is like when (3) is true, according to Gibbard.

1.5. Gibbard's Metatheory

Besides his novel formal model of states of normative judgment, Gibbard offers a distinctive kind of philosophical gloss on, or metatheory for, his model.[17] The gloss is interestingly different from the sort of gloss we find from other theorists in the tradition of decision-theoretic/possible worlds modeling, for instance Lewis or Stalnaker, who Gibbard otherwise looks to be rather continuous with. Those theorists would say that to believe that grass is green is to be in a state that rules out possible worlds wherein grass is not green, and they would hold that these possible worlds that the state rules in or out are, at least eventually, explicable independently of intentional mental states—they are not themselves explained in mental terms, not themselves metaphysically dependent on anything mental (Lewis 1994; Stalnaker 1984). Gibbard is skeptical, however, that a nonintentional notion of modality is available for this purpose (see, for instance, Gibbard 2012, 277). He prefers not to attempt reduction. He explains the "possibilities" that mental states rule out *as themselves mental states*. He has this view quite apart from his proposal to model in terms of hyperplans—this is how he would want to think about an ordinary, hyperplan-free possible worlds model of belief. Gibbard will agree that to think grass is green is to "rule out a possibility," but fundamentally he will explain this state as the state of ruling out another mental state, the state of rejecting grass is green—the

16 The "qualifications" Gibbard alludes here to have to do with the first two problems we discussed, though, so we shouldn't read this quote as directly animated by the problem we're focused on in this section. So I don't put too much weight on this passage.

17 In this section, I draw on some of the ways I put things in Yalcin (2018).

mental state of *rejecting that grass is green* is the "possibility" ruled out. Similarly, he would describe the state of believing that grass is not green as "disagreeing with believing" that grass is green—as rejecting believing grass is green. The centered worlds of the model are interpreted by Gibbard as maximally opinionated states of (factual) belief—they are not, as Lewis or Stalnaker would have it, maximally specific ways things might have been, understanding the relevant modality as fundamentally nonmental. Likewise, he glosses the hyperplans of his model as maximal states of decision—the states that idealized "hyperplanners" would be in.

This will seem to some like a dangerously tight circle: it is a model of mental states whose basic resources for modeling mental states include mental states. Out is the idea of characterizing propositional attitudes as relations to contents, if the latter are understood in a traditional way as some sort of mind-independent abstracta—sets of possibilities, for instance, as Lewis and Stalnaker would have it. We don't arrive on this picture at a conception of mental content giving us a handle on it in other terms; understanding must somehow come from the whole system. Schroeder (2008a) argues that all this renders Gibbard's view explanatorily deficient. Does Gibbard's view leave it mysterious what makes it the case, when it is the case, that one content is incompatible with another? For example, the state *believing grass is green* and the state *believing grass isn't green* "disagree" with each other; they are in logical tension. In virtue of what? Not, says Gibbard, in virtue of their having incompatible contents. What, then? Gibbard says he has no further explanation of such disagreement facts; he takes them as primitive.

Gibbard argues that this is not a disadvantage, however, because the orthodox line of explanation invokes "substantial, unexplained truth, eschewing any minimalist explanation of truth" (Gibbard 2003, 74). He expands:

> Proceeding this way might seem to be philosophical theft. The scheme amounts just to helping ourselves to the notion of disagreeing with a piece of content, be it a plan or a belief. A negation, we say, is what one accepts when one disagrees—and this explains negation. Now I wish, of course, that I could offer a deeper explanation of disagreement and negation. Expressivists like me, though, are not alone in such a plight. Orthodoxy starts with substantial, unexplained truth, eschewing any minimalist explanation of truth. I start with agreeing and disagreeing with pieces of content, some of which are plans. It's a thieving world, and I'm no worse than the others.
>
> (74)

In his more recent book (Gibbard 2012), he calls orthodoxy "Fregeanism," and puts the problem, or at any rate one key problem, for the view like this:

> Not all impossibilities make for entailment—or at least they don't make for the kind of entailment that, with enough conceptual competence, a thinker can recognize. It is this kind of entailment, inconsistency, and the like, we might well think, that the Fregean needs to explain, or the whole Fregean project fails. The thought I'M DRINKING WATER doesn't recognizably entail I'M DRINKING H_2O, unless one knows the chemical composition of water. The

> thought I'M USING THIS LECTERN likewise needn't recognizably entail I'M USING AN ORIGINALLY WOODEN LECTERN, even if it couldn't be that this very lectern was originally made of anything but wood. The metaphysical impossibility of a putative state of affairs needn't yield its conceptual impossibility.
>
> (277–78)

The hypothetical maximally opinionated agents at the bedrock of Gibbard's theory aren't explained in nonintentional terms. But Gibbard says orthodoxy requires a logical space for marking epistemic or conceptual distinctions that go beyond metaphysically possible distinctions—a space where the possibilities must be intentionally specified. So, he thinks we have at worst parity between the views here.

There is one more component to Gibbard's meta-theoretical reflections. He styles his preferred philosophical gloss on his model as part and parcel of an expressivist approach. He says things like this:

> The orthodox explain disagreeing with a claim as accepting its negation, whereas I go the other way around: I explain accepting the negation as disagreeing with the claim. Agreement and disagreement are what must ground an expressivistic account of logic.
>
> (Gibbard 2006, 73)

One gets the impression that to go expressivist in the relevant way about normative judgment, it is not enough to embrace a model of this state of mind that renders this thinking as not (just) a matter of the way one represents the world to be. It is not enough to model with fact-plan alternatives, tracing the not-fully factual character of normative language at the modeling level to their interactions with hyperplan structure. One must also adopt a certain foundational gloss on this model—an "attitudes-first" metatheory like that just described. The fact-plan possibilities of the model must be understood intentionally, as idealized "hyperdecided" states of mind; and the relations of (dis)agreement between them and between less-than-maximally decided states are to be taken as primitive.

So the picture of normative judgment I've just sketched is Plan A. Most prominently Plan A provides a formal model of normative states of mind and a certain kind of philosophical gloss on that model. I mean Plan A to be a highly selective take on Gibbard's picture of normative judgment, with an emphasis on the role of hyperplans in the story that, I should stress, is quite out of proportion to Gibbard's own emphasis. I have skirted over many nuances.[18] But if I haven't got Gibbard's view just right, I hope you will agree that Plan A is very Gibbardian, and worth assessing.

---

18 At several turns, especially in discussing "facts" and "the way the world is," I have ignored qualifications whispered to me by the quasi-realist devil on my shoulder, reminding myself that Gibbard's talk of *prosaically factual* belief is suggestive of the sort of grip on Reality my discussion seems to presuppose.

2. Plan B

Now to Plan B. Plan B is like Plan A in explaining normative judgment as plan-laden—that is, modellable with the help of things like hyperplans—and as not in the business of representing normative facts. It is an expressivist view. But it differs from Plan A at the foundational level in two key ways. First, Plan B embraces what Gibbard styles as the "orthodox" or "Fregean" metatheoretic attitude toward the formal model. Second, Plan B explicitly says that normative judgment isn't planning, understood as deciding how one will act, though it allows that normative judgment may be formally analogous to planning.

Let me now go through these differences. I don't attempt any full-throated defense of Plan B against Plan A here. My real aim is just to bring out Plan B as an option, so we have a sense of which choices on the expressivist road are separable and which come together.

2.1. Orthodox Metatheory

It is one thing to model as Gibbard does, and another thing to interpret the model as Gibbard does. You can model as Gibbard does, without interpreting the model as Gibbard does. It would be a mistake to reject Gibbard's model merely because one isn't ready yet for his distinctive metatheory. We have to separate these things.

Plan B says that when we are thinking about the ingredients of our formal model, we don't understand possible worlds, centered worlds, hyperplans, or fact-plan possibilities as themselves mental states; we don't construe the model as explaining the content of a mental state by reference to further contentful mental states; we don't treat it as a brute fact that the state of believing *p* is in logical tension with the state of believing *-p*. Possible worlds, we take it, are not maximally opinionated mental states, though we might use a possible world to model a state of maximum opinionation. Plan B says: go back to what we might think of as the more typical way of understanding possible worlds modeling, familiar from (e.g.) Lewis (1979, 1994) and Stalnaker (1984)—an approach Gibbard would group with the "orthodoxy."

As noted above, Gibbard gives some indication that he thinks that this approach really smuggles something it claims to explain through the back door—the truth-conditions of a sentence must be said to discriminate between "conceptual" possibilities, possibilities that are intentionally specified (Gibbard 2012, 277–8). This is part of why he holds his view to be not less explanatory than orthodoxy. If this is at the crux of the debate, I wish Gibbard had given the topic more airtime. While I doubt that the orthodox view does require an intentional conception of possibility—there are certainly various well-known ways to ensure, for example, that I'M USING THIS LECTERN doesn't entail I'M USING AN ORIGINALLY WOODEN LECTERN, and so forth, without such a conception—this is a suboptimal place to debate that large issue.[19]

19 I objected to Gibbard's preferred metatheory in Yalcin (2018), but I now think I didn't adequately address the appendices of Gibbard (2012), in particular the worry about whether the logical space assumed by orthodoxy can really be nonintentionally characterized.

Gibbard is playing defense, not offense. I'm satisfied to observe that the orthodox position is not subject to a new style of objection, one that Gibbard's own view is somehow immune to. Either the score is tied—both Plan A and Plan B both traffic in partly in unexplained intentional notions—or Gibbard is wrong and Plan B can be carried out with a nonintentional notion of logical space. Since I am only plumping for the view that Plan B deserves airtime along with Plan A, I can live with a tie for the purposes of this paper.

There are many questions one could raise about what it is to "explain" or "ground" an abstract model. One basic sort of question we could ask about a simple possible worlds model of belief is this:

> In virtue of what sort of facts is an agent in a state of belief well-modeled with such-and-such possible worlds content rather than some other possible worlds content?

It seems worth saying that this sort of question has hardly gone unaddressed by those who enjoy the orthodoxy style of model-interpreting. The basic issue here was the stuff of much of the philosophy of language and mind of the eighties and nineties on the metaphysics of content. Restricting attention to the broadly decision-theoretic modeling tradition Gibbard departs from, Stalnaker (1984) and Lewis (1994) both directly address this question about how facts of content should be understood to metaphysically depend or reduce to facts of another stripe. Their views differ in important ways, and resist easy summary. But the point I'm stressing is just that there is not a blithe indifference, or a failure to address the question, on the side Gibbard styles as orthodox. Instead, there's a nontrivial literature. To say that "Orthodoxy starts with substantial, unexplained truth" in the face of this work seems a touch glib—though maybe I'm missing what sense of "explanation" is intended.

There's a whole debate to have about this—elsewhere. For now, let me just repeat the point that if we truly are in the thieving world Gibbard says we're in, then Plan B and Plan A are at worst on par in the relevant respect, and therefore Plan B remains in the running as a view worth talking about. While I don't think Team Plan B should concede the antecedent of this conditional, I won't argue it now.

But can one embrace an orthodox attitude at the metatheoretic level and still be expressivist? Isn't Gibbard's style of metatheory constitutive of expressivism? Isn't "explaining in terms of mental states" what it is to be expressivist?

To repeat something I tried to say in Yalcin (2018), I think it is a mistake to take Gibbard's distinctive kind a metatheory to be the hallmark of an expressivist approach. I suggest we separate two ways of understanding how an expressivist might be described as "explaining in terms of mental states." One way is in the vein of Gibbard's metatheoretic reflections: on this way we "explain" the "possibilities" the content of a mental state eliminates as themselves mental states, and take (dis)agreement between mental states as primitive. But there is second way of understanding how an expressivist view might bring mental states into

explanation. This is the idea that expressivism is characterized by a strategy we could call *psychological ascent*. Here is a crude way to put the recipe:

> **Expressivism by psychological ascent**. To go expressivist about φ, first reject the question "What is the world like when φ is the case?" Replace it with the question: "What is the state of mind of accepting φ like?" Answer this question in such a way that the state of mind is understood as not tantamount to ordinary factual belief that something is the case—as not representing φ-facts. Then approach the target discourse from this perspective: find a way to elucidate the semantics and pragmatics of φ consistent with the idea that accepting φ is being in this not-fully-factual state of mind.

Psychological ascent is a pathway for stating an expressivist view. When I look at the various twentieth-century works in the expressivist genre, it looks to me if anything like the standard pathway. One can follow this path without also taking on Gibbard's style of metatheory.

The expressivism about epistemic modality I have defended elsewhere (Yalcin 2007, 2011) is expressivist in this sense—as is Gibbard's classic work on indicative conditionals (Gibbard 1981) when viewed at a natural angle. It is famously difficult to say what the world has to be like to make an indicative conditional true, or to make it true that something (epistemically) might be the case. The expressivist mode is to table that question and ask instead about the state of mind that goes with accepting what the sentences say. It says: don't start with questions like this:

> What is the world like when 'It might be raining' is true?
> What is the world like when 'If the marble isn't under cup A, it's under cup B' is true?

But instead with questions like this:

> What is to think it might be raining?
> What is to think that if the marble isn't under cup A, it's under cup B?

Or, semantically ascending, with questions about attitude ascriptions, like this:

> What is the world like when '*A* thinks it might be raining' is true?
> What is the world like when '*A* thinks that if the marble isn't under cup *X*, it's under cup *Y*' is true?

This kind of expressivist ends up with a conception of the truth-conditions of the target attitude ascriptions that does not resolve them into relations to propositions characterizing the world as being some way or other. (For instance, they might say that when *A thinks it might be raining*, that's because *A* is in a state of mind leaving rain possibilities open,

and not because *A* believes-true some *might*-proposition; or they might say that that when *A thinks that the marble is under cup Y if it's not under X*, that's because *A* has high credence in *Y* conditional on not-*X*, and not because there is some conditional proposition *A* believes.) Thereby this expressivist dissolves the original questions, the ones in search of facts to be expressed by the original epistemic modal sentences.[20]

Plan B takes this kind of approach to normative language. Don't ask:

> What is the world like when 'Holmes ought to pack' is true?

Instead ask

> What is it to think Holmes ought to pack? Semantically ascending: what is the world like when 'Holmes thinks that he ought to pack' is true?

Plan B, like Plan A, gives an answer that explains the ascription as not ascribing to Holmes a prosaically factual belief. In particular, it's not that Holmes locates himself in a world that includes the normative fact that he ought to pack. Rather, it's for Holmes's state of mind to be plan-laden in the right sort of way.

Obviously, "psychological ascent" is meant to remind you of "semantic ascent." They have a lot in common. Semantic ascent can be helpful in metaphysics. When two sides disagree about what some aspect of reality is like, that will often make for a difference between the two sides in respect of what they can say their terms represent. That in turn can make it hard to state the point at issue between the sides in a neutral way. Retreat to talk of sentences and what those sentences say can be a way of preventing the sides of the debate from talking past each other. But stereotypical applications of semantic ascent still do involve the assumption that the target sentences (the sentences which are the locus of ascension) have truth-conditions, that they say something or other about reality. The two sides might disagree about what is represented by sentences containing the terms key to their dispute—semantic ascent is just what will make that perspicuous—but in the standard examples, the two sides at least agree that the target sentences are in the business of describing the world as some way or other. Take for instance the "ontological debate" about whether Pegasus exists (Quine 1948). The Pegasus believer explains the meaning of 'Pegasus' in terms of a certain winged horse of their ontology, suppose. The Pegasus denier, lacking that thing in their ontology, says something else. We'd like to say that the two sides disagree about whether Pegasus exists, but annoyingly, it seems the sides must understand the meaning of a sentence like 'Pegasus doesn't exist' in different ways, owing to the different ways they will each explain the meaning of 'Pegasus'.

20 Though, of course, the expressivist faces new questions in semantics and pragmatics, about how to think systematically about the communicative roles of these sentences in the absence of factuality. I discuss this especially in Yalcin (2012, 2018).

Retreating to talking about the conditions under which 'Pegasus doesn't exist' is true can help to isolate the metaphysical dimension of the debate from the semantic dimension. But note that in standard cases like this, the two sides agree that the target sentences have truth-conditions; at worst they differ on what the truth-conditions are. Both the believer and the denier do think 'Pegasus doesn't exist' characterizes the world as being some way or other, though they may differ in what way that is.

Psychological ascent is like semantic ascent in that we are retreating, broadly speaking, to talk of things with intentional properties (and which both sides in the debate are happy to recognize), but it is a move that is free of a presupposition that the target sentences are factual, in the sense of characterizing the world as being some way. The expressivist about normative discourse wants to reject questions like "What feature of reality corresponds to its being wrong to break promises?" The expressivist is not better served by the question we'd get by semantic ascent: "Under what conditions would 'It is wrong to break promises' be true?" The question "What is to think that it is wrong to break promises?" brings us to level of description where the expressivist can start to unfold their view.

The strategies of psychological ascent and semantic ascent can be fruitfully combined. The question we get by psychological ascent—"What is to think it's wrong to break promises?"—is a metaphysical question as susceptible to semantic ascent as any other. Semantically ascending, we get the question "What is it for something like '*A* thinks it's wrong to break promises' to be true?" My Plan B expressivist favors this kind of two-step ascension. It is clarifying to get at the view this way, because as we've seen, our expressivist's abstract model of normative states of mind does not map into attitude ascription in a trivial or linear way. You don't just disquote to articulate the truth-conditions of attitude ascriptions; the right-hand side is substantive.[21] Mentioning rather than using the relevant attitude verbs prevents confusing slides between the modeling language and ordinary language. This perhaps reveals why I elected to place such emphasis, in the preceding sections, on framing Gibbard's view as a position about the truth-conditions of attitude ascriptions that embed normative vocabulary.

I don't mean to suggest that psychological ascent isn't part of Gibbard's own approach. He gives a theory of normative judgment, not normative facts, and the theory of normative judgment is supposed to dispel the need for normative facts; he makes the move of ascending to the psychological level ultimately to dispel philosophical perplexity about a seeming domain of facts. That embodies the core expressivist pattern, and the pattern my views about epistemic modality also fit. Mixed into Gibbard's development of his theory is a separable body of metatheory, though. We do better to unmix these issues, I suggest, and consider the metatheory questions as different ones.

Sometimes expressivism is conceived of as something like a special kind of semantic theory, one that identifies the compositional semantic values of sentences with mental

21 Gibbard understands his theory as combinable with a thoroughgoing minimalism about truth (e.g., Horwich 1998). Plan B, by contrast, is not so combinable.

states (Blackburn 1993; Rosen 1998; Schroeder 2008b; Charlow 2015; among others). One might understand it like this because one thinks that adopting the strategy of psychological ascent in the expressivist's way must lead, on its most plausible development, to such a semantic theory. If one has such a view, then one will be skeptical about unmixing the issues as I suggest. I argue against this view directly in Yalcin (2018). Less directly, I think the case of epistemic modality (Yalcin 2007, 2011, 2012) already illustrates that one might psychologically ascend in the expressivist's characteristic way without also signing up for a nonstandard kind of semantic theory, or a nonstandard way of interpreting standard semantic models.[22]

2.2. Normative Judgment Not Reduced to Planning

Plan B takes a stand on the thing Plan A was agnostic about. It agrees that planning and normative judgment might be alike in both calling for something like hyperplans in models of their content. But Plan B has no pretension to reduce normative judgment to planning. We earlier said, following Bratman, that planning is its own thing, not to be reduced to something else, like some combination of belief and desire. Now Plan B says the same thing about normative judgment. Normative judgment—or more specifically, one's views about what is permissible to do in various situations actual, hypothetical, and counterfactual—is its own thing, too, not to be reduced to something else, like some combination of belief, desire, and intention or planning. We can (and will) of course still attempt to *offer a model* of the state, and we can still aim to *offer an interpretation of the model* in an orthodox vein, as we might do for a decision-theoretic model of belief and desire by (for instance) clarifying the functional role of the state. It is just that we will theorize under the assumption that normative thinking is not identical to planning.

In this I seem to be on the same page as Scanlon (2006). Discussing Gibbard (2003), Scanlon writes:

> The difficulties I have described do not arise from the expressivist strategy of giving a (nonreductive) psychological account of normative attitudes, but rather from the attempt to base this explanation on the single notion of a plan. My suggestion is that Gibbard's strategy could be more plausibly carried out if he were to broaden the range of notions that figure in his psychological explanation. These will include notions of an explicitly normative character, such as the idea of seeing something as a reason. But we can distinguish here, just as Gibbard proposes, between the normative content that these notions have when one employs them in deliberation and their descriptive employment in a psychological account of deliberating agents.
>
> (726)

22 I am inclined to read the (otherwise diverse) expressivist stylings of Stalnaker (2014), Santorio (2016), Starr (2016), Willer (2017), and Moss (2018) as compatriot views here.

I won't put the sort of normative state Plan B models as the state of "seeing things as reasons"—instead I want, again, to talk about one's views about what is permissible to do in various situations actual, hypothetical, and counterfactual—but the relevant point is that one can pick out the state using normative vocabulary without apology, and compatible with embracing an expressivist pattern of explanation of some target discourse in terms of a model of that state.

Would doing this leave something unexplained? It is good to compare the situation here to the situation with (prosaically factual) belief and preference. In giving an explanatory model of these states in the broadly decision-theoretic style, we don't have to claim to be reducing these states to states of other kinds in order for the modeling project to seem like progress. (One might have a reductive aim—to reduce everything to betting dispositions, for instance—but that is hardly a prerequisite for taking models of this sort seriously.) We limn the structure of these states with our abstract modeling tools and we say how the elements of the model are supposed to connect to each other and to other features of reality. One could argue about how illuminating the abstract models of agents pursued in things like decision theory are, but it would be strange to complain that the defect in these views is that they don't come with a recipe for reducing the states they traffic in to something else.

Perhaps unsurprisingly, I am not sure to what extent Gibbard would disagree with this part of Plan B. His "possibility proof" was primarily aimed at showing that the concepts needed to handle planning and disagreement in plan recommend a path for thinking about normative concepts; and one could succeed in showing this without reducing normative judgment to planning.[23] He's clearly saying that a formal entity approximating the structure of a hyperplan is of use for modeling both planning states and normative thinking, and on that Plan B agrees. Just as a textbook possible worlds theorist might hold that two distinct kinds of intentional states, such as belief and (say) imagination, can be helpfully modeled by appeal to a formal object of a single kind (a sets of possible worlds, say), so a theorist might hold that planning and normative judgment, while different kinds of mental states, both have at an abstract level a structure that is helpfully modeled using hyperplans.

In any case, the thing Plan B sticks its neck out on is normative thinking, not planning.

3. Plan B+

Come back yet again to this sentence:

(3) Holmes thinks he ought to pack.

We said Gibbard models it like this: (3) is true iff

23 "A fully consistent planner in my sense of the term, I tried to show, would in effect deploy concepts that work much as a non-naturalist would think that normative concepts work" (Gibbard 2006, 735).

> For all ⟨c, *h*⟩ left open by *N*: *h*(c) permits only the options *o* in O_c that entail packing by the agent o is centered on.

Again, *N* is a set of fact-plan possibilities determined by Holmes's state of mind at the evaluation world for (3).

Now, the most basic Plan B position embraces the ideas of the last section (pass on Gibbard's metatheory, embrace a metatheory of a more orthodox kind, psychologically ascend and then semantically ascend to state the position using a nondeflationary notion of truth, and take normative judgment to be its own thing) but keeps this specific modeling proposal for ought-thoughts—it keeps the truth-condition above. That is, this most basic version of Plan B disagrees with Plan A not on the modeling questions but on the philosophical gloss on the model. The two plans agree about how to model normative judgment with hyperplans.

But now that we've come this far, it is interesting to consider a version of Plan B that departs from Plan A also in some of the formal modeling respects. The idea of a hyperplan is a very interesting one, and it seems to me to open up some fruitful questions on the modeling side of things. Metatheoretic issues tend to get most of the airtime in discussion of Gibbard's approach, but the formal model he gives is itself worthy of investigation quite apart from those issues. Now I want to explore some modifications to Gibbard's modeling proposal, to do with how exactly hyperplans figure in the story. So, I will be exploring ways of giving truth-conditions to (3), which invoke hyperplans but not quite in the way Gibbard does. The ultimate formal model of normative judgment we will end up with is still very much in the vein of Gibbard's, but it will come apart from the letter of Gibbard's own version in key ways. If we add the modifications of I'm about to suggest to the ideas of the last two sections, let's say what we end up with is Plan B+.

If your main interests were on the metatheoretic side of things, this is a safe time to check out of the paper.

3.1. Hyperplans as Information-Based

Gibbard models an agent's state of belief-and-normative-judgment with a set of fact-plan possibilities. Question: why exactly are centered worlds and hyperplans packaged together in this way? Let's ask:

> Why model an agent's state with a single set of fact-plan possibilities, rather than with two separate sets—a set of centered worlds (modeling ordinary factual belief) and a set of hyperplans (getting at the agent's normative views)?

To think about this, consider the following kind of example. Suppose that ⟨c, *h*⟩ and ⟨*c′*, *h′*⟩ are the only two fact-plan possibilities compatible with Holmes's state of mind. Suppose also that Holmes thinks that he ought to pack. Now we could ask: what does *h* permit relative to *c′*? And what does *h′* permit relative to *c*?

For all Gibbard says, the following could be the case:

- $h(c')$ forbids packing
- $h'(c)$ forbids packing

Suppose that is so. So Holmes thinks he ought to pack (this holds throughout his plan-laden state), and yet his state of mind leaves open plans that say not to pack relative to centered worlds that are bonafide doxastic alternatives for him.[24]

This not a contradiction, but it is a bit puzzling and counterintuitive. We might be able to make sense of this if the different centered worlds compatible with Holmes's state of mind potentially fixed different sets of options. However, it is hard to understand how they could, given the way Gibbard talks about options:

> An occasion, as I have characterized it, contains much that the agent has no way of knowing, but one's plans must respond to features of the occasion available to the agent. Alternatives must be subjectively characterized, so that the same alternatives are available on subjectively equivalent occasions. And a plan must permit the same alternatives on subjectively equivalent occasions.
>
> (57)

> "Keep away from radioactivity" would surely be a good part of a plan for living, if only we knew how to tell what's radioactive—but it's not much help if we don't. Plans, it seems, must be couched in terms of features that we can recognize: features of contingencies and features of options. Both of these must be available to the person who follows the plan. "Buy low, sell high" is no plan we can implement. Plans must be couched in terms whose application we can recognize.
>
> (99)

So, an option available to an agent in a situation—a hidden door behind the curtain, say—is no option at all if the agent centered at this situation does not recognize it as such. Gibbard does not to my knowledge define "subjective equivalence," but on one not unnatural take, one's doxastic centered possibilities are subjectively equivalent—these centered worlds are, for all the agent can tell, who they are. But if so, then a hyperplan cannot issue distinct verdicts on any pair of centered worlds left open by given a state of belief, because these will present the same options. This would remove some of the motivation for modeling

24 Of course, if the options are fixed by what is doxastically possible for a centered agent, then if an agent has exactly the same doxastic possibilities open at each of her doxastic possibilities, then for all centered worlds c, c' in any prosaically factual belief state B, $O_c = O_{c'}$. So, the problem won't arise. But then it's not clear why we need to model in terms fact-plan possibilities at all, as contrasted with the alternative model I am about to suggest.

normative thinking and factual believing together in terms of a single set of fact-plan alternatives: after all, if a hyperplan cannot issue distinct verdicts on any pair of centered worlds left open by given a state of (factual) belief, then it doesn't matter which hyperplans a plan-laden belief state pairs with which centered worlds. You might as well have two separate sets, a set of centered worlds (factual belief) and a set of hyperplans.

Moreover, although a hyperplan is a function on centered worlds, it appears a hyperplan is only sensitive to one feature of a centered world: the subjective predicament of the agent at the center. That suggests we might just as well view hyperplans as functions on subjective predicaments. If we think of a subjective predicament as representable by a body of information—a set of centered worlds—then a natural idea would be to reconstrue hyperplans as functions on sets of centered worlds, rather than on centered worlds.

Let me assemble these considerations into a concrete proposal. Suppose now that hyperplans are functions on states of information rather than centered worlds: they are *information-based*. Associate Holmes with two sets: B (a set of centered worlds—his *doxastic state*, a kind of state of information) and H (a set of hyperplans capturing his views about what is permissible relative to various predicaments). Assume that any state of information, like Holmes's belief state B, fixes a set of options O_B. Take it the hyperplans in H speak to what to do given the options fixed by the whole of B. No longer do we have any special pairing of particular hyperplans with particular doxastic alternatives. Then the idea is to use these components to give Plan B+'s truth-conditions for (3), as follows:

> For all h in H: $h(B)$ permits only the options in O_B that entail packing by the agent the option is centered on.

Informally, this says that Holmes thinks that he ought to pack when all of the hyperplans left open by his normative state require packing when evaluated relative to his state of factual information. Observe that on this proposal, the truth of (3) owes both to B and H; again, we have a kind "hybrid" view.

Let me offer two reasons to be interested in this way of modeling *ought* thoughts.

3.2. *Nonpersistence of* Ought *Thoughts*

First, this proposal respects the apparent nonpersistence of *ought* thoughts: the fact that *ought* judgments seem capable of coming and going under the impact of strict information gain. Suppose the following are the case at t:

(6) Holmes thinks he might still be able to make the train.
(7) Holmes thinks that if he might be able to make the train, he ought to pack.
(8) Holmes thinks that if it's too late to catch the train, it's not the case that he ought to pack.

So of course, at t,

(3) Holmes thinks that he ought to pack.

Now at $t + 1$ Holmes learns it's too late to catch the train. So, at $t + 1$:

(9) Holmes thinks that it's not the case that he ought to pack.[25]

It seems:

- From the point of view of his purely factual belief state, Holmes underwent a strict information gain from t to $t + 1$.
- From the point of view of Holmes's pure normative state—his views about what is allowed in various possible situations—Holmes didn't change from t to $t + 1$.

Of course, Holmes goes from thinking he ought to do something to thinking he needn't. Putting it that way makes it look like he underwent a normative change of mind. But I think we can make out a level of description where we can say that he didn't really undergo any change of normative opinion. What changed is his view about the world, and hence which aspect of his (stable, unchanging) normative view speaks to the situation he takes himself to be in.

The truth-conditions for (3) supplied by Plan A do not get this right. If Holmes's state at t is such that each fact-plan possibility it leaves open calls for packing, then no shrinking of this set moves Holmes to a new state where packing isn't the thing to do. Some kind of plan-laden belief revision happened; the state was replaced with an entirely new set of fact-plan possibilities. But this seems an unnatural way to model a case of strict gain in information.[26] Plan B+, by contrast, seems to fit the facts. (The idea that *ought* thoughts might come and go under the impact of strict information gain is a main theme of Kolodny and MacFarlane (2010), and it is built into their semantics for deontic modals.)

One way to put pressure on Plan A here is to ask: what supposed condition on fact-plan alternatives obtains throughout Holmes's state if (8) is true? It is not easy to discern a natural answer. Do we check what is to be done relative to fact-plan alternatives in Holmes's state where it's too late to catch the train (the alternatives rendering the antecedent true)? But if we are assuming Holmes thinks he ought to pack and assuming Plan A's conception of what that means, then we already know such antecedent alternatives require packing, and hence are in tension with the consequent of (8).

25 Hold fixed the time the packing is supposed to be happening, according to Holmes, across (3) and (9)—the same thing thought is at issue.

26 One could try appealing to the changes in Holmes higher-order beliefs that take place between t and $t + 1$. How best to think about that will interact with the question how *de se* updating gets handled.

Plan B+, in contrast, would allow one to say the following. If (8) true, it is because when you start with Holmes state of information B and strictly add to it the information that it's too late ($B \cap$ TOO LATE), all of Holmes's hyperplans are such to prohibit packing relative to the options this updated state of information fixes.[27]

An important thing to notice here is the following: if you take Holmes's doxastic state and intersect it with the set of centered worlds where it's too late to catch the train, the resulting set is not actually a possible belief state.[28] Yet it does not seem incoherent to suppose that this body of information nevertheless fixes a set of options. Reflection on conditional thinking seems to suggest we need the idea that a set of options can be fixed by a state of information that is not itself a possible doxastic state. This seems to encourage to move beyond the Plan A idea that the options that hyperplans interface with are always fixed by some possible subjective state of an agent.

3.3. Other-Locating Deontic Thinking

A second consideration in favor of Plan B+'s conception of the truth-conditions of things like (3) comes in when we think about how to model one's views about how others are permitted or required to act. We could separate two kinds of such thinking:

- **Self-locating deontic thinking**. Thinking what is to be done as if in the subjective circumstances of another (taking on their beliefs, desires, and all).
- **Other-locating deontic thinking**. Thinking what is to be done when in the position of another, but in the world as one (not the other) takes it to be.

Other-locating deontic thinking seems to me hard to model along the lines of Plan A. But there is something natural we can say on the information-sensitive approach. A case will help us fix ideas (discussed in Yalcin 2012):

> John's puppy has been poisoned; so too Niko's kitten. There is only enough antidote left to save one of their pets, but the price is too high. Jay know all this. His view is that the thing for John to do, given his situation, is to steal the last of the antidote and give it to his puppy. Jay also thinks that the thing for Niko to do, given his situation, is to steal the last of the antidote and give it to his kitten. That is:
>
> (10) Jay thinks Niko ought to steal the remaining antidote and give it to his kitten.
>
> (11) Jay thinks John ought to steal the remaining antidote and give it to his puppy.

---

27 Of course, one eventually requires a theory of conditionals that jives with this conception of the truth-conditions of (8). See, for example, Yalcin (2007), Kolodny and MacFarlane (2010), and Gillies (2010).

28 Assuming any relation of doxastic accessibility obeys the axioms K and J (where J: $\Box\varphi \rightarrow \Diamond\Box\varphi$). The trouble is that $B \cap$ TOO LATE of course entails TOO LATE, but the centered worlds in B $\cap$ TOO LATE are centered on an agent that does not believe TOO LATE. Thus an agent in the putative belief state $B \cap$ TOO LATE would (1) believe TOO LATE and yet (2) believe she doesn't believe TOO LATE.

Take it Niko and John can't *both* steal the antidote, and Jay knows that. But Jay's view seems coherent. How to capture this?[29]

Suppose H is Jay's normative state and B is Jay's ordinary factual-but-self-locating doxastic state. Let B_N be the state that comes from B by shifting all the centers to Niko.[30] Likewise, B_J comes from B by shifting the centers to John. Then we can say the following:

(10) is true iff for all h in H: $h(B_N)$ permits only options in O_{B_N} that entail antidote-stealing for Niko's kitten.

(11) is true iff for all h in H: $h(B_J)$ permits only options in O_{B_N} that entail antidote-stealing for John's puppy.

These can both be true. The point here is that Plan B+ shows us how Jay's state of mind is coherent.

A notable feature of the story, which we saw already with conditionals at the end of the last section, is that B_N and B_J are not even possible belief states (see fn. 14). Still, it seems natural to consider these "supposable" bodies of information as each fixing a set of options. We seem to have an ability to think what to do relative to bodies of information that are only hypothetically entertainable—bodies of information that could never even possibly correspond to a belief state.

This is a good time to ask: is there not also a Plan A+? Couldn't we package this newfangled conception of hyperplans with Gibbard's original style of metatheory? I am not so sure, and the reason is to do with the notable feature just referenced. Gibbard's metatheory wants to explain hyperplans as the states of maximally decided planners—hyperplanners. This aspect of his metatheory is part of what drives him to model with fact-plan possibilities. But it is hard to see what would correspond in his metatheory to the information states needed by Plan B+, in particular the ones that don't correspond to any possible doxastic state. One of the lessons of the cases that motivate Plan B+ is that not all thinking what to do can be theorized in terms of hyperdecided states being ruled in or out.

4. Closing

I hope those who rejected Plan A because of its metatheoretic picture, or because of its seeming ambition to reduce normative thought to planning, are interested to see a related view, Plan B, without these features. Plan B fits into a broadly 'representationalist' conception

29 On the face of it, this would seem to be a challenge for a textbook deontic logic, one appealing to the idea of *what is true in the worlds viewed as ideal* according to the agent. In the worlds ideal according to Jay, is it Niko or John who steals the antidote?

30 Or Jay could have an individual concept or role associated with Niko, and we could use that to find the person Jay takes to be Niko at his doxastic possibilities. I skip over this complication.

of the mental, though expressivism is sometimes viewed as a competitor to that position. I hope also that Plan B+ draws out the point that among those who enjoy modeling with hyperplans, there remains nontrivial space for intramural debate about the details—and that debate can help to inform metatheoretic questions.

One thing I haven't discussed is the Frege-Geach problem. I have focused on models of normative judgment and the ways these can be philosophically glossed, and not on the compositional semantics or pragmatics of normative language. But I often put Plan A and Plan B as views about the truth-conditions of attitude ascriptions embedding normative vocabulary. In this way, we were implicitly generating constraints—racking up debt, you could say—in semantics-pragmatics. As I see it, both Plan A and Plan B need eventually to vindicate their proposed truth-conditions compositionally. (Or show that compositionality doesn't matter—as expressivists who take a Horwichian path might.) This is one of the core aspects of the Frege-Geach challenge. Which view has an easier time here? Is either approach even remotely plausible on linguistic grounds? These are natural next questions.[31,32]

References

Ayer, Alfred J. (1936). *Language, Truth and Logic*, 2nd ed. New York: Dover.

Blackburn, Simon (1993). *Essays in Quasi-Realism*. Oxford: Oxford University Press.

Bratman, Michael (1987). *Intention, Plans, and Practical Reason*. Cambridge: MIT Press.

Carnap, Rudolf (1935). *Philosophy and Logical Syntax*. London: Kegan Paul.

Charlow, Nate (2015). "Prospects for an Expressivist Theory of Meaning." *Philosophers' Imprint* 15, no. 23: 1–43.

Copp, David (2001). "Realist-Expressivism: A Neglected Option for Moral Realism." *Social Philosophy and Policy* 18, no. 2: l–43.

Dandelet, Sophia (2017). "Plans, Preference, and Normative Judgment." Unpublished draft.

Gibbard, Allan (1981). "Two Recent Theories of Conditionals." In *Ifs: Conditionals, Belief, Decision, Chance, and Time*, edited by W. Harper, R. Stalnaker, and G. Pearce, 211–47. Dordrecht: D. Reidel, 1981.

——— (1986). "An Expressivistic Theory of Normative Discourse." *Ethics* 96, no. 3: 472–85.

——— (1990). *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

——— (2003). *Thinking How to Live*. Cambridge, MA: Harvard University Press.

——— (2006). "Reply to Critics." *Philosophy and Phenomenological Research* 72, no. 3: 729–44.

——— (2012). *Meaning and Normativity*. Oxford: Oxford University Press.

31 Yalcin (2018) contains some discussion.

32 Thanks to audiences at the University of Toronto, Rutgers, and the 2017 Philosophy Mountain Workshop for helpful feedback on earlier versions of this work. Thanks also to Sophie Dandelet, William Dunaway, Melissa Fusco, Anandi Hattiangadi, Joshua Petersen, and David Plunkett. I am especially indebted to conversations with Alejandro Perez Carballo for discussion of the modeling questions that arose in section 3.

Gillies, Anthony. (2010). "Iffiness." *Semantics and Pragmatics* 3, no. 4: 1–42.

Hare, Richard M. (1981). *Moral Thinking: Its Method, Levels, and Point*. Oxford: Clarendon Press.

Heim, Irene (1992). "Presupposition Projection and the Semantics of Attitude Verbs." *Journal of Semantics* 9, no. 3: 183–221.

Hintikka, Jaakko (1962). *Knowledge and Belief: An Introduction to the Logic of Two Notions*. Ithaca, NY: Cornell University Press.

Horwich, Paul (1998). *Truth*, 2nd ed. Oxford: Clarendon Press.

Kolodny, Niko, and John MacFarlane (March 2010). "Ifs and Oughts." *Journal of Philosophy* CVII, no. 3: 115–43.

Lassiter, Daniel. (2011). *Measurement and Modality: The Scalar Basis of Modal Semantics*. PhD thesis, PhD dissertation, New York University.

Levinson, D. (2003). "Probabilistic Model-Theoretic Semantics for 'want'." In *Proceedings of SALT 13*, edited by R. Young and Y. Zhou, 222–39.

Lewis, David K. (1979). "Attitudes De Dicto and De Se." *Philosophical Review* 88, no. 4: 513–43.

——— (1988). "Relevant Implication." *Theoria* 54, no. 3: 161–74.

——— (1994). "Reduction of Mind." In *A Companion to the Philosophy of Mind*, edited by S. Guttenplan, 412–31. Oxford: Blackwell.

Moss, Sarah (2018). *Probabilistic Knowledge*. Oxford: Oxford University Press.

Quine, Willard V. (1948). "On What There Is." *Review of Metaphysics* 2, no. 1: 21–38.

Railton, Peter (1992). "Nonfactualism about Normative Discourse." *Philosophy and Phenomenological Research* 52: 961–68.

Ridge, Michael (2014). *Impassioned Belief*. Oxford: Oxford University Press, 2014.

Rosen, Gideon (1998). "Blackburn's *Essays in Quasi-Realism*." *Nous* 32, no. 3: 386–405.

Russell, Bertrand (1935). *Religion and Science*. Oxford: Oxford University Press.

Santorio, Paolo (2016). "Nonfactual Know-How and the Boundaries of Semantics." *Philosophical Review* 125, no. 1: 35–82.

Scanlon, T. M. (May 2006). "Reasons and Decisions." *Philosophy and Phenomenological Research* LXXII, no. 3. 722–28.

Schroeder, Mark (2008a). *Being For: Evaluating the Semantic Program of Expressivism: Evaluating the Semantic Program of Expressivism*. Oxford: Oxford University Press.

——— (2008b). "What Is the Frege-Geach Problem?" *Philosophy Compass* 3, no. 4: 703–20.

——— (2009). "Hybrid Expressivism: Virtues and Vices." *Ethics* 119, no. 2: 257–309.

Stalnaker, Robert. (1976). "Propositions." In *Issues in the Philosophy of Language: Proceedings of the 1972 Oberlin Colloquium in Philosophy*, edited by Alfred F. MacKay and Daniel D. Merrill, 79–91. New Haven: Yale University Press.

——— (1984). *Inquiry*. Cambridge, MA: MIT Press.

——— (2014). *Context*. Oxford: Oxford University Press.

Starr, William (2016). "Dynamic Expressivism about Deontic Modality." In *Deontic Modality*, edited by Nate Charlow and Matthew Chrisman, 355–94. Oxford: Oxford University Press.

Willer, Malte (2017). "Advice for Noncognitivists." *Pacific Philosophical Quarterly* 98, no. S1:174–207.

Yalcin, Seth (2007). "Epistemic Modals." *Mind* 116, no. 464: 983–1026.

——— (2011). "Nonfactualism about Epistemic Modality." In *Epistemic Modality*, edited by Andy Egan and Brian Weatherson, 295–332. Oxford: Oxford University Press, Oxford.

——— (2012). "Bayesian Expressivism." *Proceedings of the Aristotelian Society* CXII, no. 2: 123–60.

——— (2018). "Expressivism by Force." In *New Work on Speech Acts*, edited by Daniel Fogal, Daniel Harris, and Matt Moss, 400–29. Oxford: Oxford University Press.

IV

DISAGREEMENT, OBJECTIVITY, AND REALISM

13

CONVERGENCE IN PLAN

Mark Schroeder

Moral judgments, Gibbard tells us, are plan-laden—fraught with ought, they implicate the directive, planning, side of our psychology, as well as its prosaic, representational, side. Following Stevenson, Gibbard has emphasized that such plan-laden judgments put us in disagreements with one another that are no less profound than disagreements in purely prosaic belief. To this, many would add that these claims also allow for an explanation not only of the *possibility* of moral disagreements but also of their *pervasiveness*. If moral judgments are plan-laden, it can seem to be no wonder that Cleopatra and Antony differ in their moral judgments. Each's judgments reflect plans—*choices*—about how to respond to possible situations, and the difference in their plans reflects only the fact that many choices are possible.

Gibbard's expressivism, however, does not concern moral judgments only. It is a general claim about normative judgments of all kinds. Genuinely normative judgments, Gibbard holds, are all plan-laden in this way—this is the distinctive mark of the normative. And among such normative judgments are judgments of linguistic meaning. If the plan-ladenness of moral judgments is what explains the depth and persistence of moral disagreements, then we might expect to see a similar depth and persistence of disagreement in each normative domain. But this is not obviously so. Indeed, it is obviously not so. Though there are disagreements about meaning, for example—and whether expressivism is true is certainly among them—this is against a backdrop of a great deal of consensus.[1]

1 Nicholas Laskowski (ms) makes a similar observation about the normativity of speech acts in his APA author-meets-critics comments on Cuneo (2014), which he touches on in Laskowski (2017). This paper is an extended attempt to work out the implications of Laskowski's remark.

In this paper I want to explore what can be said, from an expressivist perspective much like Gibbard's, about such consensus in normative outlook—about convergence in plan. What I will be looking for are the sorts of factors that might lead us to expect a greater degree of convergence in plan, and evidence as to whether these factors are present in greater numbers in the case of judgments of linguistic meaning than in the case of moral judgments. I will distinguish between three main kinds of engine of convergence—high-octane, medium-octane, and low-octane. High-octane engines of convergence guarantee perfect convergence over some domain, at least among rational and reflective thinkers. When these are present, there is a guarantee given meaning that every thinker is rationally committed to taking the same view, or at least to not denying it. Low-octane engines of convergence are driven by empirical or historical assumptions—convergence over some domain could be driven simply by psycho-social or etiological factors. And medium-octane engines of convergence aspire to something middling. They appeal to intrinsic features of some planning questions in order to explain why some answers are more natural than others.

1. High-Octane Convergence

Some normative convergence is pervasive and rationally compelling. According to Gibbard, for example, every rational thinker is committed to accepting the claim that the normative supervenes on the natural and, similarly, to accepting the claim that there is some natural property that constitutes what it is to be wrong. The arguments that Gibbard gives for these claims are, as I will put it, high-octane. They put convergence over these claims in the same category as convergence over truths of logic. Given the meanings of all of the terms involved in these claims, every thinker is rationally committed to the accepting them, unless they fail to have the relevant concepts at all.

High-octane explanations of convergence are a particularly essential tool in any normative expressivist's toolkit. Broadly speaking, what high-octane explanations of convergence do is to establish a kind of *analyticity* for certain claims. But we need to be careful about exactly what this means. Normally, a claim is understood to be analytic just in case the meanings of its terms guarantee it to be true. But Tappenden (1993) has taught us how to relax this conception: we might take analyticity to be the status that a sentence has when the meanings of its terms guarantee it not to be false.

But expressivist meanings are not truth-conditional in nature; they do not guarantee any claims to be true or even not to be false—at least not directly. So what expressivist meanings are capable of guaranteeing directly is that some claims are ones that everyone (or perhaps, everyone who understands their meanings) is rationally committed to *accepting*, or—on the analogue of Tappenden's generalization—at least rationally committed to not denying.[2] These expressivist analogues of analyticity play the right role for analyticity, within

2 Compare Schroeder (2010b), which I apply this expressivist conception of analyticity to the paradox of the liar.

an expressivist view—they establish analytic claims as ones that can be taken for granted (on the first formulation), or at least as undeniable (on the more relaxed formulation).

Expressivist-friendly analyticity, as I have said, is an essential tool for expressivists. Not only can it be used to characterize the status of supervenience and the natural constitution of normative properties, as Gibbard explores in *Thinking How to Live*, but it can be used to characterize the validity of arguments. The claim that it is impossible for any argument of the form *modus ponens* to have true premises but an untrue conclusion can be shown to satisfy the expressivist analogue of analyticity.[3] With care, we could extend Gibbard's system to prove that every thinker is committed to accepting this claim, and similarly, in the system of biforcated attitude semantics, developed in my *Being For*, there is a relatively straightforward proof that it is rationally undeniable.[4] But once we have this result on board, since we are ourselves rational thinkers, we may take the transcendental turn. Since every rational thinker is committed to accepting this claim (or at least, to not denying it), we should accept it, too, on pain of irrationality. And since we do, let us assert it: it is impossible for any argument of the form *modus ponens* to have true premises but an untrue conclusion. That is how we earn the right not just to some expressivist substitute for validity, as noncognitivists have sought at least since the criticisms of Ross (1938), but also to the real thing.

But even before we take the transcendental turn and assert, as theorists, the claims about validity, supervenience, and property constitution, our standing to make these claims, for the expressivist, is rooted in the high-octane explanation of why we should expect convergence. We should expect convergence over these matters for the same reasons that we should expect convergence over matters of logic or other analyticities—because anyone who disagrees is either making some mistake about meanings or making some mistake about following through on their own rational commitments. So, it is no wonder that we observe a striking asymmetry between the great divergence among those who accept utilitarianism and those who do not, and the great convergence over normative supervenience. The latter is not incompatible with the idea that expressivism helps to explain the existence of deep disagreements—rather, it is just what you would expect, given that some claims are bound to be analytic, and supervenience is (according to many, at least[5]) plausibly one of them.

2. Low-Octane Convergence

High-octane explanations of convergence explain part of the contrast between topics of great normative disagreement and topics of great consensus. But they only explain part of it.

3 Compare Schroeder (2010a), chapter 10.

4 See Schroeder (forthcoming).

5 It is worth pointing out that I myself do not think that supervenience is analytic, since that is not always obvious to readers of my (2005), (2007), (2014b), or Schmitt and Schroeder (2012).

There are many matters of significant consensus that do not lend themselves at all to such high-octane explanations.

For example, for most of human history, it was a matter of great consensus that duties to people far away are less stringent than duties to the nearby. This is hardly the sort of thing that we would expect to be analytic—indeed, it is almost certainly false. Similarly, there is great consensus that it is wrong to torture someone for fun. But though in contrast to the former claim, this is almost certainly true, it seems like an important *substantive* truth. People who deny this claim aren't linguistically mistaken—they are *evil*. So, the claim that it is wrong to torture people for fun isn't the right kind of thing to be an analytic truth—and so it's not the right kind of thing to be subject to a high-octane explanation of convergence.

Fortunately, low-octane explanations of convergence can come to the rescue. Though there is no linguistic confusion or failure to follow through on their own rational commitments that is exhibited by people who deny that it is wrong to torture for fun or that duties to people who are far away are less stringent than duties to people who are nearby, it is no surprise that such people are rare. For moral views are constituted by patterns of norm-acceptance (or by plans for what to do in a range of counterfactual circumstances), and because which norms we accept (or which plans we adopt) have strong implications for how we act, moral views are subject to powerful evolutionary forces. Altruistic behavior toward neighbors received ample evolutionary payoffs under the historical conditions of human evolution in small hunter-gatherer societies, but not so altruistic behavior toward the distant needy, and so it is no wonder that stronger moral attitudes toward helping the nearby were selected for, without selecting such strong attitudes toward helping the distant needy. Similarly, torture for fun is a paradigmatically noncooperative activity with no direct evolutionary payoff. So, the evolutionary payoffs of cooperation would naturally select against it. Obviously, these explanations could be tightened considerably, and I am merely gesturing toward how such explanations might go.

Low-octane explanations of moral convergence may also be cultural. Widespread consensus could be the result of influential films or novels, of popular trends, or even of the bare fact of unfamiliarity with certain ways of life. Low-octane explanations of moral convergence may be historical, etiological, psychological, or sociological. What they have in common is that they offer contingent explanations that aspire to explain something less than high-octane explanations of moral convergence.

Appreciating the diversity of the range of possible low-octane explanations of moral convergence is important, in order to temper our sense of what it is reasonable to expect about patterns of disagreement in some domain, given an expressivist account of that domain. Low-octane explanations show why there may be a wide variety of explanations for why not everything seems to be up for grabs, just because the meanings of the terms in some domain do not settle what views it is rational to hold in that domain.

But as our examples of low-octane explanations of convergence clearly illustrate, low-octane explanations have severely limited power. It is at least as plausible that evolutionary

considerations explain why throughout human history, the view that our duties to those nearby are more stringent than our duties to the distant needy has been nearly universal as it is that they explain why human history has been dominated by the view that it is wrong to torture for fun. But to many of those of us who reflect on the matter in the twenty-first century, when there are easy ways of helping people in all corners of the globe, it is far from obvious that distance could possibly affect the stringency of our moral obligations.[6] The fact that we now understand why many people have thought otherwise does nothing at all to cast the common view in a favorable light, let alone to make it seem compelling or natural, in any way. It explains without rationalizing.

I take away from this example the observation that low-octane explanations of convergence in normative outlook are weaker than high-octane explanations along more than one dimension. They are less powerful in that they typically explain less convergence—whereas high-octane explanations of convergence can explain convergence among every rational and reflective thinker, low-octane explanations of convergence will typically only have the right structure to explain tendencies or predominant patterns. But more strikingly, low-octane explanations of convergence provide less than high-octane explanations, because their explanations give us causes without rationalizations. Understanding a low-octane explanation for convergence will never make the converged-on view seem compelling or even appealing, in its own right.

3. Expressivism, Meaning, and Judgment Internalism

Let us return, then, to our test case that motivated this inquiry—the normativity of meaning. To say that meaning is normative is to say that meanings are "fraught with ought," and given the expressivist treatment of the normative, that means that they are in Gibbard's terms plan-laden. To hold that a word has a certain linguistic meaning is to plan for how to use it, or for what standards to hold others to, for its use.

Gibbard's (2012) primary motivation for accepting the normativity of meaning thesis is as an answer to the problem of the underdetermination of meaning, as explored particularly extensively by Kripke (1982). The core of this problem is that meanings are infinite in the distinctions that they make, but the meaning-constituting facts are finite. For every pattern that extends the totality of past finite patterns of use out into the future, there are infinitely many alternatives to that pattern that respect the totality of past use equally well. This is no problem at all, Gibbard contends, if meaning is something that we bring to the world, rather than something that we find there.

Once we adopt expressivism about meaning, moreover, our account of meaning turns in on itself. Since expressivism is itself a claim about meaning, that means that expressivism itself is a thesis fraught with ought. Endorsing expressivism, either about meaning or about

6 Compare especially Singer (1972).

morality, therefore, consists in accepting a plan for what to do with the words 'means' or 'ought', or what standards to hold others to, in their use of these terms.

The fact that expressivism turns in on itself in this way could potentially have striking implications for Gibbard's original metaethical expressivism, in turn. Gibbard's original formulation of metaethical expressivism, in *Wise Choices, Apt Feelings*, was committed to a particularly strong form of judgment internalism—on that view, it is literally impossible to think that stealing is wrong without being in a norm-acceptance state that would motivate you not to steal. And sincere speakers will always be in the mental states that their assertions express, so sincere speakers who assert, "stealing is wrong" will always be in such a norm-acceptance state. But given that expressivism itself is a plan-laden thesis, the disagreement between cognitivists and expressivists is itself a plan-laden one. So, speakers who plan to use moral language in the cognitivist way may be *sincere* in their assertion of "stealing is wrong," in that they are in the mental state that they themselves *take* it to express, without being in any state of mind with any intrinsically motivating properties. Such speakers would behave much as supposed counterexamples to judgment internalism are alleged to behave.

Of course, those of us who endorse metaethical expressivism will still say that these speakers are no counterexample to the thesis that everyone who *understands* the meaning of "stealing is wrong" and who asserts it sincerely will be in a planning or norm-acceptance state that would motivate them not to steal. And so we will still get the letter of Gibbard's original, very strong, version of judgment internalism. But the spirit of this claim now has the potential to be substantially watered down by the fact that claims about what it takes to understand what a sentence means, at least in general, are now substantive, planning, questions.

So, while we still get to say that the cognitivist who is unmotivated by her moral judgments is no counterexample to judgment internalism, because she does not understand the meanings of moral words, we *may* now also allow that it betrays no misunderstanding of how to use the term "means," to come to her conclusion about the meanings of moral terms rather than ours. This disagreement is, in some sense, to be expected, precisely *because* it is a kind of disagreement in plan.[7] So our criticism of the character who sincerely asserts, "stealing is wrong" but has no motivation not to steal is now on a par with our criticism of the character who believes that our obligations to the nearby are more stringent than our obligations to the distant needy—it is grounded in a substantive planning error.

My own view is that this is one of the subtlest and most wonderful possible upshots of Gibbard's arguments in *Meaning and Normativity*—a striking virtue that has the potential to arise when Gibbard's views about meaning are put together with his views about moral language, so long as we allow (which Gibbard may not) that the dispute between cognitivists

7 The disagreement is particularly to be expected, if the dispute between expressivism and cognitivism is itself one of the cases in which patterns of use underdetermine meaning—a claim that I think Gibbard himself rejects but which I think could be accepted along with most of his other commitments. Thanks to Billy Dunaway for discussion.

and expressivists is itself one of the questions left open for planning. Although metaethical expressivism like the variety that Gibbard advanced in his earlier work, particularly in *Wise Choices, Apt Feelings*, offers a powerful account of the practical force of moral language, one glaring worry that we should always have had about it is that its account is *too* powerful, because it is committed to a stronger claim about what this practical force amounts to than can plausibly be defended, given what we know about the vast range of actual motivational upshots of moral judgments among the diverse range of thinkers and speakers in the real world.

The idea that metaethical expressivism is itself a plan tempers this thesis, without strictly weakening it. It still comes out as true (according to the proponent of both expressivism about meaning and metaethical expressivism) that anyone who understands the meaning of "stealing is wrong" and sincerely asserts it will be in a planning state that has the right structure to motivate them not to steal, but there will be speakers who exhibit no misunderstanding whatsoever of the meaning-determining facts who can sincerely assert "stealing is wrong" *with its usual meaning* but have no motivation whatsoever not to steal. This *is* a concession to motivational externalists, but in contrast to other concessions—such as Michael Smith's (1994) suggestion that the proper formulation of judgment internalism includes a restriction to agents who are "practically rational"—it does not water down judgment internalism too much for it to still bear weight in supporting metaethical expressivism. So, it is the right kind of concession to make for an expressivist who seeks to use judgment internalism in its traditional role of helping to motivate metaethical expressivism. It strengthens the hand of the metaethical expressivist, by heading off the worry that he is committed to an unacceptably strong form of judgment internalism.

4. Convergence in Plans for Meaning

As I said, in my view the implication that I have just been discussing of the normativity of meaning thesis for the defensibility of metaethical expressivism is one of the most striking and powerful upshots of Gibbard's normativity of meaning thesis. It is just one example of how powerfully the more general perspective of the normativity of meaning can lead us to rethink how we understand the commitments of metaethical expressivism, without undermining those commitments.

But now I want to make one important observation about how we get this upshot: we get it because we were assuming that because the question of whether cognitivism or expressivism is true for moral terms is a planning question, both answers are in some sense optional, and hence to be expected. This is just on a par with the assumption that since the question of whether utilitarianism or deontology is true is just a planning question, both answers are in some sense optional, and hence to be expected. In both cases, we get the expectation of disagreements from the diagnosis of the underlying nature of the issue as a planning question.

I've been arguing that it can be a *virtue* of Gibbard's package of commitments that it opens up the possibility of allowing that the disagreement between cognitivists and expressivists is reasonable. But even if we don't say this—even if we conclude that cognitivists somehow get things more deeply wrong than either deontologists or consequentialists, and so this disagreement is not, fundamentally, reasonable, we must still recognize that it *exists*. It is an example of a deep and substantial disagreement about meanings. Like all disagreements about meanings, there is some trivial sense in which whichever party is incorrect does not even understand the meanings of the words at stake in the debate, but this sense is trivial and does not obscure the fact that this disagreement really exists.

But that brings the puzzle that I am pursuing in this paper into full view—*some* disagreements about meaning do exist and are quite arguably intelligible and natural, including the disagreement between expressivism and cognitivism about moral claims. But *many* such disagreements are not at all intelligible or natural. The view that "steal" means to give a gift is just a plan about how to use the word "steal," but we don't observe widespread disagreement about whether "steal" means to give a gift; on the contrary, this view is not merely false but also absurd. So, the challenge for the advocate of expressivism about meaning is to explain how the planning nature of meaning judgments could create space for disagreement in a very restricted range of cases, at the same time as we observe that the vast majority of possible plans about what words mean seem to be not just false but totally absurd and endorsed by no one, actual or imaginable.

So far, we have seen two models for explanations of why we might observe a convergence in plan, despite the fact that something is itself a planning question. But neither of these models does quite what we should want of an account of the vast range of agreement that we see over linguistic meaning. On the high-octane model, we get explanations of convergence because some planning questions turn out to have analytic answers—answers to which every planning agent is implicitly committed, or at least committed not to denying. But the claim that "steal" does not mean to give a gift does not seem like the right kind of thing to be analytic—if this is not obvious, it helps to remind yourself that this sentence only mentions, but does not use, the word "steal."

In contrast, on the low-octane model, all that we get are causal explanations of patterns among speakers' plans. And as we've seen, these kinds of causal explanations don't *rationalize* the views that they explain. But the claim that "steal" does not mean to give a gift is not just common, or even just universal—it is *compelling* to anyone acquainted with the use facts surrounding "steal." And it is compelling in a way that ought to be graspable by anyone who understands how meaning works.

We need, therefore, a third model for explanations of normative convergence—one that can hope to explain the strong rational pressures toward convergence about meaning and other topics of intense and pervasive normative agreement, without subsuming this agreement to the analytic. We need to see what a medium-octane explanation of convergence might look like.

5. Medium-Octane Convergence

What we need from a medium-octane explanation of convergence is that it rationalize, and not merely explain, convergence in plan. The more forcefully it rationalizes this convergence, the stronger its claim to be able to make good on the prediction that every rational and reflective thinker will have plans that satisfy the relevant constraint, and hence to predict extensive actual convergence, given minimal assumptions about real thinkers' actual levels of rationality. The answer, I take it, is that some plans are better than others. Not better merely in the sense expressed by a second-order plan to make some plans rather than others, but better in some unavoidably recognizable way—that can be appreciated by anyone, no matter what else they plan.

This is a high standard. In order to be bad in a way that can be appreciated by anyone, no matter what else they plan, a plan would essentially have to be self-frustrating, or at least conditionally self-frustrating, given pretty minimal conditions. That is what I will now argue medium-octane explanations of convergence in plan can provide.

Consider the case of tic tac toe. Some plans about how to play tic tac toe are better than others. There is no mystery about why there is so much convergence among plans about how to play tic tac toe, for the point of these plans is to solve some problem—how to *win* or at least *avoid losing* at tic tac toe. And given the rules of the game, some plans are straightforwardly dominated by others. Finally, the way in which some of these plans dominate others with respect to this goal is easily discovered and widely known. So, it is no wonder that these are the strategies on which everyone converges, and it is no wonder that these strategies seem rationally compelling to anyone who shares the goal of winning or at least not losing at tic tac toe.

Not every move at tic tac toe is mandated by the winning strategy. The first move of the second player, for example, is fairly unconstrained by the goal of not losing. Most of us, therefore, adopt a plan that is indifferent about what move to make at this stage of the game but restrictive about what move to make at later moves of the game. But another conceivable plan is to always mark the top left corner at this stage of the game. This plan is no more frustrated by the goal of not losing than the more permissive plan. And similarly for the plan to always mark the bottom right box at this stage of the game. But though there are *several* possible plans for how to play tic tac toe that are not dominated, conditional on the goal of not losing, the vast majority of such plans are ruled out. They are guaranteed to do worse than some other strategy at the goal of not losing.

So, if hypotheses about linguistic meaning are like strategies for how to communicate, as strategies in tic tac toe are strategies for how to win or at least not lose, then the background facts about patterns of use that according to Gibbard, following Kripke's Wittgenstein, underdetermine linguistic meaning could set sharp constraints on the success of these strategies. If your goal is to successfully communicate with your audience, then it takes only minimal observation of the pattern of use of "steal" in order to observe that successful communication will be difficult, if you use it to mean "to give a gift."

Using "steal" to mean "to give a gift" is not intrinsically doomed to failure—it could succeed in an environment very different from ours—and even in our environment, it works just fine if successful communication is not one of your goals. But given minimal background facts about other speakers of English, it straightforwardly fails at the goal of successful communication, and it does so in a way that is relatively transparent—obvious enough to even minimally reflective thinkers that it should be no wonder that any thinker who formulates views about linguistic meaning as part of a plan to communicate will not plan in this way.

6. Optional Meanings

Meaning, Kripke's Wittgenstein reminds us, is infinite, but use facts are finite. The use facts underdetermine how to go on well after our use facts have died out, and even an intention to go on in one way rather than another helps only if something makes it the case that one's intention has one content rather than another. Hypotheses about linguistic meaning that diverge over the infinity of cases underdetermined by the use facts, therefore, are predictably going to be unconstrained by the use facts with respect to the goal of successful communication. No hypothesis about linguistic meaning that diverges only over this range of cases can get one into trouble, given the goal of successful communication. And so these alternative hypotheses might each be a reasonable plan about how to use language.

This result makes good on the idea that the use facts underdetermine the meaning. After all, if there was only a *unique* plan that fit with the use facts, facts about linguistic meaning could simply be facts about that plan, and so meaning would not be underdetermined by use, after all. But this kind of result can also be extended, in order to make sense of evolving uses of language over time. For example, it is now clear that "water" is a natural kind term, which picks out the same chemical kind in every possible world. But it is not obvious that this was always the case. The discovery of the chemical constitution of water could have been a semantic choice point—before which none of the use facts determined how to go on in using the word "water" under the multifarious possibilities of philosophers' imaginations. If that was so, then before the advent of modern chemistry, multiple plans for how to use "water" could have been consistent, given ordinary speakers' knowledge of the use facts, with the goal of successful communication. But even if this is right about the past, we have now crossed a semantic choice point, and our own plans for how to use "water" must correspond to a broader set of constraints. This is just what we should expect, if use constrains meaning but underdetermines it—more use could more closely constrain meaning over time, yielding less and less scope for reasonable disagreements about meaning.[8]

8 By this, of course, I don't mean that use or dispositions to use are a hard constraint—merely that pressures toward success in plans about what to mean come from dispositions and patterns of use, and so if there are more developed dispositions and patterns of use, they will create more pressures on our plans for what to mean.

This view can make sense of why the vast majority of views about linguistic meaning, like the vast majority of strategies for playing tic tac toe, are not just rare but also unreasonable, because they frustrate the goal of linguistic communication, which is the point of meaning things by our words. And it can do so while leaving open a range of permissible plans about what to mean—by and large, plans that go beyond the range of the existing use facts. So what, then, of the dispute between cognitivist and expressivist theories of the meaning of normative terms?

Earlier I endorsed the idea that both cognitivism and expressivism may be reasonable plans about the linguistic meaning of normative terms as a way for Gibbard to soften the commitments of the form of judgment internalism that underlies his norm-expressivism. If, even while endorsing the expressivist plan for what to mean, we allow that the cognitivist plan is another reasonable plan, then we can allow that although every agent who understands the meaning of "ought" is motivated in accordance with their "ought" judgments, among these are agents who fail to understand what "ought" means only because they endorse a different plan for what to mean with it. This makes much more intelligible how these agents could fail to be motivated by their moral judgments, making the resulting form of judgment internalism more palatable. And this explanation requires the assumption that the disagreement between cognitivists and expressivists is itself reasonable.

I'm not yet sure, however, on the form of the medium-octane convergence strategy being considered here, whether the disagreement between cognitivists and expressivists is the right sort of thing to be stably reasonable. It contrasts with the case of the plus/quus distinction and the case of possible past semantic indecision about whether "water" was a natural-kind term in that it applies to cases that have already been considered—indeed, with which we are all, as speakers, highly familiar. So for the disagreement between cognitivists and expressivists to be stably reasonable, both patterns of use need, intuitively, to be persistent and common, even in the face of the fact that we recognize these features of each other's use. Much more would need to be said, in order to achieve clarity about why this continues to be a stably reasonable thing to disagree in plan about, even though it is not reasonable to disagree with someone who uses "Mary" to refer to a different person than you do about to whom it refers.

So I'm not sure that my solution, in the form of a medium-octane strategy for predicting convergence, gets us everything that I would have liked to have, on behalf of the view in the neighborhood of Gibbard's that I would most have liked to have been able to defend. But perhaps I have tried to get too much, and the moral is that we should let go of the idea about using expressivism about meaning in order to soften the force of the kind of judgment internalism to which Gibbard is committed. But the point remains, I think, that the space of strategies is rich, for expressivists to make sense of a wide variety of differences in the space of reasonable disagreement.[9]

9 Special thanks to Nicholas Laskowski for the idea to pursue this line of thought, to Billy Dunaway and David Plunkett for the opportunity to pursue it and for enormously helpful comments, and especially to Allan Gibbard for giving us all so much to think about.

References

Cuneo, Terence (2014). *Speech and Morality: On the Metaethical Implications of Speaking*. Oxford: Oxford University Press.

Gibbard, Allan (1990). *Wise Choices, Apt Feelings*. Cambridge: Harvard University Press.

——— (2003). *Thinking How to Live*. Cambridge: Harvard University Press.

——— (2008). *Reconciling Our Aims: In Search of Bases for Ethics*. Oxford: Oxford University Press.

——— (2012). *Meaning and Normativity*. Oxford: Oxford University Press.

Kripke, Saul (1982). *Wittgenstein on Rules and Private Language*. Cambridge: Harvard University Press.

Laskowski, Nicholas (ms). "Comments on Terence Cuneo." Comments delivered at the Pacific Division meeting of the APA, April 2016.

——— (2017). "Speech and Morality: On the Metaethical Implications of Speaking, by Terence Cuneo." *Journal of Moral Philosophy* 14, no. 6: 781–84.

Ross, W. D. (1938). *Foundations of Ethics*. Oxford: Oxford University Press.

Schmitt, Johannes, and Mark Schroeder (2012). "Supervenience Arguments under Relaxed Assumptions." *Philosophical Studies* 155, no. 1: 133–60. Reprinted in Schroeder (2014a).

Schroeder, Mark (2005). "Realism and Reduction: The Quest for Robustness." *Philosophers' Imprint* 5, no. 1. 1–18.

——— (2007). *Slaves of the Passions*. Oxford: Oxford University Press.

——— (2008). *Being For: Evaluating the Semantic Program of Expressivism*. Oxford: Oxford University Press.

——— (2010a). *Noncognitivism in Ethics*. New York: Routledge.

——— (2010b). "How to Be an Expressivist about Truth." In *New Waves in Truth*, edited by Nikolaj Jang Pederson and Cory Wright, 282–98. New York: Palgrave MacMillan. Reprinted in Schroeder (2005).

——— (2014a). *Explaining the Reasons We Share: Volume 1 of Explanation and Expression in Ethics*. Oxford: Oxford University Press.

——— (2014b). "The Price of Supervenience." In Schroeder (2014a).

——— (2015). *Expressing Our Attitudes: Explanation and Expression in Ethics, Volume 2*. Oxford: Oxford University Press.

——— (2018). "The Moral Truth." In *Oxford Handbook to Truth*, edited by Michael Glanzburg. Oxford: Oxford University Press. 579–601.

Singer, Peter (1972). "Famine, Affluence, and Morality." *Philosophy & Public Affairs* 1, no. 3: 229–43.

Smith, Michael (1994). *The Moral Problem*. Oxford: Basil Blackwell.

Stevenson, C. L. (1937). "The Emotive Meaning of Ethical Terms." *Mind* 46, no. 181: 14–31.

Tappenden, James (1993). "The Liar and Sorites Paradoxes: Toward a Unified Treatment." *Journal of Philosophy* 90, no. 11: 551–77.

14

COMIC DISAGREEMENT

Lauren Olin

1. Introduction

On September 20, 1895, the *Sheffield Independent* reported on an "Adventure of Three Distinguished Cyclists." Here are the highlights:

> Mr. George Bernard Shaw tells in a letter a "piece of catastrophic news" about three distinguished persons—to wit, himself, Mr. Sidney Webb, and the Hon. Bertrand Russell, brother of Earl Russell. "On Thursday afternoon, on the road from Trelleck to Chepstow we three rode on our bicycles down a steep hill on our way to Tintern Abbey. Russell is rather absent minded, as he is preoccupied at present with a work on non-Euclidean space. He suddenly woke up from a fit of mathematical abstraction, and jumped off his machine to read a signpost. The consequences may be imagined. G.B.S. was just behind him and there was a 'terrific smash' and the great critic and Fabian was hurled 5 yards through space (Euclidean) and landed impartially on several parts of himself . . . Mr. Shaw was able to ride home after, as he explains, lying flat on his back on the roadway for awhile, and defending himself against all proposals to poison him with brandy."

More than sixty years post-incident, Russell recalls having been concerned for Shaw, fearing the accident "might have brought his career to a premature close" (1956: 81). Russell was not hurt, but his bicycle was badly mangled, and he had to return to the Monmouthshire home of Sidney and Beatrice Webb by train. Russell recalls that the train was very slow, slow enough that Shaw could cycle at a faster pace, and that "at every

station Shaw with his bicycle appeared on the platform, put his head into the carriage and jeered" (ibid.).[1]

Shaw and Russell appeared to agree about the moral significance of the crash: Shaw was hurt, and Russell was concerned for him. In one way this is surprising: for if there are any predictably comical ways of falling, bicycle crashes are one. The bicycle crash scene has been celebrated since (at least) *Wheels of Chance*, the 1922 film adaptation of H. G. Wells' 1895 story by the same name.[2] However, Russell and Shaw seem to disagree about the significance of Shaw's antics following the crash: Russell apparently did not think that Shaw's jeers from the train platform were funny, and apparently did not think so in a way that proved consequential. Their friendship, reportedly, never recovered from the events of the weekend (Brown 2013; Russell 1956). In his autobiography, Russell explains that his admiration for Shaw "had limits" (1956: 83):

> It used to be the custom among clever people to say that Shaw was not unusually vain, but unusually candid. I came to think later on that this was a mistake . . . Shaw, like many witty men, considered wit an adequate substitute for wisdom.

Although Russell didn't consistently pursue a metaethical agenda, he was among the first to articulate an emotivist view, and some of his commitments bear striking resemblance to Allan Gibbard's.[3] Consider the following points from an early lecture on Moore's idea of an absolute good, probably delivered first in 1922 (Russell 1988, 149)[4]:

---

1 I've found no record of Webb's impression of the episode. Russell elsewhere describes him as "somewhat earnest" and disposed to dislike jokes about "sacred subjects such as political theory" (Russell 1956, 73).

2 Humor almost always—or always, according to some (Bergson 1900; Hurley, Dennett, and Adams Jr. 2010)—involves violations of the intentional expectations we have for agents acting in goal-directed ways. Moreover, very few persons, in very few circumstances, plausibly *intend* to crash their bicycles. Bicycle crashes are in fact highly dangerous, and it's not clear in many cases why onlookers are disposed to laugh rather than to rush in and help.

3 While expressives have been the subject of only marginal interest in linguistic discussions, due to Gibbard's influence expressivist views have recently been offered in diverse areas of metanormative discourse. For aesthetic versions of expressivism, see Prinz (2004) and Hopkins (2009). Barker (2006), Price (2011), and Yalcin (2012) develop versions of the view for probability claims, while Schroeder (2010a) and Yalcin (2012) develop a version of the view for epistemic modals; Blackburn (1993, 52–74) and Price (2011) develop versions of the view for other modals. Gert (2002) and Dreier (2009) remark on the possibility of developing an expressivist account of humor. Finally, Schroeder develops a version of the view for conditionals (2008, 2010a) and for logical vocabulary (cf. Brandom 1994, 2000).

4 Pidgen (2014) speculates that the neglect of Russell's emotivist work results from the fact that his most original contributions were not published during his lifetime, and to the fact that what did come out in print was often put out as part of works devoted to less theoretical topics. But he was at the forefront of the noncognitivist turn: "Is There an Absolute Good" was delivered twenty years prior to the publication of Mackie's "The Refutation of Morals," but then only published posthumously in 1988. His version of emotivism was first published at the end of a popular book on science and religion published in 1935, one year before Ayer's *Language, Truth and Logic* appeared and two years prior to the first publication of Stevenson's *The Emotive Meaning of Ethical Terms*.

> It seems to be an empirical fact that the things people judge good are the same as those towards which they have an emotion of approval, while the things they judge bad are those towards which they have an emotion of disapproval.
>
> The emotions of approval and disapproval influence our actions, whereas purely theoretical judgments do not. Therefore, in so far as ethics is concerned with what people actually do, or with how to influence action, the emotions suffice without the help of the predicates "good" and "bad."

Gibbard and Russell are united in endorsing the thesis that moral claims express attitudes. They are both committed to the idea that attitudinal approvals and disapprovals are intimately connected to motivational capacities, and to decisions about what ought to happen or be done. They are both possessed of naturalistic ambitions. And they are both committed to the idea that it is possible to formally analyze moral meaning.

Despite these points of similarity, Gibbard and Russell apparently disagree about the possibility of genuine disagreement. Gibbard has long emphasized the reality and importance of normative disagreement. His long-standing ambition has been to construct an expressivist semantics for normative terms that can (1) vindicate the validity of reasoning involving those terms and (2) vindicate realism about normative discourse. Fundamentally, for Gibbard, questions about how to live, about what to do together, are the questions about which it is possible to agree or disagree genuinely.

Most recently, Gibbard understands disagreement in terms of states of *planning*: noncognitive mental states that reflect what an individual judges is the thing to *do*; to plan to *x* is, likewise, to endorse *x*-ing. To disagree, on this view, is just to reject some plan for action in ways that reflect more than "mere personal difference" in opinions about what ought to be done (2003, 280–81). Plans and associated intentions to act are bound by the norms of consistency implicated by pro-attitudes. Norms of consistency are violated, according to Gibbard, by claims such as that "I think this is the best thing to do, but I don't favor doing it" (2003, 29). The difference between the status of an apparent disagreement, or mere impasse, and a genuine one is cashed out in the terms provided by *epistemic symmetry*: when speakers agree on the meanings of moral terms and rationalize their rejection of plans nonindexically, their disagreement is genuine. Attention to genuine disagreements is supposed to provide substantial insight into the structure of the normative systems people endorse.

For Russell, in contrast, *any* dispute about the nature of value must not be treated as "a disagreement as to any kind of truth, but a difference of taste" (Russell 1935, 249). For Russell, differences in taste can *only* be grounded in subjective desires: "when we assert that this or that has 'value,' we are giving expression to our own emotions, not to a fact which would still be true if our personal feelings were different" (1935, 242). Moral judgments, on Russell's view, involve no claims but rather express desires about how the world ought to be, or about how oneself or others ought to feel in it. Saying "hatred is bad," for Russell, is really to say "would that no one felt hatred" (Russell 1938, 247).

The distinction between genuine disagreements and subjective differences of opinion has recently received attention in the guise of a debate about the possibility of faultless disagreement. Following Kant's (1790, 52) distinction between judgments of beauty that make "rightful claim upon the assent of all men" and judgments of taste as judgments of individual preference, a majority now agrees that a viable semantics for predicates of personal taste (PPTs) must account for the possibility of disagreement about whether something is *fun* or *tasty*, for instance, without either party being in error. And while intuitions about faultless disagreement in the case of PPTs are widely shared, the semantic significance attributed to the phenomenon is increasing in other areas of evaluative discourse, including the comic domain. As Egan (2014) has argued, on the basis of examples such as the difference in preferences for the UK and US versions of *The Office*, sometimes when people's comic judgments conflict, "we are not inclined to say that anybody is in error" (Egan 2014, 74).

Comic disputes do sometimes appear faultless: they sometimes seem to reflect expressions of subjective attitudes like those prominently associated with judgments of taste and probability. But it also seems right that some disputes about the funny are genuine, along the lines of more robustly moral expressions of approval and disapproval (de Sousa 1990; Gaut 1998; Smuts 2010). People do claim that some things are genuinely funny, that other things are not or are less so, and sometimes hold others accountable for botched attempts at amusement (as Russell, apparently, held Shaw so accountable). People worry about whether they are mistaken in their judgments of funniness. It is commonly assumed that it is possible to improve one's comic sensibility through exposure to the right people, or to better jokes.[5] In the course of day-to-day life people as well as institutions are censured for their attempts at humor, or for their expressions of amusement (Alexander 1986; Bergson 1900; Cohen 1999). And when offense is given—or taken—people as well as institutions can and do apologize. There are norms regarding the appropriateness of expressions of mirth, just as there are norms concerning the appropriateness of emotional responses and attitudinal expressions in other evaluative domains (D'Arms and Jacobson 2000; Shoemaker 2015; Sripada and Stich 2006).

A viable account of the semantics of comic terms must account for both of these facts: it must explain intuitions about cases of apparently faultless comic disagreement but also account for genuine comic disagreement. This paper explores the idea that faultless-looking comic disputes are well explicated in terms of the expression of desires, on the model of Russell's emotivism. I'll argue that when people appear to disagree about the funny in ways that seem faultless, they are expressing desires concerning the overall structures of the normative systems parties to the dispute endorse. Genuine comic disagreements, in contrast

5 People also appeal to comic experts: just as people recognize moral authorities in their communities, Hollywood celebrates comedic mentors like Lorne Michaels, Gary Shandling, and Judd Apatow. Comedians have even developed their own critical vocabulary: comics deride other comics as *clowns* for making jokes that are objectionably cheap or crude.

and following Gibbard, involve conflicting opinions about whether a plan is called for, given antecedent agreement on the meanings of comic terms.

I'll first introduce existing approaches to the phenomenon of faultless disagreement, and argue that at least some comic judgments, like moral and epistemic judgments, are normative. I'll then articulate an account of comic normativity and explain how it may be usefully deployed to complement Gibbard's account of disagreement in other normative domains. Finally, I'll sketch an associated semantic framework and suggest several ways that the move toward thinking about apparently faultless comic disputes in terms of desires may help Gibbard in his efforts to reduce properly normative beliefs to states of planning.

2. Disagreement and Faultlessness

In semantic theorizing, versions of relativism have been articulated explicitly in order to explain the phenomenon of faultless disagreement in the case of PPTs (Kölbel 2002; Lasersohn 2009).[6] An intuitive relativistic account is subjectivist. According to subjectivism, taste predicates convey the point of view of some particular speaker, and their truth-conditions specify that point of view. However, subjectivist accounts encounter difficulty with cases of disagreement, since disagreements are typically understood in terms of utterances with contradictory contents. Consider:

(1a) Russell: This brandy is tasty.
(1b) G.B.S.: No, the brandy is not tasty!

This intuitive view cannot count (1a) and (1b) as contradictory, since there is no discrepancy between the set of worlds in which Russell finds that the brandy is tasty and G.B.S. fails to, though the speakers in (1a) and (1b) are intuitively disagreeing. Furthermore, disagreement typically prescribes the assignment of fault, and neither Russell nor G.B.S. seems rightly faulted for their judgments about the brandy.

In order to account for the intuition that it's possible to disagree faultlessly about matters of taste, Lasersohn (2005) follows Kaplan (1989) in holding that the content of an utterance is a function from the circumstances of evaluation—world [w] and time [t] indices—to extensions or truth-values. Lasersohn adds a third judge index [j] to the circumstances of evaluation representing the individual whose judgments fix the truth values of utterances implicating PPTs. So, for example, a claim such as that the brandy is tasty takes a positive truth-value just in case the brandy is tasty at w, t, according to j. Given this analysis, we can say that the claims in (1a) and (1b) are both true, but true only with regard to distinct judges. At the same time, it is supposed to allow us to account for the disagreement between (1a)

6 Kölbel (2002), for instance, takes the phenomenon of faultless disagreement as "basic evidence" for relativism about PPTs.

and (1b), since there is no overall index of evaluation (including *w*, *t*, and *j*) in which both contents are true.

Russell anticipates Lasersohn's insight: personal tastes are, like desires, subjective. At the same time, however, communities play a large role in the determination of the truth-conditions for predicates of personal taste (Recanati 2007), and this fact has given rise to a number of pragmatic problems noted elsewhere in the literature. For example, Stojanovic (2007) argues that, on Lasersohn's view, it is most natural to assume that the judge for an utterance involving PPTs is the speaker. However, if that's true, there should be no reason for G.B.S. (in 1b) to think that Russell (in 1a) is expressing anything other than his individual opinion, so no reason to be disagreeing with him! In order to explain the intuition that disagreements in matters of taste are faultless, proponents of a relativistic view like Lasersohn's end up explaining away the presence of genuine disagreement.[7]

The force of Stojanovic's complaint is even clearer as applied to the case of comic predicates. Consider a case like (1) but which employs comic predicates in lieu of a PPT like *tasty*:

(2a) Webb: G.B.S. is hilarious!
(2b) Russell: No, he's not!

In (2), on a relativistic view, Webb should be understood as describing his own attitude toward G.B.S., and Russell as describing his own attitude. So while there is some intuitive sense in which they might appear to be disagreeing, they are in fact expressing subjective opinions that do not conflict. Exactly the materials the relativist requires in order to account for the intuition of faultlessness must be dispensed with in order to arrive at an account of their disagreement.

For these reasons and others, some philosophers and linguists have argued against the existence of faultless disagreements: attention to the apparent phenomenon, they argue, has had a distorting influence on recent semantic theorizing (Beukens 2011; Cappelen and Hawthorne 2009; Glanzberg 2009 Palmira 2014; Rovane 2012; Stojanovic 2007). According to theorists in this tradition, a disagreement is genuine when one party gets things right and someone is at fault, or it's only a misunderstanding of some kind and no one is at fault. Intuitions about apparently faultless disagreements, on these views, must be dispensed with: they just don't justify the radical departures from traditional semantic theory required by relativism.

Realism has mostly been neglected as a viable option in the case of PPTs, but Gibbard's view already goes some distance toward accounting for the appearance of faultless disagreement. On his view, when people appear to disagree about, for example, whether something

7 Contextualist approaches to predicates of personal taste share with Lasersohn's relativism a notion of context dependency that involves a judge, *j*, whose perspective determines the truth-values for predicates of personal taste. Given contextualism's problems accounting for disagreement, and the fact that the view is semantically equivalent to relativism (Stojanovic 2007), I will not discuss the option further here.

is *tasty*, or *fun*, they may be disagreeing, or they may just be expressing subjectively held opinions. For instance, Gibbard asks us to consider the differences between two cases involving *Yum!* and *Yech!* (Gibbard 2003, 279–80):

> You say "*Yum!*" to asparagus, suppose, and I say "*Yech!*" We may treat these voicings as working towards a joint plan for distributing food. Then there is no disagreement; clearly the asparagus goes to you. Or we may treat it as a critique of food, so that there is something to be resolved as we work toward a joint standard of taste in food. The same reactions can figure in logically different ways.

As Gibbard notes, whether or not it is appropriate to treat an exchange between discussants that has the outward characteristics of a dispute as a genuine disagreement will depend upon contextual specifics: it is often not enough to mark out the content of the apparent disagreement to tell whether or not it qualifies.

When confronted with a dispute about the funny, it is sometimes natural to characterize it as a disagreement in plan. This is perhaps most obvious in cases where something presented satirically is taken seriously. After running a story designed to satirize the budget battles in the US Congress in 2011, titled "Congress takes group of schoolchildren hostage," *The Onion* released a series of twitter "updates" on the breaking news that actually led to a Capitol Police Investigation. For another example, when fundamentalist Christians were condemning J. K. Rowling's Harry Potter novels as satanic, *The Onion* published a piece according to which scientists had found empirical confirmation of their negative influence. In response, the paper reportedly received hundreds of emails thanking the paper for "telling the truth." Religious leaders re-reported the story on Christian radio stations, and from the pulpit. The "finding" was discussed as a key resource for combating the satanic influence of Harry Potter novels in the context of a meeting of southern Christian leaders. Presumably those leaders disagreed with the writers about whether the piece was funny but did so because it was accorded moral and epistemic significance where none had been intended. They we disagreeing genuinely about whether the piece was funny because they were disagreeing about whether or not it called for a plan.

However, there are many other circumstances—I think the clear majority—in which comic disputes are not naturally characterized in terms of states of planning. While many comedians have claimed that jokes are funny *if* they work, there are all sorts of reasons that jokes fail, and most of those have nothing to do with the properties possessed by jokes themselves, or by the properties instantiated by their tellings. Judgments of funniness are highly dependent on levels of antecedent arousal: for instance, when people are first asked to read an erotic passage from a novel or a graphic description of torture, rather than passages that are less positively or negatively arousing, they subsequently rate jokes as funnier (Cantor et al. 1974). And as Cohen (1999) emphasizes, all jokes presuppose a substantial amount of background knowledge. Take the following joke:

> *The thing about German food is that no matter how much you eat, an hour later you're hungry for power.*

Cohen (1999, 21–22) insists, I think rightly, that

> the joke is largely unavailable to anyone who doesn't know the old chestnut about Chinese food invariably leaving one hungry after eating, whether one believes that about Chinese food or not. But then one must also know the commonplace about Germans that they long to control others and to wield power. Now it makes some difference whether one only knows this commonplace, or whether one knows it and believes it to be true. And finally, it matters whether one has negative feelings about Germans on that count, or doesn't. If it offends one to have Germans represented in this way, then the amusement may be lost altogether.

Even setting aside considerations about the extent to which comic disputes sometimes turn on brute contextual factors, or on failures of shared background knowledge, there are many cases in which a violation of expectations elicits mirth from one person, and fails to do so for another, for reasons that don't seem blameworthy in either direction. And these disputes seem to turn on differences in the broader normative systems that parties to the dispute endorse.

3. Comic Normativity

The reasons few comic disputes seem amenable to characterization in terms of plans seem distinctive to the comic domain. As Gibbard emphasizes, evaluative judgments are typically *actionable*: epistemic judgments concern what to believe; moral and prudential judgments concern what to do; aesthetic judgments concern what deserves sustained attention. Most experiences of humor, on the other hand, don't prompt the entertainment of new candidates for belief, or the revision of existing beliefs. They don't typically suggest plans, or different courses of action, and they demand attention only very briefly (Hildebrand et al. 2013). As Apter (1991: 31, 1982, 1984, 2001; cf. Koestler 1964) has suggested, when one is engaged humorously one adopts a certain static "state of mind, a way of seeing and being, a special mental 'set' towards the world and one's actions in it" that calls for nothing. Mirth, in fact, is strongly associated with a loss of control of normal abilities to act in goal-directed ways. In laughter, muscle tone decreases and, in extreme cases, it is accompanied by nonvoluntary tear production and incontinence (Paskind 1932)! This is hardly the behavioral profile associated with a goal-directed affect program (Morreall 1983, 2011).

Yet there remains a strong intuitive sense in which humor *is* motivational from the perspective of desire. Sharing jokes, for instance, is an important mechanism for the promotion of intimacy, as well as for ostracism. This quote, again from Cohen, should help to clarify (1999: 31):

> Of course I want you to like the one about Winnie the Pooh. I want you to like it because I like you and I want you to have something you like, and I want you to be grateful to me for supplying it. But I also need you to like it, because in this liking I receive a confirmation of my own liking. I put this by saying that the joke is *funny*, as if this were an objective matter.[8]

What is it to desire that others share your comic preferences? One attractive possibility is intimated by a suggestion, due to Kotzen (2015), that the concept of humor is distinctive among normative concepts because humor always implicates violations of the norms that are constitutive of other normative concepts. Kotzen surveys different categories of humor that, on his interpretation, involve violations of practical, epistemic, and aesthetic norms. For instance, he asks that you consider the following joke from Groucho Marx:

> *One morning I shot an elephant in my pajamas.*
> *How he got into my pajamas I'll never know.*

Kotzen points out that listeners first naturally interpret "in my pajamas" as modifying "I," while the second sentence is grammatically tractable only if "in my pajamas" modifies "elephant" in the first sentence. The reasons that this joke promotes intimacy when shared seem, at least in part, to do with the fact that listeners are independently making the same mistakes, and they are making those mistakes in full knowledge that given the norms of communication to which they all subscribe, they are understandable, forgivable, perhaps even appropriate.

Kotzen seems correct in his assessment that humor always involves the violation of the norms constitutive of other normative concepts. His account does not itself, however, imply an answer to the question of why only some such violations are humorous. This question has an analogue in long-standing debates about the nature of humor itself. Humor-related cognitions involve "lateral" or "divergent" forms of thinking: in humorous manifestations connections are made, albeit if only momentarily, between very disparate ideas or ways of representing the world (Morreall 1997). While most all humor theorists now agree that incongruity is a necessary feature of humor, they also agree that it is not sufficient, because many incongruous things fail to be amusing. Bain (1859, 257) gives a long list of good examples:

> There are many incongruities that may produce anything but a laugh. A decrepit man under a heavy burden, five loaves and two fishes among a multitude, and all unfitness and gross disproportion; an instrument out of tune, a fly in ointment, snow in May, Archimedes studying geometry in a siege, and all discordant things; a wolf in sheep's clothing, a breach of bargain, and falsehood in general; the multitude taking the law into their own hands, and everything

8 The joke is: *What do Winnie the Pooh and Alexander the Great have in common? Their middle name.*

> of the nature of disorder; a corpse at a feast, parental cruelty, filial ingratitude, and whatever is unnatural; the entire catalogue of vanities given by Solomon—are all incongruous, but they cause feelings of pain, anger, sadness, loathing, rather than mirth.

As Bain's examples intimate, incongruous violations are important across normative domains. Lyall (1855, 488) discusses the relationship between incongruity and morality at length:

> The morality of the action is something more than its incongruity. Many actions are incongruous which are not wrong, and excite no moral disapprobation. Whence the wrongness? Whence the moral disapprobation? The wrongness is the *moral* incongruity. And here all the peculiarity lies in the *moral element—moral* incongruity. Incongruity we can understand; inconsistency, unfitness, but what is moral in it—the element which allows us to call it *moral* incongruity? Which allows us to speak of it as wrong! This is the very point in the question.[9]

The importance of incongruity in aesthetic contexts has also been recognized. Consider some remarks by Poe (1845, 37–38) on the relevance of humorous and nonhumorous forms of incongruity to the fantastic:

> Fancy is at length found impinging upon the province of Fantasy. The votaries of this latter delight not only in novelty and unexpectedness of combination, but in the avoidance of proportions. The result is therefore abnormal and to a healthy mind affords less of pleasure through its novelty, than pain through incoherence. When, proceeding a step farther, however, Fantasy seeks not merely disproportionate but incongruous or antagonistical elements, the effect is rendered more pleasurable from its greater positiveness—there is an effort of Truth to shake from her that which is no property of hers—and we laugh.

Whatever one makes of the prospects for the incongruity theory, it is clear that the relationship between comic value and other kinds of value is complex and multifaceted. It is not accidental that many comedians are revered for making things funny that are typically regarded as blameworthy epistemic or moral faults. It's also true that issues falling squarely in the epistemic domain are sometimes comicalized—consider experiences of naïve gullibility or attempts to excuse failures of self-knowledge in humorous ways. When attempts at humor fail in ways that are associated with genuine disagreement, it seems true without exception

9 When presented with some instance of vicious action, Hume argued in similar spirit that it was impossible to find the viciousness in it: "The vice entirely escapes you, as long as you consider the object. You can never find it until you turn your reflection into your own breast, and find a sentiment of disapprobation, which arises in you, toward that action" (Hume 1978, 468–69).

that those failures involve other sorts of evaluations. Rape jokes are routinely rejected, for instance, on the grounds that moral forms of evaluation and moral action are called for in response. And jokes can fail because still other forms of normative evaluation seem called for. As Gordon (as interviewed in Wiator 1992, 84) reminds us, for instance, Hitchcock's *Psycho* was a failed attempt at humor:

> When Hitchcock referred to Psycho, he always referred to it as a comedy. It took seeing it three or four times before I started picking up on it as a comedy. He said that there was a very fine line between getting someone to laugh and getting someone to scream.

Understood in relation to all the other kinds of attitudes people might issue in response to violations of their expectations, mirth appears to function as a "catch-all" emotional response.[10] When people share a sense of humor, they must thereby also share a broad range of values and attitudes toward what counts as morally, epistemically, aesthetically, and even grammatically significant. The hypothesis developed in the rest of my discussion here is that in cases of faultless-looking comic disputes, some incongruous violation of expectations is recognized, but neither party to the dispute regards the violation as significant enough to warrant pressure on the broader normative system already in place. Such disputes may look like disagreements but only because there are attendant subjective desires for more closely shared norms of consistency among pro-attitudes.

This regulative role helps to explain why sharing a sense of humor with someone serves to signal intimacy and shared cultural beliefs. It makes sense, that is, of the reasons that the sense of humor "is so all inclusive and highly prized that to say of another: 'He has a grand sense of humor' is almost synonymous with 'He is intelligent, he's a good sport, and I like him immensely'" (Omwake 1939, 95).

4. Comic Disagreement

From this perspective, the difference between comic disputes that appear faultless and genuine cases of comic disagreement can be brought into focus. Some claims about comic value are subjective attitude reports that can be cashed out with a simple indexical semantics. On the basic view, the truth-values of atomic comic sentences are indexed to individual judges—typically speakers—and hence do not conflict in the ways required to sustain genuine comic disagreement. In cases of genuine comic disagreement, a violation of expectations is not regarded as trivial by one party to the dispute, who is then disposed to advocate for a plan. Genuine comic disagreements in this way are always disagreements in plan, and they

10 Similar discussions regarding the relevance of incongruity appear in epistemic (Popper 1959) and aesthetic contexts (Carroll 1999; Santanaya 1896, sec. 62–64), and in discussions of norms and normative concepts in the psychological literatures (Carey 2011; Piaget 1932; Turiel 1994).

are always contrastive. If someone genuinely claims that something is not funny, they must also be expressing the attitude that the violation is normatively significant in some other domain.

The ordinary semantic content of an assertive utterance like *That's funny* can be given by a set of possible worlds; expressive elements provide corollary resources for determining the contexts in which uses of the term *funny* to mark experiences of amusement are appropriate. The source of disagreements about the funny in these cases can attach to the semantic-level content but requires additional contextual information that is provided by an expressive dimension signaling approval or disapproval.

By way of an example, consider this interpretation of the exchange first outlined in (2)

(3a) Webb: G.B.S. is hilarious → *G.B.S. is hilarious for Webb*.
(3b) Russell: G.B.S. is not hilarious → *G.B.S. is not hilarious for Russell*.

Absent expressive content, the exchange can be interpreted truth-conditionally in ways that do not imply genuine disagreement: the propositions expressed in (3a) and (3b) are both true statements about personal comic preferences and are, hence, compatible. This follows directly from an indexical interpretation of what's going on in the exchange. It also explains why in so many contexts the following kind of exchange is awkward:

(4a) Webb: G.B.S. is hilarious to me.
(4b) Russell: ## No he's not!

In other contexts, however, an interpretation of comic sentences as involving use-conditional expressions in addition to truth-conditional expressions is appropriate. Typically, following Potts (2005, 2007), the default judge in such utterances will be the speaker, but in the comic domain it appears that additional expressive dimensions can provide information suggesting that other relativization points are more appropriate. Taking into account the presence of such expressive dimensions suggests that the same exchange can be analyzed as follows:

(5a) Webb: G.B.S. is hilarious → *Mirth is an appropriate response to G.B.S. (yay!)*
(5b) Russell: G.B.S. is not hilarious → *Mirth is not an appropriate response to G.B.S. (boo!)*

More schematically, it's possible to think about the truth and use conditions for this example as follows: *G.B.S. is hilarious* is true in $\langle c, w \rangle$ if the judge in c likes the antics of G.B.S., and also expresses the desire that others experience his jeering in the same way. Genuine comic disagreement, on this reading, is generated by a use-conditional layer that

expresses approval or disapproval, but not independently of the truth-conditional layer. Cases of *apparent* comic disagreement, in contrast, involve the presence of a separate, desire-focused expressive index. In these cases, people are reporting on their subjective attitudes toward some putative object of amusement, while expressing desires about the extent to which broader normative systems are shared.

Noncognitivists are thought to owe a semantics for evaluative terms that gives meaning in cases of unembedded predication in a way that is continuous with the account required for complex sentences involving the same predicates. As Yalcin (2012, 123–24) has emphasized: while the expressivist thesis is taken to be an "empirical thesis about some fragment of natural language, a fragment including normative vocabulary" from a serious linguistic perspective "metaethical expressivism is not in the game."

One major source of difficulty is that expressivism typically treats claims like "That's funny" as expressions of attitudes. In embedded contexts, however, this is not a natural understanding of what such claims mean. If, for example, a claim about humor is embedded in a conditional like "*I wonder if that's funny*" the expressive force supposedly associated with the term *funny* in the atomic context doesn't transfer. Traditionally, expressivists have attempted to handle this problem by developing "logics of attitudes" that aim to account for relations of implication between different judgments, or sentences. This can't work in the case of comic predicates, however, because they don't always behave the same way as typical expressives when embedded or under negation. Consider, for example:

(6) If bicycle crashes are funny, then G.B.S. crashing his bicycle into Russell's is funny.
(7) If lying is wrong, then it is wrong to teach your children that lying is permissible.

The speaker in (6) is making no claims about whether bicycle crashes are funny, any more than the speaker of (7) is making any claims about the permissibility or wrongness of lying. This may suggest that it's wrong to think about funny as an expressive term, since expressives typically escape the force of conditionalization. For example, consider:

(8) If he publishes another paper, that *damn* Kaplan will get promoted.
(9) If he books another gig, that *clown* Gallagher will get a network deal.

The emotive force of *damn* in (8) and of *clown* in (9) escape the force of conditionalization. And in cases where purely expressive terms are negated, the expressive force of those terms also appears to escape:

(10) That *damn* Kaplan did not get promoted.
(11) That *clown* Gallagher did not get a network deal.

In the case of comic predicates like funny, in contrast, it appears that negation sometimes target both the truth-conditional (a) and use-conditional (b) elements.

(12) Russell: The jeering isn't funny.
   a. The jeering is not funny for Russell.
   b. The jeering is not funny for Russell → *The jeering ought not be counted as funny*.

This hybrid view makes sense of the behavior of comic terms under negation and conditionalization. For instance, consider:

(13) Webb: "G.B.S. is funny."
   a. G.B.S. is funny for Webb.
   b. G.B.S. is funny for Webb → *G.B.S. ought to count as funny*
   c. Would it that G.B.S. counted as funny for others.

A combination of the truth-conditions and use-conditions explains how it is possible to derive a positive attitude from Webb's utterance. Likewise, in the case of negation, considering both truth and use conditions allows for the determination of a negative attitude. For instance,

(14) Russell: "G.B.S. is not funny!"
   a. *not* (G.B.S. is not funny for Russell)
   b. G.B.S is funny for Russell → *G.B.S. ought to count as funny*
   c. *not* (G.B.S. ought to count as funny)

Finally, a combination of use- and truth-conditional elements can be appealed to in order to explain why when comic predicates are used as antecedents under conditionalization, they sometimes fail to implicate expressive elements.

(15) Russell: "If G.B.S. is funny, then I'll laugh when he crashes his bicycle."
   a. If G.B.S. is funny for Russell, Russell will laugh when he crashes his bicycle.
   b. G.B.S. is funny for Russell → *G.B.S. ought to count as funny*
   c. ## G.B.S. ought to count as funny

On this view, statements like "That's not funny" sometimes function descriptively as subjective attitude reports that can be cashed out in the terms of a simple, indexical semantics. On their use-conditional layers, however, they can express attitudes concerning what ought to be counted as funny in the context of the utterance, and in other circumstances to express desires that others shared the same standards for making such determinations.

Genuine Disagreements, on this model, obtain at the level of attitudes about the appropriateness of mirthful responses in the context of the overall normative system, rather than at the level of truth-conditional content. Intuitions about apparently faultless disputes are explained in terms of desires concerning the extent to which systems of norms are shared, even where there is no disagreement at the level of truth-conditional content or at the level of attitudinal approvals and disapprovals involved in decisions about what to do. Conceptualizing comic predicates as involving multiple use-conditional layers appears to solve some of the problems accounting for their behavior, while leaving open a number of different semantic and syntactic options.[11]

I have not said anything so far about the formal aspects of the desire-focused expressive dimension hypothesized to be implicated the sorts of comic disputes that appear faultless. There are, however, independent reasons for thinking that expressives are not a distinctively natural class, and that there may be multiple varieties. For instance, Geurts (2007) has suggested that many epithets are used predicatively in ways that challenge a clean division of the lexicon into expressive and descriptive categories. When two people use the same epithet, one descriptively and one expressively, they still intuitively disagree.

One attractive possibility is that the shift from the descriptivist function of comic terms to function of expressing desires for agreement in norms of consistency involves a broader relativization of the judge parameter. Wolf and Cohen (2011 cf. Recanati 2007) treat *clear* as a predicate that can sometimes become "objectivized" in such a way that its truth-conditions depend upon the opinions of whatever group of people a speaker considers competent or relevant in context. Like comic terms and PPTs, evaluatives like *clear* in the epistemic domain are gradable (cf. Cohen 2010):

(16) It is very/reasonably/sort of clear that G.B.S. is funny.

Clear also allows for comparatives and superlatives (clear, clearer, clearest; like funny, funnier, funniest) and can also be modified by an overt experiencer:

(17) It is clear to me/to you/to Webb that G.B.S. is funny.

And *clear* also gives rise to apparently faultless disagreements, such as in

(18a) Webb: It's clear that G.B.S. is hilarious.
(18b) Russell: No, it's not!

11 I've not belabored formal details for two reasons. The first is that hybrid expressivist approaches are linguistically abstract and remain the subject of scant attention from formal linguists. The second, however, is that the basic idea stands or falls in ways unrelated to decisions about how to cash out the formal details. As Potts (2005, 2007) and Guntzmann (2012, 2016) have emphasized, hybrid templates can be cashed out in pretty much any formal framework.

In at least some circumstances, a generic reading of a hidden judge argument may be required to explain the functioning of epistemic evaluatives. To claim that *It should be clear* would be to express a broader desire that epistemic standards between discussants are shared, rather than a disagreement about what plan ought to be pursued in the service of arriving at shared standards. This, perhaps, is one reason that feelings of futility are so often associated with failed attempts at humor.

In his arguments against pursuing expressivist strategies in the case of PPTs, Lasersohn (2005) considers the possibility that the judge index should be read generically, but rejects it on the basis of an example in which someone is having fun cataloging their paperclip collection—an activity that is not easily regarded as fun *in general*:

(19) John: This is fun!
(20) John: ## This is not fun at all, although I'm having fun doing it.

(20), Lasersohn reasons, is something that John *might* say if the hidden judge variable is always generically quantified, but the claim is infelicitous. However, if the hidden variable is taken to have several settings like it does in the case of clear—one specific that is typically assigned to the speaker and that is implicated in expressions of her personally held attitudes, and another layer associated with the expression of desires that can take on a generic value—the problem Lasersohn identifies disappears. If the variable is understood to have a default setting according to which the judge is the speaker, that expressive dimension may serve as a signal to change the judge value, for instance to the community of people one judges to be witty, or wise, or even to everyone.

Another interesting possibility concerns the behavior of the English predicate *one*, which functions as a generic that allows a speaker to self-ascribe a property while, at the same time, "projecting himself onto everyone in the relevant group" (Moltmann 2004, 20). For instance, consider:

(21) One can tell more sophisticated jokes → *I can tell more sophisticated jokes*.
(22) People can tell more sophisticated jokes → *I can/can't tell more sophisticated jokes*.

The generic use of *one* in (21) implies that the capacity to tell more sophisticated jokes is possessed by the speaker, but the claim that people *in general* can, in (22), leaves open the possibility that the speaker is an exception.

Moltmann (2010) argues that generic "one" has an unpronounced counterpart syntacticians call "arbitrary PRO" (PRO_{arb}) that corresponds to the empty subject position. The connection between them sometimes licenses a generalization on the basis of a speaker's experience or action that takes roughly the following form: for anyone *x*, *x* can tell more

sophisticated jokes. On Moltmann's analysis, PRO_{arb} implicates two independent strategies for the kind of generalizing self-reference implicated in usage of one, which may be called on to independently explain the involvement of PRO_{arb} in certain uses of moral and epistemic predicates like:

(23) (PRO_{arb}) To help others is a virtue.
(24) One cannot be mistaken about the content of one's own mental states.

Perhaps, as Moltmann has suggested, the range of meanings facilitated by PRO_{arb} changes when sentences involving comic predicates implicate expressive indices that do not assign the speaker the role of judge. On this view, the introduction of expressive elements that communicate desires would sometimes be connected to the adoption of a more general interpretation of the pronouns under consideration. One interesting question concerns whether this approach to disagreement and apparently faultless disputes in the comic domain might be adapted more generally, especially since similar approaches to the problem of faultless disagreement have been put forward in the aesthetics literature (Sundell 2011), in the case of epistemic modals (Stephenson 2007), and evaluative discourse more generally (Barker 2002, 2013).

5. Directions

As Price has emphasized (Macarthur and Price 2007; Price 1988, 1996, 2011), expressivism is an attractive and plausible view globally, despite semantic difficulties. The account sketched in the foregoing claims the semantic virtues of hybrid expressivist theories, since it claims all the resources of traditional truth-conditional semantics in addition to the contextual information provided by a classical expressive dimension. But it also provides a plausible account of the reasons that, as some have recently noted, it seems plausible to regard the differences between descriptive and expressive usages of normative terms along several different dimensions and degrees, rather than simple differences in kind. For example, Smith (1993, 245; cf. Barker 2011; Boisvert 2008; Copp 2001; Gert 2007, 95) has suggested that response-dependent concepts should in general be expected to fall on a continuum from obviously representational to less obviously, or obviously not, representational. It has also been suggested that some but not all such usages involve an assertive-expressive dimension (Bar-on and Long 2001; Boisvert 2008; Buekens 2011; Von Fintel and Gillies 2007).

By hypothesis, I've suggested that some purely expressive uses involve the expression of subjectively held attitudes, and that some others involve the expression of desires that norms of consistency among pro-attitudes are more widely shared. This approach suggests a way of accounting, in addition, for the differences between normative concepts that arguably belong in the same class or domain. Following Sweetser (1990), Foolen (1997, 22–25) has suggested that there are multiple domains in which descriptive linguistic forms functionally

"shift" to normative or evaluative level at which the possession of reasons is implied by contextual use conditions. For example, he argues that when a speaker says *John must be ill*, they are indicating that they are in the possession of reasons for believing that John is ill.

In some cases, standard descriptive uses are lost altogether. Some recent work on the semantics of slurs indicates that "purely expressive" linguistic usage is not empirically defensible. However, it seems right to think that some indicators in the comic domain as purely expressive, and possessed of normative force. People laugh more when together than when alone. And people laugh in different registers. The *groan* seems a very good candidate for the purely expressive communication of comic disapproval.

These considerations suggest a view of evaluative discourse on a continuum from the descriptive to the normative. In the clearly representational cases such as those in the perceptual domain, for example, agreement about standards for belief are widespread (consider: *I know, I saw it! Really, I was there!*). Cognitive capacities like vision, audition, and touch are remarkably widely instantiated, and they function in ways that are relatively impervious to the influence of cultural norms and the vagaries of individual experience. It makes sense, then, that descriptive uses in these cases are rarely coupled with expressive layers that indicate desires for joint standards. In domains like the comic where standards are widely acknowledged to be sensitive to culturally available norms, in contrast, one would expect faultless-looking disputes and associated expressions of desires for shared standards.

6. Conclusion

I have suggested that comic disagreements are at least sometimes genuine, and that apparently faultless comic disputes implicate expressive dimensions that communicate desires for shared normative standards. When people disagree genuinely about whether something is funny, they are disagreeing about whether or not mirth is an appropriate response to an incongruous violation of expectations, rather than some other plan-oriented response. In cases where comic disputes appear faultless, it is because one party to the dispute is unwilling to dismiss the putatively comic violation as unimportant from the perspective of other normative domains, in just the sense that Russell may have desired that G. B. S. shared his views about the distinction between wit and wisdom. Intuitions about comic disputes that have the apparent structure of disagreements seem very well accounted for on the basis of desires that the norms constitutive of other normative concepts are more closely shared between parties to the dispute.

References

Apter, M.J. (1991). "A Structural-Phenomenology of Play." In *Adult play: A reversal theory approach*, edited by J. H. Kerr & Michael J. Apter, 13-29. Amsterdam: Swets & Zeitlinger.

——— (1982). *The Experience of Motivation: The Theory of Psychological Reversals*. New York: Academic Press.

——— (1984). "Reversal theory, cognitive synergy and the arts." In *Cognitive Processes in the Perception of Art* edited by W. R. Crozier & A. J. Chapman, 411-426. North Holland: Elsevier.

——— (2001). Apter, M. J. *Motivational Styles in Everyday Life: A Guide to Reversal Theory*. Washington D.C.: American Psychological Association.

Bain, A. (1859). *The Emotions and the Will*. London: J.W. Parker.

Bar-On, D., & Long, D. C. (2001). "Avowals and First-person Privilege." *Philosophy and Phenomenological Research*, *62*(2), 311-335.

Barker, C., (2002), "The Dynamics of Vagueness." *Linguistics and Philosophy* 25, 1–36.

——— (2013), "Negotiating Taste." *Inquiry: An Interdisciplinary Journal of Philosophy* 56, no. 2, 240–257.

Bergson, H. (1900/2007). *Laughter: An Essay on the Meaning of the Comic*. Cincinnati, OH: Standard Publishing Inc.

Buekens, F. (2011). "Faultless Disagreement, Assertions and the Affective-expressive Dimension of Judgments of Taste." *Philosophia*, *39* no. 4: 637-655.

Boisvert, D. R. (2008). "Expressive-Assertivism." *Pacific Philosophical Quarterly* 89, no. 2: 169–203.

Brown, C. (2013). *Hello Goodbye Hello: A Circle of 101 Remarkable Meetings*. New York: Simon and Schuster.

Cantor, J. R., Bryant, J., & Zillmann, D. (1974). Enhancement of humor appreciation by transferred excitation. *Journal of Personality and Social Psychology*, *30* no. 6, 812-821.

Cappelen, H., and J. Hawthorne (2009). *Relativism and Monadic Truth*. Oxford: Oxford University Press.

Carroll, N. (1999). "Horror and Humor." *Journal of Aesthetics and Art Criticism* 57: 145–60.

Cohen, A. (2010). "A tasty mixture." Paper presented at the Subjective meaning workshop: Alternatives to relativism, Berlin.

Cohen, T. (1999). *Jokes: Philosophical Thoughts on Joking Matters*. Chicago, IL: University of Chicago Press.

Copp, D. (2001). *Morality, Normativity, and Society*. New York: Oxford University Press.

D'Arms, J., and D. Jacobson (2000). "On the Moralistic Fallacy: On the 'appropriateness of emotions,'" *Philosophy and Phenomenological Research* 61: 65–90.

de Sousa, R. (1990). "When Is It Wrong to Laugh?" In *The Rationality of Emotion*, edited by R. de Sousa. Cambridge: MIT Press.

Egan, A. (2014). "There's Something Funny about Comedy: A Case Study in Faultless Disagreement." *Erkenntnis* 79, no. 1: 73–100.

von Fintel, K., & Gillies, A. (2007). "An Opinionated Guide to Epistemic Modality." *Oxford Studies in Epistemology*, *2*, 32-62.

Foolen, A. (1997). "The Expressive Function of Language: Towards a Cognitive Semantic Approach." In *The language of Emotions: Conceptualization, Expression, and Theoretical Foundation* edited by S. Niemeier and R. Dirven, 15-32. Amsterdam: John Benjamins Publishing Company.

Gaut, B. N. (1998). "Just Joking: The Ethics and Aesthetics of Humor." *Philosophy and Literature* 22, no. 1: 51–68.

Gert, J. (2007). "Cognitivism, Expressivism, and Agreement in Response." *Oxford Studies in Metaethics*, *2*: 77-110.

Geurts, B. (2007). "Really Fucking Brilliant." *Theoretical Linguistics* 33, no. 2: 209-214.

Gibbard, A. (1992). *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.

——— (2007). "Thinking How to Live with Each Other." *Tanner Lectures on Human Values* 27: 165.

——— (2009). *Thinking How to Live*. Cambridge: Harvard University Press.

Glanzberg, M. (2009). "Semantics and Truth Relative to a World." *Synthese* 166, no. 2: 281–307.

Gutzmann, D. (2016). "If Expressivism is Fun, Go for It." In *Subjective Meaning: Alternatives to Relativism*, edited by C. Meier and J. van Wijnberger-Huitink, 21–46. Berlin: de Gruyter.

Hildebrand, K. D., & Smith, S. D. (2014). "Attentional biases toward humor: Separate effects of incongruity detection and resolution." *Motivation and Emotion*, *38* no. 2: 287-296.

Hurley, M., D. C. Dennett and R. B. Adams Jr. (2011). *Inside Jokes: Using Humor to Reverse Engineer the Mind*. Cambridge, MA: MIT Press.

Kant, I. (1790/2000). *Critique of the Power of Judgment*. New York: Cambridge University Press.

Kaplan, David (1989). "Demonstratives. An Essay on the Semantics, Logic, Metaphysics, and Epistemology of Demonstratives and other Indexicals." In *Themes from Kaplan*, edited by Joseph Almog, John Perry and Howard Wettstein, 481–563. Oxford: Oxford University Press.

——— (1999/2004). "The Meaning of Ouch and Oops. Explorations in the Theory of Meaning as Use" [2004 version—Ms. Los Angeles].

Koestler, A. (1964). *The Act of Creation: A Study of the Conscious and Unconscious Processes of Humor, Scientific Discovery and Art*. New York: Dell.

Kölbel, M. (2002). *Truth without Objectivity*. London: Routledge.

——— (2004). "Faultless Disagreement." *Proceedings of the Aristotelian Society* 104, no. 1: 53–73.

Kotzen, M. (2015). "The Normativity of Humor." *Philosophical Issues*, *25*, 396-414.

Lasersohn, P. (2005). "Context Dependence, Disagreement, and Predicates of Personal Taste." *Linguistics and Philosophy* 28: 643–86.

——— (2009). "Relative Truth, Speaker Commitment, and Control of Implicit Arguments." *Synthese* 166, no. 2: 359–74.

Lyall, W. (1855). *Intellect, the Emotions, and the Moral Nature*. Edinburgh: T. Constable and Company.

Macarthur, D., & Price, H. (2007). Pragmatism, quasi-realism and the global challenge. *New pragmatists*, *91*: 93-94.

Moltmann, F. (2004). "Properties and Kinds of Tropes: New Linguistic Facts and Old Philosophical Insights." Mind, 113 no. 449: 1-41.

——— (2010). "Relative Truth and the First Person." *Philosophical Studies* 150, no. 2: 187-220.

Morreall, J. (1983). *Taking Laughter Seriously*. Albany, NY: SUNY Press.

——— (2011). *Comic Relief: A Comprehensive Philosophy of Humor*. Malden, MA: Wiley-Blackwell.

Omwake, L. (1939). "Factors Influencing the Sense of Humor." *The Journal of Social Psychology*, *10* no. 1: 95-104.

Palmira, M. (2014). "The Semantic Significance of Faultless Disagreement." *Pacific Philosophical Quarterly*, 96 no. 3: 349–71.

Paskind, H. A. (1932). "Effect of Laughter on Muscle Tone." *Archives of Neurology & Psychiatry* 28, no. 3: 623–28.

Piaget, J. (1932). *The Moral Development of the Child*. London: Keegan Paul.

Poe, E. A. (1845). *Broadway Journal, No. 3*—January 18, 1845.

Popper, K. R. (1959). "The Propensity Interpretation of Probability." *British Journal for the Philosophy of Science* 10, no. 37: 25–42.

Potts, C. (2005). *The Logic of Conventional Implicatures*. Oxford: Oxford University Press.

——— (2007). "The Centrality of Expressive Indices." *Theoretical Linguistics* 33, no. 2: 255–68.

Price, H. (1988). *Facts and the Function of Truth*. Oxford, UK: Basil Blackwell.

——— (2011). "Expressivism for Two Voices." In *Pragmatism, Science and Naturalism* edited by J. Knowles and H. Rydenfelt, 87-113. Berlin: Peter Lang.

——— (1996). "Essays in Quasi-Realism." *Philosophy and Phenomenological Research*, *56* no. 4, 965-968.

Recanati, F. (2007). *Perspectival Thought: A Plea for (Moderate) Relativism: A Plea for (Moderate) Relativism*. Oxford: Oxford University Press.

Rovane, C. (2012). "How to Formulate Relativism." In *Mind, Meaning, and Knowledge: Themes from the Philosophy of Crispin Wright*, edited by A. Coliva, 238–66. Oxford: Oxford University Press.

Russell, B. (1935). *Religion and Science*. New York: Holt.

——— (1938). *Power: A New Social Analysis*. New York: Norton.

——— (1956/1998). *Autobiography*. New York: Routledge.

——— (1957). *Why I am not a Christian: and Other Essays on Religion and Related Subjects*. London: G. Allen & Unwin.

——— (1988). *The Collected Papers of Bertrand Russell Vol. 9*: *Essays on Language, Mind and Matter, 1919–26*. London: Unwin Hyman.

Santanaya, G. (1896). *The Sense of Beauty*. New York: Scribner's.

Shoemaker, D. (2015). *Responsibility from the Margins*. New York: Oxford University Press.

Smith, M. (1993). "Objectivity and Moral Realism: On the Significance of the Phenomenology of Moral Experience." In *Reality, Representation, and Projection*, edited by J. Haldane & C. Wright, 235-256. Oxford UK: Oxford University Press.

Smuts, A. (2010). "The Ethics of Humor: Can Your Sense of Humor be Wrong?" *Ethical Theory and Moral Practice* 13: 333–47.

Sripada, C. S., and S. Stich (2006). "A Framework for the Psychology of Norms." In *Evolution and Cognition. The Innate Mind Vol. 2. Culture and cognition*, edited by P. Carruthers, S. Laurence, and S. Stich, 280–301. New York: Oxford University Press.

Stojanovic, I. (2007). "Talking about Taste: Disagreement, Implicit Arguments, and Relative Truth." *Linguistics and Philosophy* 30, no. 6: 691–706.

Sundell, T. (2011). "Disagreements about Taste." *Philosophical Studies* 155, no. 2, 267-288.

Sweetser, E. (1990). *From Etymology to Pragmatics: Metaphorical and Cultural Aspects of Semantic Structure*. New York: Cambridge University Press.

Wiator, S. (1992). *Dark Visions*. New York: Avon.

Wolf, L., & Cohen, A. (2011). "Clarity as Objectivized Belief." In *Vagueness and language use* edited by Égré, P., & Klinedinst, N., 165-190. London, UK: Palgrave Macmillan.

Yalcin, S. (2012). "Bayesian Expressivism." In *Proceedings of the Aristotelian Society* 112, no. 2, 12-160. Oxford, UK: Oxford University Press.

15

EXPRESSIVISM AND OBJECTIVITY

Peter Railton

Introduction

Once, when Christine Korsgaard was about to begin a talk to the Philosophy Department at Michigan, she noticed that no water had been provided at the lectern. She glanced my way, making a drinking gesture, and I took off to find something. In the Department kitchen, I managed to find an unused cup, but filling it I realized that it was clearly too small to last through the whole colloquium. So, I grabbed a water-filtering jug from the fridge and brought that along, too. Korsgaard had begun her talk by the time I got back, and I tried as inconspicuously as I could to circle around to the front of the room and slip the cup and jug onto the table beside the lectern without interrupting her. But Korsgaard interrupted herself—she stopped, picked up the cup in one hand and the jug in the other, stared at them, and said, deadpan, "Isn't this just like Michigan? You ask for a cup of water and they give you a *system*."

She wasn't wrong. It had long been a feature of some of the more conspicuous Michigan philosophers to be unhappy with an answer to a philosophical question until a system could be built to support it. No Michigan philosopher has put this proclivity to better use than Allan Gibbard, who has shown us time and again that an answer many may find initially unsatisfactory can be fit into a systematic account that establishes how good an answer it really is. And no example of this is clearer than his boldly taking up the banner of expressivism at a time when very few in the ethics community thought such a view could be viable. "After all," people thought, "isn't it subject to a half-dozen well-known and convincing counter-arguments? Arguments we repeat regularly in our undergraduate ethics courses?" It seemed to Gibbard, however, that there was something profoundly right about what he took to be the core insights of expressivism—and that people hadn't really seen this.

Accordingly, he set about showing how expressivism can provide a philosophically explanatory and empirically plausible approach to moral thought and discourse. Gibbard took a step back from moral thought and discourse as we now find it, and presented a compelling account of how normative thought and discourse might arise in response to the needs and aims of humans trying to get along with each other and get on in the world. He opened *Wise Choices, Apt Feelings* with the anthropologist Lorna Marshall's arresting image of conversation flowing like a brook through a !Kung village, night and day, now loudly, now quietly, on subjects large and small. An important part of that conversation, Gibbard argued, would be—and, really, would have to be—normative. Everyone needs to figure out what to say or do, how to react to what others are saying or doing, what to hope for or fear, and whom to rely upon for what. Whether trying to divide the capture of a hunting party, or deciding whether gathering firewood is more urgent right than grinding manioc, or thinking through how to deal with a child or neighbor who has become violent, normative questions are of pressing concern, and persist even after we have gathered or agreed upon relevant facts. Gibbard wanted an analysis of normative concepts that could make sense of the !Kung's discussions and thinking as much as our own, and explain why thinking and talking about *what is to be felt or done*—which feelings are apt, or choices are wise—is just as indispensable to recognizably human lives as thinking and talking about *what is the case*.

From taking this perspective, there emerged a firm conviction that the guiding thread for understanding normativity is that people need to have ways of thinking or talking, agreeing or disagreeing, that are by their nature linked to how they will go on to feel and act—such that *making up our minds* in arriving at judgments of these kinds is no idle exercise, but settling what to think and do. Normative judgments can play the indispensable role they do because we cannot make such judgments while remaining entirely indifferent as to whether we ourselves, or others, feel or act accordingly. A term of art for this idea is *motivational judgment internalism* (Darwall et al. 1992)—somehow normative thought and discourse need to be connected *by their nature* to motivation and action.

1. An Evolving Approach

In Gibbard's first book, *Wise Choices, Apt Feelings* (1990), he took as the focus of his analysis the normative concept <*rational*>, which he glossed in terms of "what makes sense." Deciding what's rational is making up one's mind about what it makes sense to think, do, or feel. Gibbard's view is called "expressivist" because, rather than giving an outright analysis of <*rational*>, it gives an account of what it is to *think* something rational, and then analyzes rational judgment as the *expression* of such states. Of course, "making up one's mind about what makes sense" does not require freedom from doubt—one can settle one's mind that, at least for now, it makes sense to feel ambivalent, or to act tentatively. But genuinely making up one's mind, whether with high certainty or substantial uncertainty, does bring oneself into a state such that one has some internally motivated tendency to act accordingly.

By ascending from the moral concepts of traditional metaethics to the more generic level of rationality, Gibbard made it possible to raise meaningful questions about whether or how much to go in for moral assessment, or how moral concerns should relate to those of other domains, such as prudence, personal or group loyalty, aesthetics, epistemology, and so on. This contributed to a spreading recognition within the philosophical community that *normativity* can be found pervasively in our lives and thought, well beyond the scope of traditional metaethics. Philosophers in epistemology, philosophy of language, philosophy of mind, metaphysics, and beyond recognized that their particular domain was as deeply involved in normativity as ethics or moral philosophy, and the idea of looking for a common way of understanding normativity gained ground. *Wise Choices, Apt Feelings* became an essential point of reference in philosophical discussion—in the new landscape of "metanormative" inquiry, it offered a clear and compelling articulation of how normative thought and language might work. Along with Simon Blackburn, Gibbard breathed new life into expressivism, and forced those who would reject expressivism to find new arguments or strengthen old ones. At the same time, his discussion of the aptness of feelings took philosophical understanding of the nature and role of emotion to a new level, and helped give impetus to another movement in metanormative inquiry, which came to be known as "fitting attitude" theories. In that literature as well as the literature on expressivism, *Wise Choices, Apt Feelings* remains to this day a foundational contribution, continuing to bear fruit.

By the time he wrote his second book, *Thinking How to Live* (2003), Gibbard had broadened his focus still further, folding judgments of what it is *rational* think, feel, or do under a more general framework in which one is asking, simply, *What to do? or think? or feel?* "Normativity" had gone from being an obscure word—in the late 1980s an editor persuaded me to drop the word "normativity" from the proposed title of a paper of mine because no one would recognize it—to becoming a perfectly respectable and even common answer for philosophers to give when asked what problems they are working on. But the idea of normativity remained unclear, with wide variation even on the question of what a satisfactory account of normativity would accomplish. In this confusing profusion, Gibbard's work stood out as a clear-eyed way of thinking about normativity, arising from our need to plan for the future, individually and socially.

It seems clear why planning should depend upon the facts, viewed objectively, so this approach helps us to understand why normative thought *supervenes* on nonnormative reality. And, since we can make good sense of the idea of conflict or agreement in plans, we can explain normative disagreement or agreement without adding to the world nonnatural states of affairs that are the objects of our disagreements or agreements. The natural world is world enough, so long as we have normative ways of thinking about it—concepts that are not simply reducible to nonnormative concepts, and that have a special relation to thought and action. The view thus mimics aspects of the nonnaturalism of G. E. Moore and others, but without the metaphysics, and fills in the blanks earlier nonnaturalists left between thought and action.

A similar picture could be applied at higher-order levels as well. What is it different approaches to metanormative thought are differing about? Not some *fact* in the world that furnishes a common focus, but a normative question of how best to interpret normative thought, language, and action. A metanormative position becomes in effect a stance, an interpretive plan. This makes it clear, however, that settling metanormative disputes can't be accomplished by *analysis* alone—substantive stances must be taken. In *Meaning and Normativity* (2012), Gibbard pushed the boundaries of metanormative inquiry still further, by focusing on this question of how such inquiry can proceed once we recognize that meaning itself is part of the normative fabric. Few philosophers had ventured far in the direction of giving an account of meaning as a normative phenomenon, and so Gibbard mostly had to blaze his own trail. While *Meaning and Normativity* was in the works, he could regularly be found lost in thought, exploring a complex maze of possibilities full of paths that turn out to be dead ends or uninformative circles. "Some of my friends said it couldn't be done," he wrote in the preface to *Meaning and Normativity*, "and for much of two decades I worked to prove them wrong" (2012).

A true Michigan philosopher in his commitment to systematic thought, Gibbard was nonetheless not someone who constantly elaborated a fixed system. His thought underwent continuing evolution, and at no point did he represent his picture as final or, for that matter, fully thought through—even by himself. He often introduces a hypothesis with phrases like, "We might try saying . . . ," or "We can try saying . . . ," or "To my ear. . . ." He was scrupulous about documenting how and why his views had changed, and generous in giving credit to those who had influenced his thinking. Gibbard's is a living body of work, restlessly moving forward. He is one great thinker who actually listens, and he builds his systems step-by-step, noting how far he has come but also how far he still has to go. He is, moreover, a writer. He works hard to formulate his views in ways that are at the same time eloquent and accessible. Even those who have not converted to his views—as indeed most working in contemporary metaethics or philosophy of language and mind have not—recognize that Gibbard's work poses a fundamental challenge. Here is someone who seems to know what he is doing in giving an account of normativity—Can those who would defend other positions develop anything with the clarity, scope, empirical plausibility, and explanatory power of Gibbard's work? I'm not aware of anyone who has.

2. Some Bedrock Convictions

Internalism in some form or other is a widely shared position in meta-normative inquiry, but Gibbard's work is partly made distinctive by the fact that it is allied with another guiding concern: Gibbard is a person of deep and serious normative commitments, but not simply someone who seeks to express his views, the way that an expletive expresses one's surprise or fear—he wanted to do justice to the *thought*, personal and shared, that goes into our forming or revising our normative commitments. For example, for Gibbard the "Frege-Geach

problem" is not merely a technical matter, part of tidying up the expressivist interpretation. Rather, he seeks a solution that would show how we are able to *reason* normatively with ourselves and others. This doesn't mean excluding emotion, but finding a way of understanding how people reason about, and with, emotion.

One striking example of Gibbard's normative commitment is the way that he is not intimidated by an intellectual climate in which anything that smacks of "absoluteness" raises eyebrows. He is unembarrassed to use the term "wise" in his own voice, and refuses to cede the term "rational" to those who would confine it to some form of internal consistency, whether the "thin" idea of consistency found among economists and decision theorists or the "thick" idea of consistency found among many ethicists. Gibbard can be seen in action at colloquia and conferences, gently but firmly insisting that some ends simply aren't worth wanting, even if one *could* form and pursue them without inconsistency. Economists and decision theorists in the audience tend to be baffled to hear someone they thought of as one of their own (Gibbard has done highly influential work in decision theory, recognized by election to the Econometric Society) arguing that some coherent utility functions simply make more sense than others. And the philosophers in the audience seemed puzzled to hear Gibbard (given his friendliness to utilitarianism in some form) appeal directly to substantive, basic normative intuitions about particular cases. For Gibbard, if thin or thick consistency considerations can't identify anything awry with an ideally coherent Caligula bent on torturing for pleasure, so much the worse for the idea that consistency alone can capture rationality. In *Thinking How to Live*, he writes:

> Crazy plans—if formally coherent and not rooted in defective naturalistic beliefs—will stem from crazy views of what, at base, is to be sought in life. A plan can be crazy though the planner's concepts are in good order; we need intact concepts, after all, to specify what's crazy in the plan. To my ear, such plans amount to vastly mistaken views of what, in itself, is good. (2003, 140)

Gibbard therefore seeks a theory of normativity that can make room for such strong intuitive normative commitments while preserving the insights of motivational judgment internalism—indeed, the two are indissoluably linked in his thought, since part of what it is to be serious about one's normative commitments is to take them as committing oneself to action—including action to oppose cruel or unjust plans. People often assume that, because he develops his theory of the meaning of normative judgments by centering on the idea that such judgments express motivating attitudes—rather than ordinary beliefs with ordinary truth-conditions—Gibbard must be committed to some form of "attitudinalism," "subjectivism," or "relativism." After all, isn't contemporary expressivism a form or descendant of "noncognitivism," and don't noncognitivists deny that moral or normative judgments have cognitive content, or could have truth-conditions, or could be objects of knowledge or objective inquiry in the manner of genuinely cognitive judgments? And aren't some philosophers

drawn to expressivism because of an underlying suspicion of objective morality? In contrast, I will try to suggest why expressivism has no special affinity for subjectivism or relativism, and can make a firm place for strongly anti-subjectivist and anti-relativist normative commitments.

Recently, Gibbard and other leading expressivists have drawn upon developments in the theory of truth to move away from the claim that normative judgments are incapable of truth or falsity. According to a *minimalist* approach to truth, the equivalence of *'p' is true* with *p* itself means that, once one has given an account of what it is to think that *The death penalty is unjust* (say), one has given an account of what it is to think that *'The death penalty is unjust' is true*. No substantive content or metaphysical assumptions are introduced by the "ascent" from *p* to *'p' is true*. If expressivism can give us a satisfactory account of *x is wrong* (e.g.), then, on the minimalist view, nothing more is needed for them to be able to say '*x is wrong*' *is true*. Similarly, for thoughts like *It's a fact that x is wrong*. Expressivists thus can make their peace with talk of truth and falsity in normative claims, and, relatedly, can allow that there are "moral facts" and "moral beliefs." Thus quasi-realist expressivists reject the label "noncognitivist."[1]

Critics have claimed that "quasi-realism" is a philosophically unstable point between two stable philosophical positions: one can either retain the long-standing contrast between judgments of fact and judgments of value, and admit that there is no more to quasi-realism than a mock-up of realism, or, efface that contrast, and lose the distinction between quasi-realism and realism. I propose to set these questions aside here, for two primary reasons. First, for many philosophers, the expressivists' introduction of quasi-realism does not remove their suspicion that there is, lurking somewhere in the heart of expressivism, a form of subjectivism—and a number of philosophers have attempted to show as much.[2] And second, what I hope to emphasize about the relation of expressivism to anti-subjectivism and anti-relativism will apply whether or not the expressivist opts for minimalism about truth and quasi-realism.

One source of the suspicion that expressivism is "in the end" a form of subjectivism could well be the fact that standard presentations of expressivism—for example, in the foundational writings by Ayer and Stevenson, in textbook recapitulations, and so forth—often explain and motivate the view by starting with subjectivism, noting the problems it faces, and then showing how noncognitivism or expressivism can fix them. However, if one digs a bit more deeply into these origins, one can also see something else going on, and this takes us to a core insight of expressivism. This "something else" gives less prominence to such issues as the difference between factual and normative judgments, and, perhaps surprisingly,

1 The conceptions of "cognitivism" and "noncognitivism," and what is central to each, are varied, ranging from the idea of a statement that is susceptible to truth or falsity to the idea of a distinctive faculty of the mind. For discussion of the various elements that have figured in philosophical conceptions of the "cognitive," see Bedke (2018).

2 For some recent attempts, and a reply on behalf of expressivism, see Schroeder (2014).

pulls into the foreground the problem of understanding *first-person normative deliberation*, which has been, as we will see, an apparent weakness for expressivism.

3. A Tendentious History: A. J. Ayer

Let's start with A. J. Ayer's *emotivism*, which can be thought of as a form of expressivism *avant la lettre*. This is because Ayer is careful to contrast his emotivism with traditional subjectivist theories precisely by emphasizing the difference between *descriptive* and *expressive* uses of language. Consider what might be the simplest version of subjectivism as an account of the meaning of moral terms:

(1) *x* is wrong $=_{\text{def.}}$ I disapprove of *x*.

Let us call (1) and views like it—for example, views in which "I" is replaced with "my group," "my God," and so forth—*descriptive semantic subjectivism*. This terminology might seem redundant except that part of our problem is that "subjectivism" has so many different uses and connotations. Who has held such a view? Hume is sometimes cited, having written:

> Take any action allow'd to be vicious. Examine it in all lights, and see if you can find that matter of fact, or real existence, which you call *vice* . . . The vice entirely escapes you, as long as you consider the object. You never can find it, till you turn your reflexion into your own breast, and find a sentiment of disapprobation, which arises in you toward this action . . . So that when you pronounce any action or character to be virtuous, you mean nothing, but that from the constitution of your nature you have a feeling or sentiment of blame from contemplation of it. (Hume 1888, 3.1.1; SBN 468–469)

However, it seems fair to say that Hume, in discussing moral judgment, was more focused on psychological and explanatory questions than on semantics. As David Wiggins has noted in his discussion of "A Sensible Subjectivism?" (1987, 187), Hume did not really come very near to an account of the *semantic* content of ethical claims—"even at the points where he speaks of 'defining' this or that." My sense is that Hume's informal use of "means" here is not intended in a strictly semantic way. Rather, it is closer to the ordinary language use of "means" when someone explains, "When he says that this worn-out toy has sentimental value, that doesn't mean that there is some feature, sentimental value, inherent in the toy, to which he is responding. Rather, it simply means that, given his personal history, he has strong feelings of attachment to it."

So Hume, I think, was not a descriptive semantic subjectivist in sense (1)—he certainly was clever enough to see the obvious problems of taking (1) as an account of the semantic content of '*x* is wrong'. Ayer summarizes these problems briefly before going on to his own, emotivist view. First, (1) would have the result that, if you say that the death penalty is wrong

and I say it is not wrong, we do not really disagree: you are reporting your attitude toward the death penalty, and I am reporting mine. Understood as reports, your remark and mine are entirely compatible. If we had both mastered the concept <*wrong*> as defined in (1), neither one of us would think the other mistaken—so if we wish to understand moral practices of agreement and disagreement, we won't get much help from (1).

On Ayer's view, saying "*x* is wrong" *evinces* rather than reports or describes one's disapproving attitude when one focuses on *x*, and such evincings do not contribute any content to the claim. Instead:

> . . . if I say to someone, "You acted wrongly in stealing that money," I am not stating anything more than if I had simply said, "You stole that money." In adding that this action is wrong . . . I am simply evincing my moral disapproval of it. It is as if I had said, "You stole that money," in a peculiar tone of horror, or written it with the addition of some special exclamation marks. The tone, or the exclamation marks, adds nothing to the literal meaning of the sentence. It merely serves to show that the expression of it is attended by certain feelings in the speaker. (1937, 107)

"Show" here is distinct from "tell" in just the way that evincing is distinct from describing. If you approve of stealing in a given instance while I disapprove of it, we have a common subject, the stealing, but have conflicting *attitudes* toward it—your *pro-attitude* pushes one way, my *con-attitude* pushes the other. Moral language allows us to *signal* such conflicting attitudes to one another.

However, this is not what Ayer views as the chief problem with orthodox subjectivism. He writes:

> . . . the main objection to the ordinary subjectivist theory is that the validity of ethical judgements is not determined by the nature of their author's feelings. (1937, 110)

This is no doubt a nontechnical use of "validity"—one might think of it as a matter of correctness or well-foundedness rather than logical validity. Such a remark on Ayer's part might sound like a nod to objectivism—certainly most objectivists would agree that whether someone's judgment that stealing is wrong is correct or well-founded is not simply determined by the individual's feelings on the matter. Yet this is not Ayer's explanation of what has gone wrong with (1), and why it points the way toward his emotivism. The "main objection" to subjectivism:

> . . . is an objection which our theory escapes. For it does not imply that the existence of any feelings is a necessary and sufficient condition of the validity of an ethical judgement. It implies, on the contrary, that ethical judgments have no validity. . . . (1937, 110)

Apart from their motive force, ethical judgments are "cognitively meaningless" according to the version of logical positivism Ayer was defending. According to Ayer, if a moral debate

persists once all disagreements in fact have been resolved, then all that is left is the expression of conflicting noncognitive attitudes toward the objects of moral assessment. Such differences in attitude are akin to matters of individual "taste," and no more subject to any objective determination than individual taste. This, he thinks, explains why moral debates can go on endlessly.

Thus Ayer, who very much wishes to distinguish his view from descriptive semantic subjectivism, is in effect upping the subjectivist ante: for him, descriptive semantic subjectivism is not "subjective" enough, since it would allow ethical judgments to be empirically verifiable, so long as one has psychological methods capable of using observational evidence to "probabilify" or "validate" (in Ayer's sense) claims about individual mental states such as approvals and disapprovals. By contrast, Ayer's emotivism pulls the rug from under talk of "validity" for ethical judgments, since they are "pseudo-propositions" lacking any content that could be probabilified by experience or proven by logic and concepts. And this, Ayer argues, is why fundamental ethical disagreements seem so irresolvable. Since ethical judgments on this account lack any possibility of verification or validation, and serve only to express the speaker's emotions, Ayer notes, his account "might fairly be said to be radically subjectivist" (1937, 109)—even in comparison with orthodox subjectivism.

But if fundamental ethical disagreements are akin to differences in individual taste, why do we treat them so differently? For example, in many matters of individual taste, we are content simply to "agree to disagree," and perhaps to seek out the company of those with whom we share tastes. Ayer's answer is that ethical language, while lacking any validity, nonetheless has a "practical justification" tied to the *shaping of others' conduct* in matters where we are not content to—or cannot—go our own separate ways, for example, in matters of family life, political governance, or social policy. In ethical disagreements, one uses the "emotive language" of "good" or "bad," "right" or "wrong," to try to "affect another person in such a way as to bring his sentiments on a given point into accordance with one's own" (1937, 22).

Why, though, if what underlies our ethical disagreements were akin to differences in taste, would my ethical judgments have any relevance for you if you do not happen to share my ethical "taste"? Unless I had some independent power over you, what pressure would expressions of my tastes put upon you to align your tastes with my own? And how is Ayer's "practical justification" supposed to be applied in first-personal ethical thought? When I wonder whether a given action would be wrong, am I simply trying to have a kind of rhetorical effect on myself? And why would such effects yield any conviction in me?

Ayer's emotivism, then, despite clearly rejecting (1) and insisting that moral judgments express rather than report our feelings, is nonetheless, as he puts it, "radically subjectivist." Here the notion of "subjectivist" in play is a broad, familiar sense in which "subjective" is contrasted with "objective." Consider such marks of *objectivity* in a domain of thought and practice as (1) the existence of conditions or standards of truth, accuracy, correctness, or "validity" for judgments in that domain, (2) that are not dependent upon any person's or group's *perspective*, *interests*, or *opinions*, and where (3) there exist methods of subjecting our judgments in

that domain to some form of *critical scrutiny* or *testing* that is appropriately linked to these truth-conditions or standards such that there are nonarbitrary ways for us to attempt to resolve disputes or gain knowledge. Ayer, then, is quite right in calling his account of moral language "radically subjectivist," since he denies all three of (1)–(3). Ethical judgments according to him are devoid of truth-evaluable or cognitive content, not subject to any standards of accuracy or "validity," and we have only pragmatic means for bringing about agreement in the face of fundamental ethical disputes. It is worries about "subjectivism" in this broad sense, I think, that have animated many of the philosophers who have criticized emotivism or expressivism for its "subjectivism." And at the same time, some philosophers with a "debunking" agenda in ethics have been drawn to emotivism or expressivism precisely because they see it as a good fit with their skepticism about the possibility of objectivity in ethics.

In short, emotivism and expressivism are clearly distinct from any orthodox metaethical subjectivism that takes a form akin to (1). But Ayerian emotivism is, by its author's admission, "radically subjectivist" in the broad sense that has been the chief concern of many critics of emotivism or expressivism. Once we fix upon the relevant sense of "subjective," then it isn't a simple failure to grasp the distinction between expressing and reporting one's attitudes (prevalent as that error may have been) that explains philosophical concerns about the "subjectivism" of Ayerian emotivism.

4. A Tendentious History: C. L. Stevenson

Ayerian emotivism may lead in the end to a "radical subjectivism" about ethical thought and practice, but is this essential for all emotivist accounts? What about C. L. Stevenson's more sophisticated version of emotivism? Stevenson takes as a starting point problems that arise for Ayerian emotivism: How to make meaningful ethical disagreement possible? How to explain the "magnetism" of moral judgments, or why we take ethical agreement or disagreement so seriously?

The development of Stevenson's proposed solutions starts off with a wide-angle view of language:

> Broadly speaking, there are two different *purposes* which lead us to use language. On the one hand, we use words (as in science) to record, clarify, and communicate *beliefs*. On the other hand we use words to give vent to our feelings (interjections), or to create moods (poetry), or to incite people to actions or attitudes (oratory). . . . [T]he distinction depends solely upon the *purpose* of the *speaker*. (1937, 75)

And what is a typical purpose in ethical discourse?

> When you tell a man that he oughtn't to steal, . . . [y]ou are attempting . . . to get *him* to disapprove of it. Your ethical judgment has a quasi-imperative force which, operating through

> suggestion, and intensified by your tone of voice, readily permits you to begin to *influence*, to *modify*, his interests. (1937, 74)

So far, we don't seem to have moved much beyond Ayer's "radical subjectivism." However, Stevenson goes on to claim that ethical debate typically involves more than bald quasi-imperatival utterances—we are expected to give others *reasons* to agree with us:

> When you point out to [the person who steals] the consequences of his actions—consequences which you suspect he already disapproves of—these *reasons* which support your ethical judgment are simply a means of facilitating your influence. If you think you can change his interests by making vivid to him how others will disapprove of him, you will do so; otherwise not. So the consideration about other people's interest is just an additional means you may employ, in order to move him, and not a part of the ethical judgment itself. . . . Thus ethical terms are *instruments* used in the complicated interplay and readjustment of human interests. (1937, 74–75)

However, this description makes it sound as if reasons are on the same footing as all of the other ways of attempting to influence others: using a sanctioning tone of voice or moralistic language, invoking the approval or disapproval of others, making vivid the consequences of actions, and so forth, are all at base mere instruments of influence, with no logical relationship to the ethical judgment itself. One simply selects whichever means of influence would be most effective in a given context.

This would seem to be a serious departure from our normal talk of "reasons," which distinguishes between *causal* or *motivating* reasons, on the one hand, and *justificatory* reasons, on the other. My boss might very much want me to submit false invoices, and demand this of me in an imperatival tone of voice; he might bring vividly to my attention the consequences of being fired and jobless if I do not comply; and these considerations might move me to reconsider my initial resistance to his command and alter my attitudes and behaviors to align them more closely with his demands. But effective as these considerations might be in altering my attitudes and conduct, would we say that he has given reasons for a judgment that submitting false invoices is the ethical thing to do in the circumstances? Stevenson writes:

> *Any* statement about *any* fact which *any* speaker considers likely to alter attitudes may be adduced as a reason for or against an ethical judgment. (1944, 114)

Stevenson sees his account of the meaning of ethical expressions as "reforming" so that he is not bothered if there are *some* elements of our intuitive notions they cannot capture. After all, there are probably incoherencies in our intuitive ethical notions. But he does want his "reforming definitions" to be "relevant" in the sense that, once people have clarified their

minds, they would be willing to accept these defined terms as clearer ways "to say all that they then want to say" in the domain (1937, 15). Yet distinguishing reasons that are genuinely justificatory from those that are simply casual or motivating *is* a chief concern of ethical deliberation, and of normative discourse in general. It would seem that, despite the invocation of reasons and reason-giving, we are back in the purely pragmatic territory of Ayer's account. Yet what if we look further into Stevenson's account?

Stevenson develops his reforming definitions for ethical expressions by starting with a (1)-like subjectivism and modifying it piecewise to address (1)'s problems. First, he suggests that ethical judgments are closer to "*We* like *x*," where "*we*" and "like" are being used "persuasively" by the speaker, rather than merely descriptively. His example is "a mother who says to her several children, 'One thing is certain, *we all like to be neat*.'" As Stevenson notes, "If she really believed this [in the descriptive sense], she wouldn't bother to say so." She instead is using her words to try to *make* it so. So, the judgment is more like:

> "This is good" has something like the meaning of "I *do* like this; do so as well." (1937, 78)

However,

> certainly this is not accurate. For the imperative makes appeal to the conscious efforts of the hearer. Of course he can't like something just by trying. He must be led to like it through suggestion. Hence an ethical sentence differs from an imperative in that it enables one to make changes in a much more subtle, less fully conscious way. (1937, 78)

Does this help with worries about the difference between *influencing* and *justifying*? If anything, it would seem to make ethical language a bit underhanded—operating "subtly" and less-than-fully consciously to gain influence over others one could not achieve with a bald imperative. He offers the following gloss:

> Strictly speaking, then, it is impossible to define "good" in terms of a favourable interest. Yet it is possible to say that "This is good" is *about* the favourable interest of the speaker and the hearer or hearers, and that it has a pleasing emotive meaning which fits the words for use in suggestion. This is a rough description of meaning, not a definition. But it serves the same clarifying function that a definition ordinarily does; and that, after all, is enough. (1937, 79)

Still, while Stevenson is careful to say that this way of identifying the moral "good" is "rough and inaccurate," he insists that, at bottom:

> It is disagreement in *interest* which takes place in ethics. When *C* says "This is good" and *D* says "No, it's bad," we have a case of suggestion and counter-suggestion. Each man is trying to redirect the other's interest. (1937, 79)

Once again, it would seem, we are being led by a founder of expressivism to a view of ethical language that rules out objectivity in senses (1)–(3), above, and renders ethical disagreement really a matter of conflicting interests, where there is no distinctively ethical cognitive content at stake—no question of ethical truth, accuracy, evidence, or knowledge. It would appear that "subjectivism" in the broad sense is still what's on offer.

However, Stevenson opens an avenue for moving beyond such subjectivism by attending more carefully to the kind of favorable attitudes expressed in ethical judgments. As Stevenson points out, the mother who says "One thing is certain, *we all like to be neat*" is using "like" as well as "we" in a prescriptive rather than descriptive sense—a sense closer to what we *approve* of rather than simply prefer. He writes:

> When a person *likes* something, he is pleased when it prospers, and disappointed when it doesn't. When a person *morally approves* of something, he experiences a rich feeling of security when it prospers, and is indignant or "shocked" when it doesn't. (1937, 79)

So, I would fail to understand something important about ethical thought and discourse if I resorted to moral language simply for the purpose subtly influencing others to get their preferences to align with mine. To make sincere ethical judgments, I must not simply *like* the action in question, I must *approve* of it in a way that is important for my sense of security in the world, and that would leave me shocked, indignant, and insecure if others were not to share my approval.

Note that this point of Stevenson's is more fine-grained than generic motivational internalism about ethical judgments—the state of mind expressed in ethical judgments must not only have motivating force, but it must have *the right kind* of motivational force. Personal preferences or tastes, for example, have motivating force. Yet, unless I *moralize* my preferences and tastes, I do not resort to ethical language to express such motivating attitudes. When I sincerely moralize my attitudes, I am *invested* in them emotionally in distinctive ways, and concerned to *advocate* them in ways that could evoke in others a similar kind of emotional investment. This very much narrows the kinds of considerations I can appeal to. For example, the considerations my boss used to influence me to submit false invoices are such that I may be willing to go along, and to feel fear about failing to do so, but they won't evoke in me a rich security in a world where such compliance is required and "prospers." Neither will I feel shocked indignation if the boss changes his instructions to drop the demand to submit false invoices. This is still to identify relevance solely in terms of potential *influence*, but finding considerations that genuinely engage feelings of security, indignation, moralized concern, and so forth, is a substantial constraint on the kinds of considerations that have ethical relevance—especially since, as Stevenson notes, these feelings are not subject to direct manipulation by the will. Contemporary descendants of emotivism have explored further down this avenue opened by Stevenson, drawing upon the fact that distinctive attitudes owe their individual character to the kinds of *information* to which they

are responsive, the ways in which they *represent* the world, and the ways in which they *guide* thought and action. Here, then, is a path for cognitive content to play an essential role in ethical discourse and disagreement.

Moreover, here is a path to enable the emotivist or expressivist to give a more plausible account of first-personal deliberation. In his intriguing 1950 paper, "The Emotive Conception of Ethics and Its Cognitive Implications," Stevenson argues that emotivism is especially well-placed to explain the seriousness with which we take the cognitive dimensions in moral inquiry, deliberation, and judgment—indeed, better placed to do so than many self-described "cognitivist" approaches in metaethics.

To show us why, Stevenson asks us to begin, not as emotivists typically have, with the standpoint of someone whose ethical opinions are settled and whose primary purpose in using ethical language is to influence others to share his own attitudes, but rather with the standpoint of someone who needs to make a "personal decision"—whose mind is at present divided on some question of real practical significance for himself and others, and who needs to make up his mind and act.

> My conception of a personal decision will not be new: I shall borrow most of it from John Dewey. . . . My hope is simply to see this old conception in a new relationship. Some of you may feel that an emotive analysis of ethics, of the sort I shall later defend, is too simple—that it must be insensitive, in particular, to the role of cognition in ethics. Now I think that is far from the case. So I shall take a conception of a personal decision which, by common consent, has cognitive elements that are highly complex; and I shall then endeavor to show that an emotive analysis, so far from ignoring them, is actually of interest in throwing them into greater relief. (1950, 291–92)

There is a risk, Stevenson notes, of emotivism seeming "trivial" in first-personal ethical deliberation, little more than a matter of "self-exhortation" (1950, 299) rather than serious examination of the ethical question posed. He adds,

> Although self-exhortation is interesting enough, it is scarcely a matter to be dwelt upon. (1950, 299)

Yet ethical quandaries are precisely the kind of thing we do dwell upon. So, what *is* someone doing when she engages in such deliberation, if the Stevensonian emotivist is to be believed?

> A part of the answer is this: he is trying to make up his mind whether to approve or disapprove of something . . . So long as he is ethically undecided, his attitudes are in a psychological state of *conflict*; half of him approves of a certain object or action, and the other half of him disapproves of it. And only when he has resolved his conflict, making his attitudes speak with one voice, will he have made his decision. (1950, 292)

If the two halves of my mind are evenly divided, there is little hope that introducing moral language to raise the rhetorical temperature will help, since it would help both sides equally. So what could help? In such cases, Stevenson believes, the agent is driven to look *behind* or *beyond* these conflicting attitudes:

> When a man has conflicting attitudes, he is virtually forced to think—to recall to mind whatever he knows about the alternatives before him, and to learn as much more about them as he can. (1950, 292)

This process of examination of his alternatives is plainly cognitive, a process of "establishing, cognitively, the varied beliefs that may *help* him resolve" the conflict. However, on Stevenson's view, the relation of these beliefs to his attitudes cannot be *logical*, since attitudes do not enter into logical relations in their own right. The relation instead will be "causal," taking the form of "influence" rather than implication (1950, 302).

That might seem to land us back where we started—trying to conceive first-personal deliberation and interpersonal moral discourse simply on a model of using language as a source of *influence*, not reasoning. The fact that the bearing of one's beliefs on one's attitudes is not a matter of reasoning may seem to the cognitivist to be a way of saying that this bearing isn't sufficiently "serious" or "thoughtful" or "rational." But the fact that a chain of mental inference follows logical relations doesn't tell us whether it is "serious," "thoughtful," or even "rational"—if we start with absurd premises, following logic may simply lead us to absurd conclusions. After all, orthodox subjectivism of form (1), above, interprets one's moral judgments as cognitive claims susceptible to evidence and reasoning, but such views conspicuously fail to capture what goes on in moral thought. As Stevenson observes, someone who sought to answer the question "Would this action be wrong?" by replacing it with "Do I disapprove of it?" *wouldn't* be making up his mind about the ethical quandary; rather, he would be making up his mind about what mental state he is in. And, having done that, he would still need to answer the question, *What is to be done?* It could be perfectly accurate and rationally justified to believe that one's mental state is divided about a given personal decision, but this in itself would generate no pressure toward reaching a more univocal mental state. The flaw of orthodox subjectivism in the interpersonal case applies equally intrapersonally, once we have a divided mind.

Instead, Stevenson argues, a "serious," "thoughtful," and "rational" person faced with his own conflicting attitudes on an important practical question calling for action is not "*looking at* his conflict," but "*living through* it"—being torn in a way that *does* create pressure to become "of one mind." "The man is trying to *resolve* a conflict" over what to do, and resolution will not occur unless he succeeds in "actually mak[ing a] change in his attitudes." So causal influence *is* part of the point, after all. The quandary won't be resolved unless an actual change in action-guiding attitudes occurs. Thus:

> the man's efforts, throughout his decision, are to change his very attitudes. (1950, 300)

The question becomes: How might such self-influence occur in such a way as to count as responsiveness to reasons? Stevenson gives the example that "our approval of anything is strengthened or weakened depending upon whether we approve or disapprove of its consequences" (1950, 293). So, by looking carefully at the likely consequences of an act, I can influence how I regard it.

This kind of influence-seeking, Stevenson writes, is a form of "'practical reason' in the only sense of that term that seems to me intelligible: . . . ordinary reasoning made practical by its psychological context," that is, by how the reasoning will affect me, strengthening or weakening my attitudes and, therefore, my dispositions to act. For the agent:

> reasoning serves . . . purely as an *intermediary* between his attitudes: by connecting his thought of [an act, say] with his thought of [a likely consequence of the act], it also connects his attitude toward [the act] with his attitude toward [the likely consequences], letting the one be reinforced by the other. (1950, 293)

This might sound passive—you are "letting" thoughts influence your attitudes rather than *deciding* what that influence is to be (compare Korsgaard 2008). But that objection just pushes the problem up a level, since *deciding* what the influence is to be would be yet another case of trying to "make up one's mind," and so to have an effect upon one's attitudes. Importantly, we *cannot* just decide what our attitudes will be, or change them by fiat—one must instead identify considerations that really *can* influence one's attitudes, and given the attitudes they are this will not be arbitrary. Just as I cannot simply decide what to believe, and must instead give myself evidence or arguments, I cannot simply decide to be indignant, for example, when I see a boss abusing his power—I must instead bring to mind the reasons why an abuse of power warrants indignation. Reasoning that remains strictly cognitive, that issues not *in* attitude change and action, but in judgments *about* attitude change and action, is not yet practical, and so insufficiently serious, hard-minded, or thoughtful to resolve our personal decisions.

This is how practical reasoning can be no mere idle exercise. In a way reminiscent of Aristotle, practical reasoning for Stevenson concludes, not in an *ought* judgment but in "the beginning of action," passing through deliberation as an intermediary that makes one aware of what is really involved when one translates ends into action (*De Anima*, 433a). An agent equipped with practical reason is someone capable of this kind of deliberative self-influence upon motivation and action.

Stevenson asks us to consider a monk, who has run through arguments from religious principles and concluded that he *ought* to be chaste—it is his "moral duty," as he would put it. Has this person really reached a practical conclusion?

> Suppose that his ordinary preferences constantly outweigh his peculiarly moral attitudes, leading him along a path that is not so straight and not so narrow. I suspect, in that case, that we shall be less interested in his code of morality than in his code of preference. (1950, 295)

There is a flaw in the seriousness, thoughtfulness, and rationality of this man's thinking, but it is not a logical flaw in his reasoning—it is a failure of his reasoning to influence his attitudes. Stevenson writes, if ethics "is to be 'practical' philosophy, and not a mockery of what is practical," then reaching an ethical conclusion cannot leave the agent so little moved to take the path he alleges to be his duty. Why study ethics, Stevenson asks, rather than some other "innocuous subject like the stamp issues of Andorra" (1950, 304), if you will rest content with forming conclusions that make so little mark on your antecedent preferences and dispositions to act?

Now some contemporary internalists will argue that it is *part of* practical rationality that *if* one arrives at an *ought* judgment concerning action one will be, other things equal, motivated to follow suit[3] (for discussion, see MacPherson and Faraci 2018). Stevenson would note, however, that redescribes what he himself has been saying—an exercise of rationality won't be practical unless it can actually influence one's attitudes and behavior. And this cannot be a purely logical enterprise, since attitudes aren't entailed by propositional arguments. That is, practical rationality is not simply the ability to construct a practical syllogism—it requires the ability to causally *influence one's attitudes* through thinking about what to do. Or, as Gibbard would later put it, "thinking how to live."

But doesn't Stevenson also say that *any* consideration that could influence action-guiding attitudes *could* be a reason for deciding? Isn't it then an arbitrary matter after all? On the contrary. This is a way of recognizing that we could always be wrong, perhaps deeply so, about what is relevant and why. Stevenson asks us to consider someone who has decided on the basis of *meaning* to take certain considerations as determinative in practical thought. For example, imagine someone who thinks the very meaning of <*x* is valuable> establishes that:

(2) *x* is valuable $=_{\text{def.}}$ *x* is conducive to *E*

where *E* is some definite end, perhaps "social survival, . . . the social integration of interests, or the greatest happiness of the greatest number, or the maximal presence of a unique, indefinable quality, or any other impersonal aim", or even "*E* is what I or we would settle upon after careful deliberation" (1950, 297). The trouble with such definitions, he argues, is not that they fail to invoke considerations that are relevant to ethical decision-making—as far as he can see, such considerations are very relevant. Rather, the trouble is that such definitions "lead us to suppose that the effect of *x* on [*E*] is *all* that we have to consider" and deprive us of language for inquiring meaningfully about whether *E* itself is really or uniquely valuable.

What if, Stevenson continues, this individual discovers that the society in which the greatest happiness is enjoyed by the greatest number is like Huxley's *Brave New World*? At this point, a serious, thoughtful, rational person would wonder whether something has gone wrong. But what? One might say, the price of happiness turns out to be too high. But this

3 For discussion, see McPherson and Faraci (2018).

shows that one could all along understand "value" in a way that is not tied to the particular content picked out by (2). Stevenson writes:

> And should anyone argue that the price was irrelevant, being foreign to the [definition of "valuable" given], I think we should answer: "So much the worse for the definition." (1950, 297)

If this individual faces such a situation in an open, objective frame of mind, rather than a frame of mind limited by his prior convictions, he will at this point experience genuine conflicts in attitude, and be "virtually forced to think" about what he needs to know to try to resolve it if he is to go forward in conducting his life.

We need a vocabulary of terms like "good" and "right" and "rational" that allows for this kind of open-ended examination and inquiry, tied not to specific content but to a connection to how we will go on to live. Definitions akin to (2) will not give us such a vocabulary, for they will simply settle such disputes *a priori* one way or another. This is another way of interpreting Moore's argument that we should understand the open-question argument as a bulwark against someone attempting to "foist" his ethics upon us as a matter of definition (Moore 1903, 7).

Stevenson argued that the emotive theory *does* give an account of ethical terms that would enable us to resist such foisting and to escape the open-question argument, not because of any metaphysical indeterminacy about what is good but because, when we seriously, thoughtfully, and rationally face ethical questions, we aren't engaged in metaphysics. Instead, we are, in earnest, "trying to make up our minds about whether to approve or disapprove," in a way that is not idle or "unsanctioned," but rather that will shape the course of what we do. And that is how Stevenson understands the way meaning works in our fundamental ethical vocabulary, and how this approach to meaning enables us to understand first-personal ethical deliberation and decision.

6. A Tendentious History: Allan Gibbard

This why there is no real tension between expressivism and a commitment to objectivity in our thinking about value or ethics. One manifestation of a commitment to objectivity in a domain of inquiry is precisely the sense that any account or criterion we develop could still be inadequate to capture all that is at stake. We cannot take our current point of view or commitments as guaranteed to be correct, but must be remain open to the idea that we could wrong even in our most central convictions or assumptions. Recall that among the marks of objectivity in a given domain is that correctness or accuracy in that domain is independent of our opinion, and not beholden to our particular point of view or interests. However, if we must treat ethical inquiry as forever open-ended in this way, we are nonetheless not without guidance—we have the vast history of humanity to draw upon when thinking about how to live, and the potential for trying new ways of living. But we are, or should be, prepared to

be wrong, to be taught something new, to find that have been insufficiently sensitive or reflective. Unlike approaches to ethics that insist upon a particular rational or empirical method, we need to be able to stand back from those methods and raise the question of their appropriateness. We cannot turn this kind of decision-making over to metaphysics, facts, or unquestionable principles.

Throughout the time I have known him, which now is, I blush to admit, *quite* a long time, I have admired in Allan Gibbard his appreciation of this Stevensonian point. Indeed, he has elaborated it in systematic and powerful ways that seemed impossible at the time when Stevenson wrote. Like Stevenson, Gibbard wants to see normative decision-making as a natural phenomenon—an aspect of the human condition and vital to the well-functioning or our lives and societies. But also like Stevenson, Gibbard does not want normative decision-making to be supplanted by such a descriptive enterprise. No one is clearer than Gibbard about the costs of various departures from, say, decision-theoretic or utilitarian principles in personal or social decision-making. Yet he has steadfastly refused to overlook the unintuitive implications of following these theories as guides to life. He wrote *Wise Choices, Apt Feelings* from the standpoint of someone who genuinely believes there are wise (and unwise) choices and apt (and inapt) feelings. His is therefore not an expressivism founded upon the disparagement of that which isn't a matter of belief, logic, and truth. Desires and other kinds of feelings or affective states might not be strictly speaking subject to truth or falsity, but they can certainly go wrong. "Desires ought to reflect what is worth wanting in life; otherwise they are misdirected" (2008, 170). Gibbard wants to show that there is a logic to desires, preferences, and plans, which makes it possible to think sensibly *with* and *through* desires so that the upshot will be genuinely practical. But this logic runs out before worthwhileness is established. "There are many things it would be logically coherent to desire but crazy," he writes (2008, 170). If this must be a matter of "intuition," then Gibbard is prepared to make his peace with intuition, but "cautiously and critically" (2008, 180). Once I ran into him after just writing up a paper meant to refute Rawls' "original position" argument against utilitarianism, a paper with which I was rather pleased. "This should help alleviate people's anti-utilitarian intuitions," I said. I didn't expect the answer I recall receiving. "Well, at this point I'm more interested in understanding my anti-utilitarian intuitions than in pushing them aside."

Gibbard has several times mentioned to me that his philosophical thinking was deeply shaped by an early reading of Sartre's *Existentialism Is a Humanism* (1947/1996). Sartre and Gibbard might seem like an odd couple, but they do share this: to understand what it is to be human we must understand what it is to decide and act, and the logical or conceptual independence of this from any "facticity" or "definition" to which we might attempt to resign our capacity to deliberate, decide, and act. As Sartre would be the first to say, choice, unnerving as this can be, is something we cannot evade.

References

Aristotle (2016). *De Anima*, translated by C. Shields. Oxford: Clarendon.

Ayer, A. J. (1937). *Language, Truth, and Logic*, 2nd ed., 1946. London: Victor Gollantz.

Bedke, M. (2018). "Cognitivism and Non-Cognitivism." In T. M. McPherson and D. Plunkett (eds.), *The Routledge Handbook of Metaethics*, 292–307. New York: Routledge.

Korsgaard, C. M. (2008): "Acting for a Reason." In C. M. Korsgaard (ed.), *The Constitution of Agency: Essays on Practical Reason and Moral Psychology*, 207-229. Oxford: Oxford University Press.

Darwall, S. L, A. Gibbard, and P. Railton (1992). "Toward *Fin de Siècle* Ethics: Some Trends." *Philosophical Review* 101: 115–89.

Gibbard, Allan (1990). *Wise Choices, Apt Feelings*. Cambridge, MA: Harvard University Press.

——— (2003). *Thinking How to Live*. Cambridge, MA: Harvard University Press.

——— (2008). *Reconciling Our Aims*. New York: Oxford University Press.

——— (2012). *Meaning and Normativity*. New York: Oxford University Press.

Hume, David (1888). *Treatise of Human Nature*, edited by L. A. Selby-Bigge. Oxford: Oxford University Press.

McPherson, T., and D. Faraci, D. (2018). "Ethical Judgment and Motivation." In T. McPherson and D. Plunkett (eds.), *The Routledge Handbook of Metaethics*, 308–23. New York: Routledge.

Moore, G. E. (1903). *Principia Ethica*. Oxford: Oxford University Press.

Sartre, Jean-Paul (1947/1996). *L'Existentialisme est un humanisme*. Paris: Gallimard.

Schroeder, M. (2014). "Does Expressivism Have Subjectivist Consequences?" *Philosophical Perspectives* 28: 278–90.

Stevenson, C. L. (1937). "The Emotive Meaning of Ethical Terms." *Mind* 46: 14–31.

——— (1944). *Ethics and Language*. New Haven, CT: Yale University Press.

——— (1950). "The Emotive Conception of Ethics and Its Cognitive Implications." *The Philosophical Review* 59: 291–304.

Wiggins, D. (1987). "A Sensible Subjectivism?" In D. Wiggins (ed.). *Needs, Values, Truth: Essays in the Philosophy of Value*, 185–214. Oxford: Basil Blackwell.

16

THE METAPHYSICAL CONCEPTION OF REALISM[1]

Billy Dunaway

1. Preliminaries

Whether we should be "realists" about a particular subject matter is one of the central questions of philosophy: the issue arises in debates about morality, numbers, material objects, and consciousness, just to name a few. There is, however, a prior question, which is the question of what the realist view *is*. Perhaps one of the most serious challenges to any attempt at an answer to this question comes from Allan Gibbard's development of expressivism about normativity in Gibbard (2003), and his subsequent extension of the view to expressivism about meaning in Gibbard (2013). Gibbard rigorously develops a "quasi-realist" view of these domains, which captures many of the claims traditionally associated with realism while appealing only to the distinctive explanatory resources of expressivism. The expressivist, Gibbard shows, can allow that there are normative truths, facts, beliefs, and properties (among other things), but explains this with a purely naturalistic story about what state of mind is involved in accepting such claims.[2]

As he notes at the outset in Gibbard (2003), this raises questions for traditional taxonomies in metaethics. The realist view about normativity is supposed to be a view about its metaphysical status, but the expressivist seems to agree with all of the metaphysical claims

1 Thanks to Allan Gibbard, Ezra Keshet, David Manley, Eliot Michaelson, David Plunkett, Peter Railton, Timothy Rosenkoetter, Kenny Walden, and participants in the "Realism, Objectivity, and Meta-metaphysics" seminar at Dartmouth College for helpful comments and discussion. This paper was written with support of the John Templeton Foundation.

2 See also Blackburn (1993) on quasi-realism and Dreier (2004) for discussion.

the realist might make. As Gibbard says at the outset in Gibbard (2003), "In many ways, I'll end up sounding like a non-naturalist, and in some ways, like certain kinds of naturalists. Am I, then, really a descriptivist in disguise, a moral realist?."[3] He then asks, "How does my position fall short of full ethical realism?"[4] but declines to give a definitive answer. Instead, he goes on to note how his expressivist view of normative judgment merely makes additional claims over what the realist claims—for instance, that normative concepts are planning concepts—but goes out of his way to avoid asserting that this amounts to a *denial* for realism. Instead, Gibbard simply points out that this "would contrast with a standard realist's mode of explanation."[5]

In this paper, I will suggest an answer to Gibbard's question. My primary aim is to characterize realism in a way that does not leave it susceptible to quasi-realist accommodation. The strategy is to argue that there is a genuine metaphysical component of realism, beyond what the quasi-realist accepts. In short, my thesis is that the properties that play the realist's explanatory role are, to some extent, *fundamental* in the metaphysician's sense. This notion of fundamentality is, I claim, the central component of realism.

To argue for this thesis, I will not start with the hard case, namely the contrast between realism and Gibbardian quasi-realism. Instead, I will argue for the thesis on the basis of straightforward examples. For instance, most will not hesitate to label Berkeley's idealism an irrealist view of material objects, to be contrasted with our ordinary, pre-theoretic realist view. Or again, the realist about the unobservable posits of our scientific theories is easily distinguished from the irrealist Instrumentalist. And in ethics, most will not hesitate to label subjectivists as irrealists, distinguishing them from the realist Moorean nonnaturalist.[6] It is examples like these, I argue, which suggest that a substantial metaphysical notion such as fundamentality is central to realism.

If this argument is successful, we will have a metaphysical characterization of what realism amounts to. Showing that Gibbard, or other expressivists, cannot adopt the quasi-realist program to accommodate this claim is a further task. I will not attempt it here but will gesture at how the issue should be viewed in the conclusion.[7]

Before proceeding, a few caveats are in order.

---

3 Gibbard (2003, 18).

4 Gibbard (2003, 19).

5 Gibbard (2003, 187).

6 To fix ideas, I understand these versions of irrealism in roughly the following way. The idealist, following Berkeley (1710), holds that material objects are nothing more than collections of ideas. Instrumentalists such as Duhem (1954) hold that scientific statements are shorthand for statements about the actual or counterfactual properties of instruments of measurement. A Subjectivist, as I will use the term, holds that ethical statements report on some agent's attitudes of approval and disapproval. One example is the Emotivist view in Stevenson (1937), but the more canonical version of Subjectivism I will have in mind throughout hold roughly that 'it is wrong for *S* to Φ' is true just in case *S* disapproves of Φing.

7 For a more detailed treatment, see Dunaway (2016).

First, we need to be careful to distinguish the present project from that of performing conceptual analysis on a term of art. 'Realism' in the relevant sense belongs almost exclusively to the lexicon of philosophers, and bears no straightforward connection to our pretheoretic vocabulary. All we have to guide our use of the term is our own dispositions and intuitions, which we learned as we became fluent with a term by imitating the usage of the term of art by others in the philosophical community. An account of realism that simply describes community-wide usage is nothing more than a redescription of the linguistic habits of a relatively small community. That might be of some sociological interest but is of no direct philosophical significance.

The project here is importantly different in a number of respects. There is a possibility of failure—there might not *be* a theoretically interesting natural kind that plausibly counts as the referent of 'realism'.[8] We aren't prejudging the question of whether we will be successful in finding an analysis by embarking on the investigation. An attempt at an analysis in terms of a joint-cutting natural kind, moreover, allows for some divergence from our intuitive judgments. If, in our final analysis, there is a highly natural kind that provides a close-but-not-perfect fit with our intuitive use of 'realism', we might well say that the views properly called "realist" differ from those we initially applied the term to.

The second caveat is that, in what follows, I will be assuming that a view is realist (or not) primarily in virtue of its *metaphysical* consequences. This is a natural idea—Berkeley's idealism seems irrealist precisely because of its consequences about the nature of material objects since they are, according to Berkeley, merely collections of ideas. Similarly for realism about scientific unobservables: part of what makes the realist view objectionable to some is that it takes a stance on the metaphysics of unobservables, which makes them unknowable. Moorean nonnaturalism about ethics is often claimed to be metaphysically extravagant. So, it is quite natural to take one interesting project to be one of asking which metaphysical consequences of a view are necessary and sufficient for realism.[9]

---

8 Crispin Wright (1987, 3–4) indicates sympathy with this pessimistic conclusion: "The fact is that realism, as implicitly characterized by the opinions of writers, in whatever area of philosophy, who regard themselves as realists, is a syndrome, a loose weave of separable presuppositions and attitudes."

9 Lewis (1984) echoes this idea when he objects to Putnam's own characterization of the model-theoretic argument in Putnam (1981, Ch. 2) as suggesting the denial of the core thesis of realism. Lewis explicitly says this is because it leaves the metaphysics of traditional realism unscathed:

> Even if the model-theoretic argument worked, it would not blow away the whole of the realist's picture of the world and its relation to theory . . . There would still be a world, and it would not be a figment of our imagination. It would still have many parts, and these parts would fall into classes and relations . . . There would still be interpretations, assignments of reference, intended and otherwise. (Lewis 1984, 231)

See also Miller (2003). For a dissenter from this idea, see Dummett (1977, 383), and other citations in Miller (2003, 196).

Such an understanding of realism is not the only one available. Many claim to find additional, nonmetaphysical aspects to realism. Some are *epistemic*: Boyd (1988, 181–82) takes realism to imply that our cognitive faculties afford us a means of "obtaining and improving" knowledge in the relevant domain. Dummett (1982, 55), on the other hand, claims to find in realism a distinctive commitment to the truth of claims "independently of whether we know, or are even able to discover" their truth.[10] Other characterizations of realism involve *semantic* properties like truth, literalness, and so forth.[11]

I will proceed in what follows by ignoring these nonmetaphysical dimensions to realism. The most straightforward motivation for this is methodological—insofar as it is clear there is some metaphysical component to realism, a characterization that posits additional epistemic or semantic dimensions to realism will thereby be less natural and more gerrymandered. A search for the natural kind underlying talk of 'realism' then does best by beginning with a purely metaphysical characterization; other dimensions should be added only if a purely metaphysical conception of realism is unavailable.

The final caveat is that there may well be specific domains in which use of 'realism' has spun off from its general philosophical use, and in these contexts has a specialized meaning. So we should not be that surprised if, for instance, 'legal realism' turns out to denote a kind that has little to do with the general philosophical sense of 'realism'. Of course, this is not an argument that it is impossible to assimilate the subject of these specialized uses to a core property of realism that applies across domains. The present point is just that we should be prepared to admit the existence of such specialized uses, and that this would not amount conceding defeat in the project of giving a purely metaphysical characterization of realism.[12]

The aim of the present paper, then, is to find a metaphysical characterization of the natural kind picked out by our general use of 'realism' (if any such kind exists). §2 outlines three popular accounts of realism in the literature: these are the *Existence View*, the *Mind-Independence View*, and the *Fundamentality View*. §3 argues that there are structural features of realism that cannot be accommodated by these views. These objections together are, I think, decisive against the Existence and Mind-Independence views. The situation with

10 Strictly speaking, Boyd's and Dummett's claims are consistent: we might have good cognitive resources for arriving at knowledge of most claims in a domain, while *some* of its claims are nonetheless in principle unknowable. Nevertheless, a tempting diagnosis of these divergent emphases is that Boyd and Dummett are latching onto merely accidental features of different realist views.

11 See Miller (2003) for a discussion of the relationship between semantic and metaphysical characterizations of realism.

12 Wright points to the same phenomenon but is not very optimistic about the prospect for separating specialized uses of 'realism' from its common core. He says:

> Of course, if there ever was a consensus of understanding about "realism," as a philosophical term of art, it has undoubtedly been fragmented by the pressures exerted by the various debates—so much so that a philosopher who asserts that she is a realist about theoretical science, for example, or ethics, has probably for most philosophical audiences, accomplished little more than to clear her throat. (Wright 1992, 1)

respect to the Fundamentality View is different: these objections only cause trouble for views which take give an account of realism in terms of *absolute* fundamentality, which is the standard approach in the literature. There is, however, another approach in the neighborhood, which instead characterizes realism in terms of *degrees* of fundamentality. §4 sketches how an account along these lines yields a promising account of the structural features of realism outlined below. I conclude by showing how the Existence and Mind-Independence accounts, though inadequate, are in many cases good heuristics for settling questions of realism. The quasi-realist explanatory strategy appears to preclude the kind of explanatory role for normative properties, which would make them highly fundamental.

2. Three Conceptions of Realism

There are three main metaphysical conceptions of realism: Existence, Mind-Independence, and Fundamentality views. In this section, I introduce and elaborate on each.

2.1. Existence Views

Existence views hold that a theory is realist just in case it entails that entities of an appropriate kind exist. What *kind* of entity is required is variable: some versions hold that realist theories entail that *properties* of the relevant kind exist; other versions hold the same for the relevant kind of *facts*. Existence views are prominent in the literature on ethical realism.[13]

One instance is found in J. L. Mackie (1977), where he intends his metaethical view, which he calls "moral skepticism," to be the denial of ethical realism. He characterizes this view in the following way:

> What I have called moral skepticism is a negative doctrine, not a positive one: it says what there isn't, not what there is. It says that there do not exist entities or relations of a certain kind, objective values or requirements, which many people have believed to exist.[14]

Mackie's "negative thesis" is a denial of an existence claim—namely, the claim that certain "entities or relations" exist and is supposedly in conflict with standard realist conceptions of ethics on this count. This presupposes that the realist view entails the existence of certain things: "values," as Mackie says.

---

13 See also one disjunct of the definitions in Cameron (2008), Devitt (1991), and Miller (2003). Pettit (1991) is more coy: "Realism in any area of thought is the doctrine certain that entities allegedly associated with that area are indeed real" (Pettit 1991, 588). He is explicit that *one* way of rejecting this thesis is to deny existence to the relevant entities (589–90). But he subsequently discusses other ways to deny realism, which suggests that he would not consider a bare existence claim to be adequate to characterize realism.

14 Mackie (1977, 17).

Another case is found in Shafer-Landau (2003), which offers a broad taxonomy of metaethical positions. The first position is the eliminativist view, which "is represented by error theorists and non-cognitivists. Such philosophers do not believe that there are any moral properties, and believe that all appearances to the contrary are either founded on error, or can be otherwise explained away" (66). The other options are reductionism, which holds that "moral properties, if they are to exist, must be (in the sense of be identical to) one of these kinds of natural property" (66–67), and nonnaturalism, which rejects "the identity of moral and descriptive properties" (72).

On a standard classification, only the last two views—the reductionist and nonnaturalist views—are the views that are consistent with realism. The eliminativist view, represented by error theorists and noncognitivists, is not. What separates these realist views from others in Shafer-Landau's taxonomy is that they entail the existence of moral properties. This strongly suggests that Shafer-Landau takes the existence of these properties to be the key ingredient for realist views about ethics.[15]

These existence-based conceptions of realism about ethics can be thought of as generalizations on a standard characterization of realism about unobservables in scientific theories. In van Fraassen (1980), the characteristic claim of realism is that *there are* electrons and other unobservables posited by scientific theories. This makes sense in the context of realism about unobservables: the primary motivation of the irrealist is to avoid what she believes to be an unwarranted ontological commitment to an unobservable world of electrons, and the way to avoid this commitment is to decline to believe that they exist.[16] Existence views of realism in other areas are then natural extensions of this idea. In some cases, there are no prosaic entities that are the subject matter of a theoretical enterprise. In these cases, the Existence View makes realism a question of whether the relevant *properties* exist.[17]

---

15 Elsewhere, he says that what is definitive of realist views is that they entail the existence of moral *facts* (see, for instance, Shafer-Landau 2003, 15). Shafer-Landau may either be undecided between one of two Existence views, or may think that they amount to the same thing. The latter view would make sense if one thought that facts are structured set-theoretic entities with properties (among other things) for constituents. Then, the failure of ethical properties to exist would by itself give rise to a lack of existence in ethical facts. I won't work with these distinctions in the main text, since it will not matter for much of what I say whether the Existence View is primarily concerned with properties, facts, or similar entities.

16 In the case of van Fraassen's irrealist alternative, the irrealist doesn't take on the contrary commitment by denying that there are electrons. Rather, she withholds belief and (in van Fraassen's terms) merely *accepts*, rather than *believes*, scientific theories for the purposes of carrying out scientific investigation. I take this to be an instance of the Existence View of realism, even though van Fraassen nonetheless recommends *acceptance* of an existence claim. This is because the difference between acceptance and belief concerns whether bearing the relevant attitude to the claim that electrons exist brings along an ontological commitment to the existence of electrons. van Fraassen recommends mere acceptance over belief precisely because it does not bring about this kind of commitment.

17 We might worry about this motivation: not all metaphysical commitments are *ontological* commitments in the sense that they are commitments concerning which objects or entities exist. Theories can contain unwanted metaphysical commitments by including an unnecessarily complex primitive *ideology* as

2.2. Mind-Independence Views

Another common way to characterize realism about a domain is to claim that all realist views hold the domain to be *independent* of the mental. Examples of mind-dependence (and accompanying irrealism) are familiar from the history of philosophy: think of Berkeley's claim that ordinary objects are collections of ideas, or a version of the Humean view of causation on which it consists in nothing more than constant conjunction plus expectation on the part of observers.[18] This thought seems especially apt when considering irrealism the ethical domain, as many paradigmatic instances of irrealist ethical theories enlist mental states of approval, disapproval, and the like, to play important explanatory roles.

One way to articulate this approach is found in Sharon Street (2006). She says:

> The defining claim of realism about value, as I will be understanding it, is that there are at least some evaluative facts or truths that hold independently of all our evaluative attitudes.[19]

For Street, then, metaethical theories are realist just in case they entail that ethical facts are independent of certain attitudes. (Admittedly, Street suggests her characterization should be taken as "stipulative." But it wouldn't be a natural stipulation if there weren't some plausibility to the claim that realism—in a nonstipulative sense—requires Mind-Independence. It is this latter claim, not any stipulated definition, that will be the focus of the following discussion.) Similarly, Brink (1984, 111) characterizes moral realism as a view that entails that the moral truths are independent from "those beliefs which are our evidence" for them.[20]

A Mind-Independence View can be extended to other domains in various ways.[21] The basic idea is that just as facts about value are mind-dependent if they depend on our evaluative attitudes, so likewise other domains are mind-dependent if they depend on attitudes in some way.

---

well—see Sider (2012, Ch. 6). Much of Lewis (1986), for instance, is motivated by the desire to eliminate any primitive modal ideology in the form of terms like 'possible', and Lewis is willing to pay a high ontological cost to do it. One might also think of Moorean nonnaturalism as sacrificing ideological simplicity, by retaining an unanalyzed normative notion, in order to achieve greater explanatory power. The Existence View, on the other hand, locates the distinctive metaphysical commitments of the nonnaturalist in her ontological (and not ideological) commitments.

18 See Goodman (1955, 59–65) for an interpretation along these lines.

19 Street (2006, 110).

20 For more discussion of Mind-Independence and realism, see in other contexts Cameron (2008), Devitt (1991), Jenkins (2005), Pettit (1991), Putnam (1981), and Wright (1992).

21 The "various ways" of spelling this out can be obtained by either (1) specifying different *kinds* of mental states for the purportedly dependent domain (beliefs, desires, etc.); (ii) specifying *whose* mental states are at issue (the speaker's, the ascribee's, etc.), or (iii) specifying *how* the dependence relation is to be construed (viz., the difference between modal and essential dependence in Jenkins (2005)). Of course, these options aren't mutually exclusive, and one might combine (say) an ascribee-dependence view with the claim that the dependence is mere modal independence.

2.3. Fundamentality-Based Views

A final metaphysical approach to realism proceeds in terms of the notion of *metaphysical fundamentality*. There is a family of related notions in the literature; these include "Reality" in Fine (2001); "Structure" in Sider (2012), and "perfect naturalness" in Lewis (1983). Ralph Wedgwood (2007) articulates the relationship between this idea and realism in the following passage:

> What exactly is realism? Following Kit Fine (2001) I shall suppose that a realist about the normative is a theorist who says that there are normative facts or truths—such as the fact that certain things ought to be the case, or that it is not the case that certain things ought to be the case—and that at least some of these normative facts are part of reality itself.
>
> The notion of *reality* invoked here is a notion that has its home within a certain sort of metaphysical project—namely, the project of giving a metaphysical account or explanation of everything that is the case in terms of what is real . . . [I]f certain normative facts are real, then . . . these normative facts, properties or relations may also form part of the fundamental account or explanation of certain things that are the case.[22]

Wedgwood and his predecessor Fine primarily use the term 'Reality' to signify the metaphysically privileged layer at which gives "a metaphysical account or explanation of everything that is the case." For terminological uniformity, I will instead use the term 'fundamental'. In the sense in which I intend it, then, it is a blanket term for the family of notions employed by Fine, Sider, and Lewis. It stands for a metaphysically privileged or basic category that stands in a privileged, explanatory relationship to other nonbasic facts.

This implies a distinction between "everything that is the case" and what is fundamental in the relevant sense. In most cases, something that is the case will *not* be fundamentally the case. Hence, there is a straightforward difference between a conception of realism that appeals to fundamentality and the Existence View, since simply existing does not guarantee fundamentality.

Realism about the ethical, as Wedgwood describes it, is the view that the most fundamental explanation of everything that is the case makes reference, in part, to ethical facts or properties. The ethical features in basic metaphysical explanations of the relevant kind. This conception of realism generalizes easily to other domains: realism in general is the view that the domain in question is fundamental. This is the Fundamentality View.[23]

---

22 Wedgwood (2007, 1–2).

23 Fine does not appear to accept the position suggested by the Wedgwood quote above. Fine's view on Fine (2001, 28) allows that some nonfundamental truths may also be Real, although there is a presumption in favor of their not being Real. Fine also allows that there may be some basic truths that are not Real because they are "nonfactual," though he is explicit that these nonfactual truths are also no fundamental. I will work with the simpler version of the Fundamentality View from Wedgwood.

3. Realism and Reduction

§2 outlined three prominent metaphysical approaches to realism. In this section, I will present some arguments against each view. In keeping with the methodology set out at the beginning of this paper, I will not simply cite intuitive counterexamples to each view. Rather, I will outline plausible and general structural features that are characteristic of realism. Any account that does capture all of these features will have a good claim to not achieving the right results in particular cases by objectionable gerrymandering. Each of these structural features, however, can be motivated and illustrated by reference to particular examples. In particular, each of these examples is a case of a reduction that intuitively is (or is not) consistent with realism about the reduced domain. For example, we can compare the structural features with intuitive judgments about the Russellian reduction of physical objects to logical constructions of sense-data,[24] the Lewisian reduction of modality to quantification over maximally complete chunks of concrete spacetime,[25] and the Logicist's reduction of mathematics to logic.[26] The concern here will not be with whether reductions like these are correct; the purpose of these alleged reductions—correct or not—is to illustrate some structural features of realism.

3.1. Structural Features

The structural features in question are the following:

Truth Independence Irrealism about a domain *D* is compatible with the existence of substantive truths about *D*.

Domain Neutrality For any domain *D*, 'realism' and 'irrealism' can apply nontrivially and univocally to *D*.

Reduction Compatibility For some domains *D*, some reductive views are irrealist about *D* while other reductive views about *D* are realist.[27]

Truth Independence says that irrealism is compatible with the existence of substantive truths. Just because it is *true* that Jane is in pain, realism about mental states doesn't follow. We can adopt an irrealist understanding of Jane's pain. **Domain Neutrality** requires that there is a sense of 'realism' that applies across domains: one can be realist (or not) about physical objects, mental states, and God. According to **Domain Neutrality**, there is something each of these positions has in common. (This is not to say that in addition to this univocal sense of the term, there are other distinct and domain-specific senses as well.) Finally, **Reduction**

24 Russell (1912).

25 Lewis (1986).

26 Whitehead and Russell (1910).

27 These structural features are also discussed in Dunaway (2017).

Compatibility says that a reduction of a domain does not thereby imply irrealism. Given a proposed reduction, it is a further question whether the denial of realism follows.

The three prominent characterizations of realism outlined above have difficulty accounting for all of these structural features.

3.2. Existence Is Futile

Begin with the distinction between reductions that have been labelled "vindicating" versus those that are "eliminative." At a first pass, the difference is something along the following lines. Vindicating reductions give an informative characterization of the reduced property or domain—they tell us something about the nature of the reduced thing. Other reductions—the eliminating reductions—show us that what we thought we were talking about isn't really there. A scientific reduction of water to H_2O is a vindicating reduction; someone who claimed to have a "reduction" of God to a naturally occurring phenomenon would hold a view on which God does not exist. This would be an eliminating reduction of the theological. I return to this example in the concluding section.

Railton (1989, 161) discusses another alleged instance of this contrast:

> The successful reduction of H_2O reinforces, rather than impugns, our sense that there really is water. By contrast, the reduction of "polywater"—a peculiar form of water thought to have been observed in scientific laboratories in the late 1960's—to ordinary water-containing-some-impurities-from-improperly-washed-glassware contributed to the conclusion that there really is no such substance as polywater. Whether a reduction is vindicative or eliminative will depend on the specific character of what is being reduced and what the reduction basis looks like.

This is an intuitive difference—it really does seem like, upon learning of the relevant reductions, beliefs about polywater are discovered to be mistaken, while no widespread error is revealed for beliefs about water. The Existence View would hold that, on discovering these facts about polywater, we would be irrealists about polywater because we accept an eliminating reduction of polywater. Irrealism about polywater can be read off from ordinary claims about whether polywater exists. Thus, the Existence View denies **Truth Independence**.

Prima facie this difference is superficial at best. It would be quite natural to go on, after learning of the relevant discoveries, to speak as if polywater *does* exist, but only fails to be the natural kind we thought it to be. We could say things like the following:

P1 There is polywater in this glass, since it contains water-plus-impurities-from -improperly-washed-glassware;
P2 Polywater is a very unnatural, gerrymandered chemical kind, and does not have any place in good chemical explanations.[28]

28 I do not claim that we *do* speak this way, only that we easily could make these claims using our word 'polywater'.

This, in broad outline, is a problem for the Existence View. For we are clearly not realists about polywater, even if we accept **P1** and **P2**. Their intelligibility is a consequence of the **Truth Independence** feature outlined above. The Existence View is incompatible with it.

It is worth digging deeper into whether proponents of the Existence View can mount a defense. The terms 'water' and 'polywater', like many terms with a life in a theoretical discipline, are associated with a "theoretical role" that determines as referent the property that best satisfies a set of theoretical constraints. These constraints include the observed properties of the relevant substance (that it is wet, clear, drinkable, etc.), the role it plays in explanations (that salt dissolves in it), among other things. While these theoretical constraints tolerate *some* divergence in a candidate referent, if the best candidate strays too far from the intended role, the term fails to refer.[29]

This observation gives us an existence-based explanation of the difference between the water and polywater reductions. The water reduction supplies a property (H_2O) that sufficiently approximates the theoretical role associated with 'water'; the polywater reduction supplies a property (water-plus-impurities) that, we might claim, does even approximate the theoretical role associated with 'polywater'. **Truth Independence** is still false on this revised version of the Existence View—irrealism about polywater can be read off from the falsity of statements associated with the theoretical role for 'polywater'.

Some care is needed in making the case that water and polywater reductions do differ in the truth of associated role-statements. Railton notes that even the water reduction doesn't provide a *perfect* satisfier for the relevant theoretical role—"even the reduction of water to H_2O was in part revisionist . . . of both common-sense notions and previous chemistry"[30]—the difference is, the polywater reduction is much *more* revisionary.[31] There must be some leeway for a theoretical term to have its role imperfectly satisfied, and still refer, but not too much leeway—if the theoretical role is almost entirely unsatisfied, then we will have to say that it does not refer. Wherever this cut-off point lies, it will spell trouble for the Existence View. If the theoretical role associated with 'polywater' goes unsatisfied, 'polywater' does not refer, and hence it isn't true that it is in a glass containing water-plus-impurities. **P1** and **P2** are, despite surface appearances, incoherent. If, on the other hand, we concede that the

29 See Lewis (1970).

30 Railton (1989, 161)

31 Here is a sketch of a story about why the reduction is too revisionary. Presumably those who originally introduced the term thought they discovered a new, interesting form of water with a molecular basis similar to that of water. This supplies a theoretical role for 'polywater', one that places requirements on the molecular structure of its referent. (Compare, for instance, the difference between the gerrymandered molecular basis for polywater and ^2H_2O, or "heavy water". This has a molecular basis similar to that of water; scientists presumably thought they were discovering a similar molecular variant of water when they coined 'polywater'.) But upon discovering that the "substance" in question was really just water-plus-impurities, we learn that the theoretical role isn't even close to being satisfied; a substance that is water-plus-impurities does not have molecular basis similar to that of water.

polywater-role is satisfied enough, to secure a referent for 'polywater', then the Existence View will have to concede that polywater is real.[32]

The point can also be extended to other domains. Mackie, for instance, is naturally interpreted as claiming that there are no properties that come close to satisfying the theoretical role for 'wrong'. This is because the theoretical role for 'wrong' requires that its satisfier be objectively prescriptive, and nothing (according to Mackie) comes close to satisfying that role. The consequences of this claim for the reality of ethics should be kept separate from questions of whether ethics exists *at all*. Mackie goes to an extreme in committing to an irrealist view of ethics by going in for moral skepticism. In short, realism should be *independent* of truth, in the sense that there are some truths that are about the real, but not all truths are about the real.

There are additional examples that illustrate this structural problem for the Existence View. A Vitalist view of living organisms (such as can be found in Bichat (1801, §1)) is substantially revisionary in view of the theoretical role we at present associate with 'life'. Quite plausibly, the role actually associated with 'life' is one that includes the claim that life is explained by biological and chemical processes. Hence, the life role requires that its satisfier not be an unexplained, primitive life force, as the Vitalist view holds. Thus, it fails to qualify as realist according to the Existence View, since some core role statements associated with 'life' are false. The Vitalist view, however, *is* realist; its vices stem in part from the fact that it is unnecessarily realist about life, giving it a basic explanatory role when none is needed.

Similarly, a Thomistic view about value identifies goodness with metaphysically foundational facts about teleology (in this case, with facts about the ends necessarily sought by members of a kind). A similar view can be found in Avicenna, who holds since God as the only necessary existent, is responsible for and "completes" the existence of all other things. Since each thing desires its own existence, it thereby also desires God.[33] It is another highly realist but—judged by our modern nonteleological picture of the cosmos—is a substantially revisionary account. In short, whenever the revision goes in the direction of giving a too much of a fundamental explanatory role to the reduced property, we will have a case of a revisionary view that is nonetheless realist. Such views suggest that an understanding of realism tied to substantial satisfaction of the associated role statements is bound to fail.[34]

---

32 Thanks to David Manley for suggesting this reading of the Existence View.

33 Aquinas (1920, 1a 2e Q. 1 A. 8); Avicenna (2005, 284).

34 Quine (1960, 265) appears to raise a different worry when he suggests that the distinction between vindicating and eliminating reductions itself fails to be substantive:

> For a further parallel consider the molecular theory. Does it repudiate our familiar solids and declare for swarms of molecules in their stead, or does it keep the solids and explain them as subvisibly swarming with molecules? . . . The option, again, is unreal.

We might explicate Quine's thought as follows: it is indeterminate whether we associate with 'solid' a theoretical role that is adequately satisfied by subvisible swarms of molecules. What exists isn't in question: it is swarms of molecules. But whether this is sufficient for the truth of the sentence 'solids exist' is just a

To sum up: the question of whether realism about a domain holds cannot be answered simply by looking at ordinary claims about the existence of the domain (or from the truth of the associated role statements). Given **Truth Independence**, the project of finding general connections between ordinary claims and realism fails; it is possible to adopt an irrealist construal domains that exist (e.g., polywater), and realist views might entail radically false claims about a domain (e.g., life).

3.3. Too Much Mind-Independence

On the Mind-Independence View, only views that entail that a domain is mind-independent are realist. This view violates the **Domain Neutrality** constraint. The violation is most obvious when we consider the reductive Behaviorist view of mental states, which is a paradigmatically irrealist view of the mental. Mental states, according to the Behaviorist (as I will understand the view), are just disjunctions of behaviors or dispositions to behave. The mental state *pain* on this view reduces to either clutching one's arm, or screaming, or . . . (the disjunction of behaviors will need to go on for quite some time in order for the Behaviorist view to be truth-conditionally adequate). Similarly for other mental states. Again, the resulting view is intuitively an irrealist one.

It also satisfies the conditions imposed by any reasonable construal of Mind-Independence. That Jane is exhibiting the behavior of clutching her arm doesn't depend on the mental—arm-clutching is just a movement of the body. And on the Behaviorist view, Jane's pain just is an occurrence of the behavior of arm-clutching. So, on the Behaviorist view, Jane's pain is an occurrence that is as objective and mind-independent as any, since the fact that Jane is moving her body in a particular way is objective and mind-independent. The Mind-Independence View makes Behaviorism a realist view about the mental. The Mind-Independence View is highly questionable on this score.[35]

---

matter of whether we choose to associate a more or less strict theoretical role with 'solid'. The truth (or falsity) of this sentence doesn't reflect a deep metaphysical fact—just a choice about our language.

On this way of explicating Quine, his comments suggest that talk of existence isn't sufficiently metaphysically robust to capture the metaphysical dimension to realism.

35 Cameron (2008) defends Mind-Independence views from objections from realism about the mental. Following a distinction from Jenkins (2005), he says that realism should be understood in terms of "essential dependence": irrealism about a domain holds that its "existence or essence is constitutively dependent on mental activity" (Cameron 2008, 7). As Cameron notes, this distinction helps with avoiding the allegation that realism about the mental is trivially false, since it is trivially true that the mental depends on the mental. Cameron's point is that this doesn't follow when 'depends' is glossed as essential dependence: a mental entity can essentially depend on a nonmental event.

But the problem posed by the Behaviorist remains, even with this distinction in place. If we explicitly add that, according to the Behaviorist view, the mental constitutively depends on nonmental behaviors, the view remains intuitively irrealist. The true worry for Mind-Independence accounts applied to the mental, then, isn't that realism is too hard to come by; rather, it is *too easy*, letting even the Behaviorist in. See also Reynolds (2006, 481) and Rosen (1994, 286–9) for more discussion of the relationship between Mind-Independence and realism.

This is a rejection of **Domain Neutrality**, since Behaviorism must be treated differently from other irrealism about other domains, on the Mind-Independence View. Failures of **Domain Neutrality** will arise for other broadly mental phenomena.[36]

Some authors such as Miller (2010, §1) claim that the Mind-Independence conception of realism cannot fail to account for realism about psychological phenomena precisely *because* the dependence is trivial. The Mind-Independence View, these authors claim, should be understood as holding that what makes theories irrealist is that they entail that their domain to be (1) mind-dependent but (2) not *trivially* mind-dependent. Adding a nontriviality constraint for irrealism doesn't, in the first instance, make Behaviorism out to be irrealist: the view is a nontrivial claim that mental states depend on nonmental phenomena. Second, it threatens to make 'realism' apply with different senses to different domains. A standard realist view of ethics holds that ethical facts obtain independent of what we think about them, and so, (according to the Mind-Independence View) it entails that ethics is real. A nonreductionist view of belief, by contrast, makes belief trivially dependent on other beliefs, and thereby counts as a realist theory on these grounds. In order to use Mind-Independence to adequately characterize these views, we would need to violate the univocality condition in **Domain Neutrality**.

Others construe the Mind-Independence condition to be a condition on the mental states of an assessor, rather than the mental states of the subject of the assessment.[37] On these views, the fact that Jane is in pain, if it is not a real fact, depends on what *we* believe about Jane, and not on Jane's beliefs or other states. Behaviorism, then, won't count as irrealist simply because it holds that the reduction base for pain includes nonmental behaviors only. If Behaviorism is an irrealist view, it must be because it entails that whether Jane is in pain depends on what we think about Jane's pain states. Behaviorism doesn't have such an entailment. Instead, Jane is in pain according to the Behaviorist if she is clutching her arm, even if as assessors of whether Jane is in pain, we never know, or believe, or have evidence that Jane is clutching her arm. Understanding the Mind-Independence View in terms of independence from the assessor's states fails to accommodate the irrealism of Behaviorism.

---

36 Consider the account of syntactic principles, or "grammars," in Noam Chomsky's *Knowledge of Language*, which is one on which they are accurate descriptions of a psychological state realized in the brains of competent language users. Grammars are, in Chomsky's terms, "psychologically real." This is a distinctive and (strikingly) realist position about grammars; its competitors include views on which grammars are merely the simplest set of axioms whose theorems are all and only the grammatically acceptable sentences making no claims about psychological reality in the process. (See Chomsky (1986, 39) and Soames (1989) for competing sides of the debate.) The Mind-Independence View fails to find any relevant difference between these views, since both views make grammars mind-dependent.

37 This kind of assessor-sensitivity is present in the formulation of Mind-Independence from Brink in §2.

3.4. Fundamental Failings

The Fundamentality View holds that realist views about a domain are just those that take the domain to be fundamental. This view nicely accommodates our first two structural features. It retains **Truth Independence** because ordinary claims about a domain do not entail whether the domain is fundamental or not. It is an ordinary, prosaic fact that Sally is in pain, but whether this fact is part of fundamental reality is not settled by ordinary claims about pain alone. And it accepts **Domain Neutrality** since any domain (including the mental) might, or might not, be fundamental. These structural advantages are bolstered by reflection on certain examples. By analyzing mental states in terms of behaviors, the Behaviorist view entails that mental states are not most fundamental. And, by analyzing wrongness in terms of speakers' attitudes of disapproval, the Subjectivist view entails that wrongness is not most fundamental. Something is more fundamental than each, namely behaviors or attitudes of disapproval. They are both irrealist views, as the Fundamentality View predicts.

This way of getting the right results in some cases for the Fundamentality View gets them for the wrong reasons: *any* analysis of a domain will entail that it is not fully fundamental, and hence will be an analysis that entails irrealism about the analyzed domain. There are plenty of examples of analyses that are consistent with realism.

Here are two. An identity theorist such as Place (1956) reduces mental states by identifying them with neurophysiological states. Hence, according to the identity theorist, pain is not most fundamental; some neurophysiological state is more fundamental than it. Likewise, the view of moral properties like wrongness presented in Railton (1986) is one on which they reduce to facts about what promotes human interests from the "social point of view." Hence, human interests are more fundamental than moral properties. Both views are, however, intuitively consistent with realism—Railton even presents this view in a paper called "Moral Realism."[38]

These examples illustrate the failure of our third structural feature of realism on the Fundamentality View: it cannot accommodate **Reduction Compatibility**. Any reduction of a domain will, by the logic of fundamentality, entail that the domain is not fundamental. The Fundamentality View will classify the domain as irrealist. This is a rejection of **Reduction Compatibility**, and as this structural feature is quite plausible, alternatives to the Fundamentality View should be considered.

It is worth mentioning as an additional point that certain moves to avoid this result will be unhelpful to the fundamentality theorist. It might seem promising to adopt a different approach to the relationship between reduction and fundamentality. For instance: begin

38 Readers familiar with that paper will note that Railton acknowledges on pp. 200–1 that his view lacks some of the characteristic features of realism (though he nevertheless claims that it resembles realism enough to deserve the name). I will return to the question of *how* realist Railton's view is in later sections. But as a purely structural point, it would be highly surprising if Railton's view failed to qualify as realist simply because it is reductive in character.

with the idea that there are some fundamental terms: perhaps those standing for basic entities and properties such as 'quark', 'spin', and so forth. A *fundamental fact*, we can then say, is one that can be specified with fundamental terms only. Facts about the identity theorist's reduction basis will, presumably, be fully fundamental since they can be specified using the relevant microphysical terms. The fundamentality theorist might then add that reductions are identities—to reduce the fact that Sally is in pain to the fact *F* is just to claim that the fact that Sally is in pain is *identical* to *F*. It then follows that facts about mental states are, according to the identity theorist, fundamental.[39]

This approach to reduction is available to the fundamentality theorist and, if she adopts it, she can claim that some reductions *are* compatible with realism. The problem, however, is that a fourth structural feature on realism is just as plausible as **Reduction Compatibility**. We can call this fourth feature ***Reduction Independence***; according to it, a reduction of a domain does not *entail* realism about the domain, some reductions in fact imply irrealism about the reduced domain. **Reduction Independence** is quite plausible; Behaviorism as sketched above provides one example of a reduction of mental states is not realist about mental states. The revised version of the Fundamentality View captures **Reduction Compatibility** at the expense of **Reduction Independence**. For the revised logic of reduction will apply to any reduction and, as a result, realism will follow for any reduced domain—including the Behaviorist's mental states.

4. Relative Fundamentality and Realism

The best solution requires a distinct metaphysical category: *degrees* of fundamentality. For instance: it is entirely natural to say that if acids are electron-pair acceptors, then there is something that is more fundamental than acidity, namely electrons. Likewise, if gravity is curvature in spacetime, then spacetime points are more fundamental than gravity. And if galaxies are collections of stars and other celestial objects surrounded by an interstellar medium, then stars are more fundamental than galaxies.

I will call these claims of the form '*A* is more fundamental than *B*' claims about *relative* fundamentality, or claims about *degrees* of fundamentality.[40]

For the sake of clarity, I will briefly sketch a theory of relative fundamentality that can then be put to use in giving an account that captures all of the structural features of realism. However, it is important to note that other understandings of relative fundamentality are available, and can in principle be adapted to providing an account of realism along the lines

39 See, for instance, Sider (2012, Ch. 7) for a similar idea.

40 Strictly speaking, these do not amount to the same thing: it could be that *A* is more fundamental than *B*, while there are no specific *degrees* of fundamentality, d_A and d_B, such that *A* is fundamental to degree d_A, *B* is fundamental to degree d_B, and $d_A > d_B$. At times, I will speak as if these degrees exist, but much of what I say below can be rephrased (albeit in somewhat more complicated language) using only the comparative 'more fundamental than' and without reference to degrees.

sketched here. Much of what we said by way of introducing the notion of absolute fundamentality in §2.3 applies to relative fundamentality as well: electrons, for example, provide a kind of "metaphysical explanation" for facts about electron pairs; stars provide the same kind of explanation for facts about galaxies, and spacetime points provide the same kind of explanation for facts about gravity.

Two clarifications are in order here. We can understand the absolutely fundamental with a helpful slogan from Fine (2001); it is that which provides the "most satisfying" metaphysical explanation for some fact. Obviously, this kind of gloss applies to that which is absolutely fundamental, and cannot be applied directly to explain relative fundamentality. Since electrons have further explanations in terms of the subatomic, electron-pair acceptors do not provide the most satisfying metaphysical explanation of acidity. Still, we can say that what is more fundamental provides the same *kind* of metaphysical explanation; it simply need not provide the most satisfying version of this kind of explanation. Thus, the electron-based explanation of acidity is still a metaphysical explanation of the same kind, even if it isn't the final explanation.

An analogy with causal explanation may be helpful here: one can causally explain the breaking of a window in terms of the fact that the ball that was thrown, its trajectory, the fragility of the glass, and so forth. This is a perfectly legitimate causal explanation if filled out appropriately. It isn't, however, the *final* causal explanation: that would make reference to the causal precursors of the throwing of the ball, and the causal precursors of the precursors, and so on, perhaps only terminating in a description of the Big Bang. A most satisfying causal explanation of this kind doesn't preclude the existence of more proximate, nonfinal causal explanations. That which is *more* fundamental similarly provides more proximate nonfinal metaphysical explanations.

The second clarification is that the examples of differences in relative fundamentality mentioned above all represent discoveries from the physical sciences—in particular, chemistry, physics, and astronomy. This might be thought to distinguish relative fundamentality, as I have described it here, from the notion of absolute fundamentality as developed by Fine and others. On these approaches, the absolute notion is approached through metaphysical or philosophical theorizing—and not through empirical science. The relative notion of fundamentality, as described here, appears not to be well-suited to feature in a metaphysical account of realism.

The appearance of an important difference may, however, be misleading. There are some approaches to absolute fundamentality where empirical science does play a central role, but in which the notion of fundamentality retains its metaphysical character. Lewis's conception of "perfect naturalness" (his terminology for what amount to absolutely fundamental properties) also assigns a central role to empirical science. He says:

> To a physicalist like myself, the most plausible inegalitarianism seems to be one that gives a special elite status to the 'fundamental physical properties': mass, charge, quark colour and

> flavour . . . (It is up to physics to discover these properties, and name them; physicalists will think that present-day physics at least comes close to providing a correct and complete list.)[41]

Lewis thus gives physics (or something close to it) a close relationship to the absolutely fundamental. The "close relationship" isn't one that undermines its metaphysical character. Lewis isn't proposing to *define* the absolutely fundamental in terms of the practices of physicists. Instead, this picture is one on which physics provides at best an epistemology of absolute fundamentality. That physics makes reference to quarks doesn't make quarks most fundamental; rather, it is simply the means by which we know that they are. Similarly, then, for other sciences and relative fundamentality: these sciences provide an epistemic window into the facts about relative fundamentality, but do not constitute them. Once we separate the epistemic from the metaphysical dimension to fundamentality, the relative version is in no worse shape to feature in a metaphysical account of realism.[42]

We can extend these ideas to degrees of fundamentality: by engaging in the right kinds of first-order inquiry in science, philosophy, and elsewhere, we can come to learn about what is fundamental, and to what degree. These degrees of fundamentality are, metaphysically speaking, primitive: they are not grounded in, or determined by, facts about first-order inquiry.

With these clarifications in place, we can investigate a positive proposal concerning the natural kind that underlies out talk of 'realism'. The account begins by introducing a new resource: relative fundamentality. I will argue that this resource is more promising than the existing alternatives, as it has additional structural features that allow it to capture the **Truth Independence**, **Domain Neutrality**, and **Reduction Compatibility** features of realism.

4.1. Accounting for the Structural Features

I will focus here on developing a relative fundamentality-based account to deal with the most difficult feature, **Reduction Compatibility**. Then I will note in closing how the resulting account also handles **Truth Independence** and **Domain Neutrality**.

Begin with the difference between an identity theorist about mental states (a realist) and a reductive Behaviorist (an irrealist). It is very natural to say that the difference between the two views lies in how fundamental, according to each view, mental states are. If pain is a particular neurophysiological state, it is a fairly natural psychological kind and hence is more fundamental than it is if it is a disjunction of behaviors. Pain, if it is metaphysically explained by a highly disjunctive state (such as the Behaviorist's disjunction of behaviors), is thereby not very fundamental. This suggests that the difference between Identity Theory

41 Lewis (1984, 228). See also Schaffer (2004) for an extension of this position to sciences beyond physics.

42 For more on the epistemology of nonabsolute fundamentality, see Dunaway and McPherson (2016) and Dunaway (2020, Ch. 5).

and Behaviorism is that pain, on the Behaviorist view, fails to meet a particular *threshold* of fundamentality. A crude account of realism about mental states runs as follows:

> **Mental State Realism (MSR)** There is a degree of fundamentality *d* such that a theory *T* is realist about mental states just in case *T* entails that mental states are fundamental to (at least) degree *d*.

The assumption behind **MSR** is that only the identity theorist's view is realist because only it entails that pain meets some threshold of fundamentality. Note, however, that this threshold needn't require a very *high* degree of fundamentality. Identity theory might imply that mental states aren't very fundamental at all. All **MSR** requires is that competing irrealist views, like Behaviorism, imply that mental states are even less fundamental.

In order to maintain **Domain Neutrality**, we should adopt analogues of **MSR** for other domains. The question arises of whether the threshold for realism is the *same* for each domain. That is: is it the case that there is a single degree of fundamentality *d* such that realist views about *any* domain entail it to be fundamental to degree *d*?

Here is a simple argument that the answer is 'no'. An identity theory of mental states holds that mental states are neurophysiological states. A number theorist might identify numbers with similar entities—perhaps the synaptic firings that correspond to counting operations in normal human minds. Thus, the number 2 on this view reduces to the neurophysiological state that occurs when normal humans count to the second item in a normal counting sequence. The reduction base for pain and the number 2 are then very similar in kind according to these views; plausibly pain and numbers are, on these views, fundamental to the same degree. Identity theory, however, seems clearly to be a realist view of pain, while our psychological reduction of numbers entails an irrealist view about numbers. So, the threshold for realism about numbers and mental states must be set at different points on the scale of degrees of fundamentality. (Perhaps even for a single domain the threshold can be set at different levels in different contexts. I won't take a stance on this question here.)

If there is variability in where the threshold for realism is set, one approach to accommodating it is to take another aspect of the analogy with gradable adjectives seriously. For 'loud' and other gradables, the threshold is set by conversational context. Exactly what features of context are relevant, and how they conspire to set a standard for loudness, is a tricky matter. It is clear that my coffee grinder counts as loud in some contexts and not others, and that the difference between these contexts in part has to do with the *comparison class* at issue.[43] The comparison class contains contextually and conversationally salient objects, and determines in some way where on the scale of volume the threshold for

43 See Klein (1980) for discussion of the notion of a comparison class and Ludlow (1989) for more on the ways in which comparison classes are fixed.

loudness is to be set. In contexts where the comparison class contains only chirping crickets, my coffee grinder counts as loud; in contexts where the comparison class contains only train whistles, it does not.

The comparison class in a discussion of realism is naturally taken to include other salient views about the domain in question. Thus, when realism about mental states is at issue, the comparison class includes theories of mental states that conversational participants take to be relevant. This comparison class then sets a threshold for fundamentality. Quite plausibly, the salient views about mental states will constitute a comparison class that determines a degree of fundamentality more demanding than the degree to which mental states are fundamental on the Behaviorist view. There are many salient and plausible theories of mental states that make them out to be more fundamental than the Behaviorist does.

The story about realism in other domains is a variation on this theme. When domains other than the mental are at issue, the comparison class is different as well: if we shift to a discussion of realism about numbers, then salient theories of numbers populate the comparison class, not theories of mental states. This shift may well determine a different threshold for realism. Without taking a stand on the exact mechanisms by which competing views populate a comparison class, and the precise way in which a comparison class determines a threshold, the account of realism is, in general form, as follows:

> **Realism** For any domain *D*, the comparison class for *D* determines a degree of fundamentality *d* such that a theory *T* is realist about *D* just in case *T* entails that *D* is fundamental to (at least) degree *d*.

Realism is natural as an account of what makes the Instrumentalist an irrealist about the unobservable entities of scientific theories, and what makes the Vitalist view of life highly realist, to take a few examples. The Intstrumentalist is plausibly construed as holding that instruments of measurement are more fundamental than unobservables, since on her view, facts about the reading of instruments *explain* facts about unobservables. And the Vitalist is plausibly construed as holding that biological processes are *not* more fundamental than life—and hence that the latter is highly fundamental since they are not explained by biological processes at all. With a normal comparison class for each domain, **Realism** entails that the Vitalist has a realist view of life, while the Instrumentalist does not have a realist view of unobservables.

4.2. Return to the Structural Features

I have sketched how Realism plausibly categorizes some examples of realist and irrealist views. We could run through more examples, asking (e.g.) whether Moorean nonnaturalism and Thomism imply that ethics is highly fundamental, or whether Subjectivism implies that it is not. Instead of doing this, however, I will turn to the structural features of realism from §1. **Realism** is well-suited to accommodating all of these; I will briefly sketch how.

Some reductions imply that a domain is not very fundamental at all—for example, this is true of the reduction of polywater. But not all do. Many settings of a threshold for realism will result in a classification of some, but not all, reductive views of the domain as realist. Thus, Realism plausibly implies **Reduction Compatibility**.

Realism is consistent with **Truth Independence**. Some theories will entail truths about a domain but make these truths out to be not-very-fundamental and hence not reaching the threshold of fundamentality required for realism. The same truth might be very fundamental according to one theory and not-very-fundamental according to another. So, ordinary truths by themselves won't settle questions about realism.

Finally, Realism can accommodate **Domain Neutrality**: the mental might, just like any other domain, be more or less fundamental depending on which theory of the mental is in play. The same considerations will then apply in assessing the realism of a particular theory, namely whether the mental meets the required threshold for realism according to the theory. In each case, it is a comparison with the degree of fundamentality of the relevant domain according to other salient theories that settles the question of realism.

5. Conclusion

Realism, I have argued, represents a much-improved attempt at an account of the metaphysically natural kind that underlies philosophical talk about 'realism'. Whether it provides a fully satisfactory account, given the parameters set out in §1, can be debated.

It is worthwhile to recall the desiderata set out in §1: we want a single metaphysical kind that plays the structural roles distinctive of realism. I have argued that **Realism** gives an account that entails all of the structural features outlined in this paper. There are substantial questions to be asked concerning whether **Realism** articulates a *metaphysical* kind, and whether there is a *single* kind underlying the account. These are worthwhile questions to ask but in the interest of space I will not address them here. Instead, I will only note that **Realism** seems to be the best approximation of the realism-role on offer, as it does better than the other conceptions of realism from §2.

To close, I will mention one more point in favor of the account presented here. Many philosophers have found the Existence, Mind-Independence, and Fundamentality views to be very compelling accounts of realism. If the objections of §§3–4 are correct, these views fail for very straightforward reasons. What can explain their appeal? The **Realism**-based account has a simple answer: existence, Mind-Independence and absolute fundamentality often stand proxy for a greater degree of fundamentality.

Take Existence views first. There are some cases where we, with good reason, restrict the theoretical roles we associate with a term to require that its referent be highly fundamental. One example is found in Schroeder (2005): with theological terms like 'God', one doesn't count as holding that God exists if one accepts a reductive account of their referents. If one identifies the referent of 'God' (as in Schroeder's example) with the strong nuclear force that

holds positively charged protons in atomic nuclei together, one does not thereby count as someone who holds that God exists. Finding just *any* existing referent for 'God' does not suffice for theological realism.

This is not obviously a failure of the reductive account to give a truth-preserving interpretation of the theoretical claims associated with 'God'. If the reductive account starts by assigning the strong nuclear force as the referent of 'God', it can go on to give interpretations of 'God is a person', 'God created the universe', 'God loves humankind', and so forth, on which these sentences are true.[44] Why then does the reductive account fail to be theologically realist? The problem is that, by beginning with an assignment of the strong nuclear force as the referent of 'God', its interpretation of theological language will be highly gerrymandered and contrived elsewhere. (Consider the interpretation of 'person': it must be a property that applies to not only the strong nuclear force but also to the referent of 'human', but not to the referent of 'rock', 'planet', or 'number'. Such maneuvers will inevitably require a significant amount of gerrymandering in order to preserve the truth of many ordinary claims.) This suggests that theological terms are not only connected via theoretical role to claims expressed by 'person', 'create', 'love', and so forth; the theoretical role attached to theological terms in addition requires that the properties of personhood, creation, love, and so forth, be *highly fundamental*. Since any interpretation that starts by assigning the strong nuclear force as the referent of 'God' will end up with not-very-fundamental referents for other terms that are closely connected to the theoretical role for 'God', the reductive view does not count as holding that there is a God. 'God exists' only comes out as true when the relevant terms are assigned sufficiently fundamental referents.

The question of theological realism then goes hand in hand with the question of the existence of the theological. This, however, is a special case: it is only because the theological and associated subject matters can be expected to be fundamental, if they exist at all. This needn't, however, be true for every domain for which the question of realism can arise. Discovering the constitution of polywater needn't show that polywater doesn't exist, as there need not be an expectation that any referent for 'polywater' is a highly fundamental one.[45] More generally, once the domain in question isn't one that can be expected to be highly fundamental if it exists at all, it can still be properly said to *exist* even if it turns out to be highly gerrymandered

44 More generally, permutation arguments inspired by Hilary Putnam (1981) claim to show that there will be many truth-preserving interpretations of a language. See also Button (2013, Chs. 1–4) for more detail on Putnamian permutation arguments.

45 Note that there *might* be such an expectation: as we filled out Railton's example earlier, one might, prior to the relevant discovery, associate with 'polywater' a theoretical role that requires it to be of the same kind of molecular constitution as ordinary water. But, we emphasized, this isn't required: one might also continue to use the term 'polywater' with the same meaning after the discovery. One can consistently do this so long as one associates a less strict theoretical role with the term. This kind of case shows how, at least in principle, discovery of a not-very-fundamental reduction basis need not require a denial of the existence of the reduced domain or property.

and unnatural. In these cases, mere existence won't be sufficient for realism about the relevant domain. From the perspective of Realism, the existence conception provides a sometimes (but not always) useful heuristic for when a view is realist.

Realism also explains why the Mind-Independence and Fundamentality accounts are tempting. Often, something that is mind-dependent is thereby not-very-fundamental: after all, the mind-dependence claim itself is a claim that there is something more fundamental, namely the mind. Views that entail the mind-dependence of a domain will, in general, also fail to be realist views in the sense Realism. There are exceptions: when the domain in question is explicitly mental, correlations between mind-dependence and comparatively lower degrees of fundamentality go out of the window.

Finally, absolute fundamentality will always imply realism—views on which a domain is absolutely fundamental are guaranteed to be views on which the domain meets the contextually set degree of fundamentality required for realism.[46] The converse need not hold: some views that meet the contextually set degree of fundamentality required for realism need not be views on which the domain is absolutely fundamental. The Fundamentality view, like the Existence and Mind-Independence views, provides in some cases a useful proxy for what is at issue in discussions of realism. None of these views provide a complete picture; for this we need relative fundamentality.

The result is a plausible, metaphysically substantial, conception of realism. I have not attempted to show in this paper that a Gibbardian quasi-realism is not a genuine version of realism. Rather, all I have shown is that there is more to realism than the claims that quasi-realists have explicitly shown that they can accommodate. Quasi-realists have shown that they can accept that there are normative facts and properties. And they have shown that they can coherently accept that murder is wrong, no matter what we happen to think about it. While these are claims about the existence and Mind-Independence of the normative, they are not, I have argued, the core commitments of realism.

Instead, the core commitment of realism concerns the degree of fundamentality of the relevant domain. It is, of course, possible that quasi-realists can show that this is also something which they can coherently accept. But perhaps not: Gibbard, along with other quasi-realists, acknowledges that how the quasi-realist *explains* paradigmatically realist claims will differ from how the typical realist explains the same claims.[47] I will leave the details of the quasi-realist explanations to the side, but for the expressivist these are at bottom explanations that interpret the realist-sounding claims to be expressions of plans, and then go on to show that it is coherent to plan in the relevant ways. Fundamentality is, at bottom, an explanatory notion: something is fundamental when it features in a

46 Moreover, it will always imply bivalence if there is no indeterminacy at the fundamental level. So, given this assumption the Dummettian view always correctly categorize views that are realist by virtue of taking their domain to be fundamental.

47 Gibbard (2003, 187).

"metaphysically satisfying" explanation of why something is the case, in Fine's language. So, perhaps, the expressivist has already implicitly denied a significant degree of fundamentality to the normative, by appealing to naturalistically acceptable planning states to do all of the explanatory work.[48]

I will not pursue this argument here.[49] At the very least, the discussion of realism here leaves an explanatory challenge to the quasi-realist, by giving a metaphysically robust conception of realism.

References

Aquinas, T. (1920). *The Summa Theologica of St. Thomas Aquinas*, 2nd and revised ed., translated by Fathers of the English Dominican Province. London: Burns, Oates and Washbourne.

Avicenna (2005). *The Metaphysics of the Healing*, translated by Michael E. Marmura. Provo: Brigham Young University Press.

Berkeley, G. (1710). *A Treatise Concerning the Principles of Human Knowledge, Part I*. Dublin: Aaron Rhames.

Bichat, M. F. X. (1801). *Anatomie générale appliquée à la physiologie et à la médicine*, Paris: Brossom, Gabon et Cie.

Blackburn, S. (1993). *Essays in Quasi-Realism*. Oxford: Oxford University Press.

Boyd, R. N. (1988). "How to Be a Moral Realist." In *Essays on Moral Realism*, edited by G. Sayre-McCord. Ithaca, NY: Cornell University Press.

Brink, D. O. (1984). "Moral Realism and the Skeptical Arguments from Disagreement and Queerness." *Australasian Journal of Philosophy* 62, no. 2: 111–25.

Button, T. (2013). *The Limits of Realism*. Oxford: Oxford University Press.

Cameron, R. (2008). "Turtles All the Way Down: Regress, Priority and Fundamentality." *The Philosophical Quarterly* 58: 1–14.

Chomsky, N. (1986). *Knowledge of Language: Its Nature, Origins and Use*. New York: Praeger Publishers.

Devitt, M. (1991). *Realism and Truth*, 2nd ed. Oxford: Basil Blackwell.

Dreier, J. (2004). "Meta-Ethics and the Problem of Creeping Minimalism." *Philosophical Perspectives* 18: 23–44.

Duhem, P. (1954). *The Aim and Structure of Physical Theory*. Princeton, NJ: Princeton University Press.

Dummett, M. (1977). *Elements of Intuitionism*. Oxford: Oxford University Press.

——— (1982). "Realism." *Synthese* 52: 55–112.

Dunaway, B. (2016). "Expressivism and Normative Metaphysics." In *Oxford Studies in Metaethics*, vol. 11, edited by R. Shafer-Landau. Oxford: Oxford University Press.

——— (2017). "Realism and Objectivity." In *The Routledge Handbook of Metaethics*, edited by T. McPherson and D. Plunkett. New York: Routledge.

48 Gibbard (2003, Ch. 10) discusses normative explanations but is primarily concerned with causal explanation.

49 See Dunaway (2016) for one way of developing the argument.

——— (2020). *Reality and Morality*. Oxford: Oxford University Press.

Dunaway, B., and T. McPherson (2016). "Reference Magnetism as a Solution to the Moral Twin Earth Problem." *Ergo* 3, no. 25: 639–79.

Fine, K. (2001). "The Question of Realism." *Philosophers' Imprint* 1, no. 1: 1–30.

Gibbard, A. (2003). *Thinking How to Live*. Cambridge, MA: Harvard University Press.

——— (2013). *Meaning and Normativity*. Oxford: Oxford University Press.

Goodman, N. (1955). *Fact, Fiction, and Forecast*. Cambridge, MA: Harvard University Press.

Jenkins, C. (2005). "Realism and Independence." *American Philosophical Quarterly* 42, no. 3: 199–211.

Klein, E. (1980). "A Semantics for Positive and Comparative Adjectives." *Linguistics and Philosophy* 4: 1–45.

Lewis, D. (1970). "How to Define Theoretical Terms." *Journal of Philosophy* 67, no. 13: 427–46.

——— (1983). "New Work for a Theory of Universals." *Australasian Journal of Philosophy* 61, no. 4: 343–77.

——— (1984). "Putnam's Paradox." *Australasian Journal of Philosophy* 62, no. 3: 221–36.

——— (1986). *On the Plurality of Worlds*. Oxford: Basil Blackwell.

Ludlow, P. (1989). "Implicit Comparison Classes." *Linguistics and Philosophy* 124: 519–33.

Mackie, J. (1977). *Ethics: Inventing Right and Wrong*. New York: Penguin.

Miller, A. (2003). "The Significance of Semantic Realism." *Synthese* 136: 191–217.

——— (2010). "Realism." In *Stanford Encyclopedia of Philosophy*, edited by E. N. Zalta. http://plato.stanford.edu/archives/sum2010/entries/realism/.

Pettit, P. (1991). "Realism and Response-Dependence." *Mind* 100, no. 4: 587–626.

Place, U. (1956). "Is Consciousness a Brain Process?" *British Journal of Psychology* 47: 44–50.

Putnam, H. (1981). *Reason, Truth and History*. Cambridge: Cambridge University Press.

Quine, W. V. O. (1960). *Word & Object*. Cambridge: MIT Press.

Railton, P. (1986). "Moral Realism." *The Philosophical Review* 95, no. 2: 163–207.

——— (1989). "Naturalism and Prescriptivity." *Social Philosophy and Policy* 7: 151–74.

Reynolds, S. (2006). "Realism and the Meaning of 'Real'." *Noûs* 40: 468–94.

Rosen, G. (1994). "Objectivity and Modern Idealism: What Is the Question?" In *Philosophy in Mind: The Place of Philosophy in the Study of Mind*, edited by M. Michael and J. O'Leary-Hawthorne, 277–319. Dordrecht: Kluwer Academic.

Russell, B. (1912). *The Problems of Philosophy*. Oxford: Oxford University Press.

Schaffer, J. (2004). "Two Conceptions of Sparse Properties." *Pacific Philosophical Quarterly* 85: 92–102.

Schroeder, M. (2005). "Realism and Reduction: The Quest for Robustness." *Philosophers' Imprint* 5, no. 1: 1–18.

Shafer-Landau, R. (2003), *Moral Realism: A Defense*. Oxford: Oxford University Press.

Sider, T. (2012). *Writing the Book of the World*. Oxford: Oxford University Press.

Soames, S. (1989). "Semantics and Semantic Competence." *Philosophical Perspectives* 30: 575–96.

Stevenson, C. (1937). "The Emotive Meaning of Ethical Terms." *Mind* 46: 14–31.

Street, S. (2006). "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127, no. 1: 109–66.

van Fraassen, B. (1980). *The Scientific Image*. Oxford: Clarendon Press.

Wedgwood, R. (2007). *The Nature of Normativity*. Oxford: Oxford University Press.

Whitehead, A. N., and B. Russell (1910). *Principia Mathematica*. Cambridge: Cambridge University Press.

Wright, C. (1987). *Realism, Meaning and Truth*. Oxford: Basil Blackwell.

——— (1992). *Truth and Objectivity*. Cambridge, MA: Harvard University Press.

V

THE NORMATIVITY OF MEANING

17

THE NORMATIVITY OF MEANING REVISITED[1]

Paul Boghossian

Introduction

In this essay, I revisit a topic that Allan Gibbard and I have been debating off and on for about three decades and which was ignited by Saul Kripke's (1982, 37) remark:

> The relation of meaning and intention to future action is *normative*, not *descriptive*.

Ever since Kripke made this remark, philosophers have been arguing over what this 'meaning is normative' slogan means and whether it is true on any of its legitimate interpretations.

Kripke attributed the insight to Wittgenstein (1953), although he makes it clear that, at least as far as this aspect of Wittgenstein's thought is concerned, he is sympathetic to it.

Kripke saw in the slogan a powerful weapon with which to combat proposed naturalistic reductions of meaning, in particular dispositional analyses. If the relation between the meaning of a word and its use is normative, so Kripke seems to have thought, then meaning can't consist in dispositions to use the word in a certain way, because the relation between a disposition and its exercise is descriptive not normative.

Gibbard joins Kripke in finding the slogan important; and he has been exploring it sympathetically over the past few decades. Recently, he has published a magisterial volume,

---

1 I've had the pleasure of discussing this and other topics with Allan Gibbard ever since I had the good fortune of becoming his junior colleague at the University of Michigan at Ann Arbor in 1984. Allan was an inspiring interlocutor, an insightful critic, and a generous mentor. This contribution to a volume in his honor is a totally inadequate expression of thanks.

Meaning and Normativity, which pulls together a lifetime's worth of profound reflections on the nature of normativity, the nature of meaning, and the relations between them.

Prima facie, it is somewhat surprising that Gibbard should be such a fan of the 'normativity of meaning' slogan. His other views, you might think, would disincline him from embracing it.

For one thing, Gibbard is a committed naturalist, who thinks that all facts are ultimately grounded in natural facts and properties. But doesn't the normativity of meaning make trouble for naturalism, as Kripke thinks?

For another, one of the ways in which Gibbard tries to secure a naturalistic outlook is by being an expressivist about normative discourse. Seemingly, though, expressivist meaning requires a contrast with nonexpressivist factual meaning. However, if the very notion of meaning is itself normative, then meaning talk must itself be given an expressivist characterization. And the result will be that the very distinction between expressivist and nonexpressivist meaning will itself have to be a distinction with merely expressivist meaning. Won't this be a problem? Wouldn't an expressivist want the contrast between expressivism and factualism to be itself a factual contrast?

Despite all this, Gibbard insists that the slogan is importantly true.

Normativity of Meaning as Correctness

In order to set some context for the debate, I will say a little about my own relation to this slogan.

In the literature, I have seen it said by some that I am a strong *proponent* of the normativity of meaning thesis, and by others that I am a strong *opponent*. This is not usually a very good sign about the degree of clarity that one has brought to the topic. I will try to clarify what might be going on here.

In my earliest writing on Kripke's book, "The Rule-Following Considerations," of 1989, I put forward a particular view of what the slogan meant, an interpretation that in the literature is apparently referred to as the 'orthodox interpretation.' To quote a bit from that paper:

> Suppose the expression 'green' means *green*. It follows immediately that the expression 'green' applies *correctly* only to *these* things (the green ones) and not to *those* (the non-greens) . . . The normativity of meaning turns out to be, in other words, simply a new name for the familiar fact that, regardless of whether one thinks of meaning in truth-theoretic or assertion-theoretic terms, meaningful expressions possess conditions of *correct use* . . . Kripke's insight was to realize that this observation may be converted into a condition of adequacy on theories of the determination of meaning: any proposed candidate for the property in virtue of which an expression has meaning, must be such as to ground the 'normativity' of meaning—it ought to be possible to read off from any alleged meaning constituting property of a word, what is the correct use of that word. (513)

On this view, clearly, the label 'normativity of meaning' is being used fairly lightly. Most of the philosophers who get excited about Kripke's slogan think that he is pointing to something that (even if not entirely without precedent) would be considered highly controversial, and which would be denied by most naturalistic reductionists about meaning.

By contrast, I say that it is just a new label for a very familiar phenomenon that the meaning of a linguistic expression—say a predicate—is inextricably linked to some notion of correct application: the main candidates being either true application or warranted application. And this latter claim no one denies.

For all that this use of 'normativity' is light, it is not entirely without interest or import. As Gideon Rosen (1997) has rightly emphasized, the notion of *correctness* here is a quite *general* one that applies even to performances of scores of music, or of dances, or of ways of folding a foldable bike, cases where there is no truth or warranted application at issue.

This is importantly right and one of the main reasons why *all* of these topics fall under the general heading of 'rule-following': they all have the following structure in common:

> A certain *standard* for doing something—playing a piece of music, dancing a dance, playing a game, using a word—which in and of itself may not be *rationally mandated*, is *adopted*. Once it is adopted, it serves as a standard for subsequent performance, pronouncing that performance as correct or incorrect.

Is Correctness Real Normativity?

A propos of this way of thinking about the slogan 'meaning is normative', many have complained that it construes 'normativity' too lightly (see, e.g., Wikforss 2001, Glüer and Wikforss, 2015).

The complaint is: 'correct' as it is being used here is not a genuinely normative term; hence the label 'normativity of meaning' is inappropriate and misleading. This is not real normativity and hence does not pose even a prima facie problem for naturalism's ability to account for meaning.

There are three questions here that I would like to look at briefly (too briefly, given the complexities involved):

1. What is it for a notion to be genuinely normative?
2. Does 'correct' count as genuinely normative?
3. How interesting is the previous question (2) for the purposes of evaluating the naturalistic program in the theory of meaning?

The critics tend to assume that for a notion to be genuinely normative, claims involving it should analytically entail subjective ought-claims. Furthermore, they point out, semantic

correctness claims do not imply subjective oughts. Hence, they conclude, semantic correctness is not genuinely normative.

I think that the critics are right that semantic correctness claims are not subjective ought-entailing (see my 2005). If I mean addition by '+' it would be *correct* for me to say '68 + 57 = 125'. But it doesn't immediately follow, without further normative assumptions, that if I am asked about that sum, I *ought* to say '125.' If I mean addition, and I want to tell the truth, I ought to say '125.' But if I want to mislead my audience, I ought to say something else, something incorrect. So, correctness is not necessarily ought-entailing or 'action-guiding.' (Notice that I am assuming here that if linguistic meaning were the source of norms, it would have to be the source of norms on assertion. This is something that, as we will see later on, Gibbard wants to deny.)

At any rate, my view, along with that of, among others, Rosen and Peter Railton, is that correctness claims count as normative even though they are not subjective ought-entailing. The notion of correctness is *evaluative*—it appeals to a standard—even if it is not prescriptive.[2]

The distinction between these two kinds of normativity shows up clearly when we say that *a false belief is incorrect*. That's not just a way of repeating the claim that the belief is false; it implies that something has gone wrong with that belief, that it is defective. And, yet, it could simultaneously be true that the thinker subjectively *ought* to have come to that belief given the (misleading) evidence available to her.

So, meaning claims are correctness-entailing. And semantic correctness is a kind of normativity, although it is not of the 'action-guiding' variety.

Some of this helps explain why I have been regarded by some as pro normativity of meaning, and by others as anti. It is partly about where you stand on the 'correctness as normative' issue.

Another potential confound is that while I do regard correctness as normative, I am also inclined to think that the question whether semantic correctness is genuine normativity is an issue that is not all that interesting in connection with the naturalistic program in the theory of meaning. I will explain why I think that later on in this essay.

Before doing that, however, I want to take a look at Gibbard's recent attempt to revive the claim that meaning is normative in the full-throated sense of being analytically subjective ought-entailing.

Gibbard on the Normativity of Meaning

Gibbard's overall view is, as you might expect, complicated, subtle, original, and impressive. I'm not entirely confident I understand all of its different parts, or how they fit together.

In this section, I am just going to look at one central argument that Gibbard provides for thinking that the concept of meaning is normative. Gibbard intends the argument to be preliminary. He takes the real test of his theory to be whether, when taken as a whole,

2 For a related distinction, see McPherson and Plunkett (2017).

it illuminates the vexed topic of meaning in a way that other theories have failed to do. So, what I will be assessing here is not his overall view but the preliminary support for it that he offers. A full assessment of his theory would be a daunting task.

Gibbard starts by distinguishing between a weak Normativity of Meaning thesis and a strong one. According to the *Weak Thesis* (22):

> Claims about meaning, all by themselves, without the help of further normative assumptions, analytically imply *ought* claims.

The *Strong Thesis* (22), by contrast, asserts not only the Weak Thesis, but also that:

> Meaning can be fully defined through some combination of normative and naturalistic concepts.

Gibbard accepts both theses. As a result, according to Gibbard:

> Every means implies an ought.

And

> For every means, there is an ought that implies it.

Notice that Gibbard is making these claims about the *concept* of meaning, not the *property* of meaning. It's one of the distinctive features of Gibbard's view that he works with a very strong concept/property distinction. While he insists that the concept MEANING is normative, he equally insists that the property of meaning, the property in virtue of which things fall under the concept MEANING, is as naturalistic as you please. Indeed, he believes that the idea of a normative *property* makes no sense. Only concepts can be normative or not. This explains why he doesn't fear that his normativity of meaning theses will have anti-naturalist consequences. By limiting the normativity theses to the *concept* of meaning he leaves it open that the property of meaning is constituted purely naturalistically.

I will raise a question about whether we have any good reason to believe even the Weak Thesis.

Gibbard's Argument

Gibbard offers the following argument in its support:

> A chief reason to believe the weak normativity thesis . . . is that a certain basic kind of ought follows from a means *invariably*. (16; emphasis added)

Gibbard explains that the 'ought' he is talking about here is Ewing's (1939) ought: an ideal, primitive ought that ignores costs and limitations on our powers of reasoning.

What's an example of the sort of entailment at issue? Gibbard says:

> With Ewing's primitive ought, we can say, an ought does follow from the meaning of 'nothing', and follows invariably. One ought not to believe both that snow is white and that nothing is white. This follows in a way not dependent on any normative principles it would make sense to doubt. This is explained if oughts are built into characterizing the very meaning of a word like 'nothing,' oughts that comprise the logic of the word. It is explained if the logic of the word 'nothing' is a matter of certain oughts that govern the beliefs couched with the word. (15)

Gibbard's point in emphasizing that the entailment is *invariable* is to rule out *prudential* explanations for why one ought not to believe contradictions. Prudentially, one might well have reason to believe both that snow is white and that nothing is white (suppose you know that an evil demon will blow up the world unless you believe both those things).

But Gibbard thinks that, even if this were so, it would still be true, using Ewing's ought, that one ought not to believe those two things together.

Let us grant, then, that it is an analytic truth that

> (Belief Nothing, BN) One ought not to believe both that snow is white and that nothing is white.

Still, even granting this, there is a prima facie problem with this way of arguing for the normative nature of the MEANING concept.

Let's agree that (BN) is true. How does anything follow about the meaning of *linguistic expressions*? (BN) doesn't involve the meaning of anything linguistic at all, it would seem. It just says that one ought not both *believe* that snow is white and that nothing is white. Perhaps when we talk about beliefs we are talking about *propositions* or *thoughts* (although even this is controversial); still nothing about language has entered into the picture, yet. So, how did we get from "one ought not . . . believe both that snow is white and that nothing is white" to "presumptively . . . a *means* seems to qualify as normative" (16)?

Let me come back to the issue of how we are going to get *language* to enter the picture.

Perhaps what we have here is, at least in the first instance, an argument that *mental content* is normative, rather than that linguistic meaning is. Should we at least agree that it follows analytically from

> p is the proposition that snow is white and nothing is white

that

> One ought not to believe p?

While this may seem more plausible, it is also plausible that whatever norm is in the vicinity here, it stems from the nature of *belief* rather than from the nature of mental content.[3]

We have already seen reason to maintain that belief is a normative concept, from the fact that for someone to properly understand the concept of belief they need to understand that

(Norm on Belief-1) A false belief is defective.

Another plausible norm on belief, stated let's say in terms of Ewing's ideal primitive ought, might be this:

(Norm on Belief-2) If X has undefeated evidence that p, then X ought to believe that p.[4]

Since, in a suitably broad sense of 'evidence,' I have undefeated evidence that it can't be the case both that snow is white and that nothing is white, I have an explanation for why (BN) holds that appeals only to norms on belief and not any norms on meaning.

Not only is there no *positive* reason to attribute the normativity here to mental content, rather than to belief, there is reason *against* doing so. After all, mental content features in a host of *other* propositional attitudes besides belief, desiring and hoping, for example, to which *different* norms apply, if any. Wouldn't it make more sense, then, to think of these different norms as deriving from the distinct attitudes that they govern, rather than from the mental contents that they have in common?

In any case, how are we going to get from these observations about *belief* to the normativity of *linguistic meaning* (here I am returning to the question I had earlier postponed). Referring to our earlier exchange, Gibbard says:

> In arguing against the normativity of content and in favor of the normativity of belief, Boghossian dismisses the normativity of linguistic meaning as something that would have to be put in terms of norms of assertion. My own attempt in this book goes by a different route, which Boghossian does not consider. (17, n32)

We saw above that there are no analytic entailments from a meaning claim to an ought claim for *assertion*. From

I mean plus by '+'

it follows that

3 I am drawing here on my 2005.

4 I borrow this formulation from Hill 2013.

It is *correct* for me to say '68 + 57 + 125'

but *not* that

I *ought* to *say* '68 + 57 = 125'

even if this were Ewing's ideal ought. At best we could claim that

I ought to *believe* that 68 + 57 = 125.

Now, Gibbard's idea is that we can convert this into a defense of the normativity of linguistic meaning by claiming that, in many of these cases,

> Talk of believing thoughts . . . amounts to talk of accepting sentences in the thinker's own language, along with what those sentences mean. (27)

Let's formulate this as the following view:

> (Belief as Sentence Acceptance, BSA): To believe that p amounts to accepting a sentence of one's own language that means that p.

If (BSA) were true for all beliefs, then we would have an answer to the objection I was pressing that the norms in this area stem from belief and not meaning or content. If belief just amounts to acceptance of a sentence of one's own language, along with what it means, then the norm that

If X has undefeated evidence that p, then X ought to believe that p.

just amounts to:

If X has undefeated evidence that p, and X's sentence S means that p, then X ought to accept S.

And this does seem to be an example where a meaning claim implies an ought claim.

The trouble, of course, is that (BSA) is not true for beliefs in general. There are prelinguistic beliefs (as in infants) and nonlinguistic beliefs (as in many animals).

Indeed, even in linguistic creatures like us, I doubt very much that most belief is linguistic in this sense. As I visually survey the scene in front of me, I am forming many new beliefs. But as I now survey this scene, I am certainly not busy accepting sentences of English. If I were asked, I could express most of those beliefs in English, but that is to give voice to them, not to explain what they consist in.

Gibbard says he takes this talk of belief as sentence acceptance over from Paul Horwich. But, for Horwich, belief is constituted by the acceptance of sentences in a 'language of thought,' not by the acceptance of sentences of a public language. However, even if one found talk of a 'language of thought' acceptable, it would do very little to make plausible a claim like (BSA). To say that belief takes place in a language of thought is just a metaphorical way of saying that the states that constitute belief have something analogous to a syntactic structure, something with parts that can be combined and recombined in certain ways. That is very far removed from supposing that belief consists in some sort of psychological relation to the independently identifiable sentences of one's public language.

In any event, as long as we have to concede that belief talk in general is not reducible to talk of the acceptance of sentences, we will have to recognize norms for belief that are not reducible to norms on linguistic meaning.

But then, once we have done that, we may as well see all the norms in this area as emanating from the norms on belief, just as we have previously said.

This is all by way of questioning the chief initial motivation that Gibbard gives for taking seriously the 'meaning is normative' slogan. For all I've shown, it might be that once we appreciate the virtues of his overall view, we will be willing to pay whatever price these remarks have identified. However, I'm not there yet.

Back to Correctness

Let's go back to meaning as setting a standard of correctness for the application of words. As I've said, I am inclined to regard this as a genuine type of normativity; but I don't attach the same importance to this claim that Kripke did.

For Kripke, the question whether the notion of meaning is normative seems to have been of paramount importance because he thought that, if it could be shown to be true, it would have an immediate and negative impact on the *naturalistic reducibility* of meaning.

Thus, after giving many specific criticisms of the dispositional theory, Kripke says that it should have been obvious from the start that a dispositional theory could not be right because the relation between meaning and use is normative whereas the relation between a disposition and its exercise is descriptive (37).

As this remark shows, Kripke is thinking of the normativity of meaning as something that would *immediately* show that no dispositional analysis of a meaning claim could be right, because it so obviously misconstrues the relation between meaning and use. There are two respects in which I disagree with Kripke here.

The first is that the most one might extract from the point is something about the a priori reducibility of meaning to dispositions, via some sort of conceptual analysis. And that would fall very short of showing that meaning is not ultimately grounded in naturalistic facts, in the way that Gibbard favors.

The second is that even if we allow that correctness is normative, I don't believe that that would immediately show that a dispositional analysis is ruled out because, as I explained in previous writings (Boghossian 1989), any dispositional theory would have to include some appeal to a set of conditions that are 'ideal' conditions under which our cognitive mechanisms are incapable of making mistakes. Our ordinary dispositions with respect to a given term are bound to include dispositions to apply that term to things to which it does not apply. Thus, reading off the meaning from our ordinary dispositions is bound to lead to incorrect verdicts about what our terms mean. Hence, any dispositional theory will have to appeal to the dispositions that we have under ideal conditions. However, if meaning is constituted by dispositions under ideal conditions, then it seems possible to capture whatever normativity there exists here in the difference between how a person would ideally respond as compared to how she actually responds.

Meaning's Normative Role

Let's go back to the crucial passage in which Kripke enunciates his 'meaning is normative' slogan and let me quote it more fully.

> Suppose I do mean addition by '+' What is the relation of this supposition to the question how I will respond to the problem '68 + 57'? The dispositionalist gives a *descriptive* account of this relation: if '+' meant addition, then I will answer '125'. But this is not the proper account of the relation, which is *normative*, not descriptive. The point is *not* that, if I meant addition by '+' I *will* answer '125', but that . . . I *should* answer '125' . . . The relation of meaning and intention to future action is *normative*, not *descriptive*.
>
> In the beginning of our discussion of the dispositional analysis, we suggested that it had a certain air of irrelevance with respect to a significant aspect of the sceptical problem—that the fact that the sceptic can maintain the hypothesis that I meant quus shows that I had no *justification* for answering '125' rather than '5,. How does the dispositional analysis even appear to touch this problem? Our conclusion in the previous paragraph shows that in some sense . . . we have returned full circle to our original intuition. Precisely the fact that our answer to the question of which function I meant is *justificatory* of my present response is ignored in the dispositional account and leads to all its difficulties. (37)

Notice that Kripke clearly regards talk of the normative nature of meaning as *equivalent* to saying that our grasp of the meaning of a word must be able to *justify* our use of that word.

There are two possible problems with this claimed equivalence.

First, and in general, to say of X that it is constitutively normative and to say that it has a justificatory role are two different things. Something can have a *normative role* without itself *being* constitutively normative. It can have that role as a matter of necessity, and yet not itself *be* normative.

To see a clear example of this, consider perception. Perception has a normative role: it can justify belief. Indeed, if perception has a normative role, it necessarily has that role.

But there is no intuitive sense in which perception is a normative concept or state. You don't need to understand anything normative (in the prescriptive sense) to understand the idea of something being perceptually presented to you as being the case.

Second, in the particular case of meaning, meaning is constitutively normative, because of its necessary connection to correctness conditions; but its *justificatory* role is a *further* aspect of its nature. It follows, therefore, that we might be left with a problem about meaning's justificatory role even *after* we have said all that needs saying about its relation to correctness conditions.

To explain. Suppose I mean *green* by 'green,' and, on seeing a green thing under good lighting conditions, I utter the sentence 'Here is a green thing.' In this case, let's suppose, I correctly and justifiably believe that the thing is green.

Now, when I voice my belief by uttering the sentence 'Here is a green thing,' what I say is not only correct, given what I mean by 'green,' but is also *justified* by my grasp of the meaning of 'green.'

The point is that language use is a *rational activity*.[5] When I use the word 'green' to express my belief that something is green, I have a *reason* for using that word as opposed to another, a reason that is provided by what I mean by the word.

So, our grasp of the meaning of an expression has to be able to justify our use of that expression (just like perception has to be able to justify our beliefs). Call this the *Justificatory Thesis*.

Notice how all these claims about meaning are entirely parallel to the intuitive way in which we would characterize an instance of rule-following:

> If S is following a rule R on a given occasion by doing A, then (1) S has *accepted* R, (2) S's acceptance of R determines whether what S did was *correct*, (3) S's acceptance of R *explains* why S does A, and (4) S's acceptance of R rationalizes or justifies her doing A.

This is why, intuitively, meaning something by a word of a public language looks like a special case of following a rule with respect to that word. They have identical structures.[6]

In both cases, it looks as though there is a time at which a meaning is grasped (a rule internalized or accepted); and that grasp then stands in the following three relations to subsequent behavior: it sets a *standard* of correctness for that behavior; it explains it; and it *justifies* it.

I think these three aspects of both rules and meaning deeply puzzled Wittgenstein. And I agree that they are indeed deeply puzzling.

---

5 A point I think of as emphasized by Christopher Peacocke among others.

6 For reasons explained in my 1989, meaning things by words of mentalese, as opposed to a public language, could not have this structure.

They are undoubtedly puzzling on a dispositional view. In particular, the dispositionalist will clearly have a problem with the justificatory requirement. For how does my being disposed to apply 'green' to green things *justify* me in applying it to this green thing? Just because I am disposed to do something doesn't imply that I am justified in doing it.

One might think that ideal conditions could be invoked here to help out the dispositionalist. Couldn't we say that the justification for what I am disposed to say here is grounded in the fact that this is what I would say if conditions were ideal?

The problem with this reply, though, is that the justification that we intuitively have for the rational use of language is some sort of internalist justification: it must be something that I am aware of and whose relevance to my action I can somehow see. That is why talk of being *guided* by one's understanding of the meaning of a word is so intuitively compelling.

But this would not be true for a justification grounded in some counterfactual about what I would do if conditions were ideal. I don't normally know what those conditions are and don't know how to compute what I would do if I were in them.

In any event, the problem here is not unique to the dispositionalist but may be thought to present a quite general problem for any account of meaning. The general problem is this:

There is one case of the understanding supplying a justification for our using a word in a certain way that we sort of understand reasonably well. This is when my grasp of the meaning of a sentence, S, consists in my grasp of some sort of explicit definition for S; and my basis for assenting to S consists in my inferring S from its definition (what I have elsewhere called 'Frege analyticity'). We could give such a story for why I assent to "All squares are four-sided figures."

But this is clearly a very special case—special both in that grasp rarely consists in grasping an *explicit definition*, and in that basis rarely means *inferential basis*.

But if the relation between grasp and assent is not like that, what else could it be like?

The cases of one thing justifying another noninferentially that we understand well enough are the cases where a perceptual state that p justifies our belief that p. A perceptual state can serve as an epistemic basis for assent because it is a *presentation* of the world as being a certain way, a *seeming*—that feature is essential to its ability to justify belief.

But how could the relation between grasp of the meaning of 'green' and the application of 'green' be analogous to the relation that obtains in the perceptual case?

It is hard to see what could count as an internalistically acceptable answer to this question, if not a story that appeals, in the way that philosophers and linguists have found quite natural to do, to the notion of an intuition or an intellectual seeming. Our grasp of the meaning of 'green' supplies us with an intuition to the effect that it correctly applies to the presented object.

Of course, to say this is to supply the merest hint of where an answer to our question should be sought. We are still some distance from knowing exactly how to fill this story out.

Conclusion

My current view, then, is that the really difficult problem in the theory of meaning concerns how my grasp of the meaning of a word could *justify* future applications of that word.

This has often been conflated with the issue whether meaning is constitutively normative.

But the Justificatory Thesis and the Normativity of Meaning thesis are distinct claims, each interesting in its own right.[7]

References

Boghossian, P. (1989). "The Rule-Following Considerations." *Mind* 98, no. 392: 507–49.

——— (2005). "Is Meaning Normative?" In C. Nimtz & A. Beckermann (Eds.), *Philosophie und/als Wissenschaft—Hauptvorträge und Kolloquiumsbeiträge zu GAP.5*. Paderborn: Mentis.

Ewing, A. C. (1939). "A Suggested Non-Naturalistic Analysis of Good." *Mind* 48, no. 189: 1–22.

Gibbard, A. (2012). *Meaning and Normativity*. Oxford: Oxford University Press.

Glüer, K., and Å. Wikforss (2015). "Meaning Normativism: Against the Simple Argument." *Organon* 22: 63–73.

Hill, C. S. (2013). "Review of Allan Gibbard, Meaning and Normativity." *Notre Dame Philosophical Reviews*.

Kripke, S. A. (1982). *Wittgenstein on Rules and Private Language: An Elementary Exposition*. Cambridge, MA: Harvard University Press.

McPherson, T., and D. Plunkett (2017). "The Nature and Explanatory Ambitions of Metaethics." In *The Routledge Handbook of Metaethics*, edited by T. McPherson and D. Plunkett, 1–28. London: Routledge.

Rosen, G. (1997). "Who Makes the Rules Around Here?" *Philosophy and Phenomenological Research* 57, no. 1: 163–71.

Wikforss, Åsa (2001). "Semantic Normativity." *Philosophical Studies* 102, no. 2: 203–26.

Wittgenstien, Ludwig (1953). *Philosophical Investigations*, G. E. M. Anscombe and R. Rhees (Eds.), G. E. M. Anscombe (trans.), Oxford: Blackwell.

7 I am grateful to the audience at the GibbardFest in Ann Arbor in spring 2016 and to Billy Dunaway, Paul Horwich, Josh Peterson, and David Plunkett for helpful comments.

18

OBLIGATIONS OF MEANING

Paul Horwich

1. Introduction

Is the word "meaning" a *normative term*, one (like "ought", "evil", "beautiful" and "delicious") with a prescriptive or evaluative function? Or is it a purely naturalistic term (like "electron", "big", "kill" and "red")? And, regarding the *phenomenon* of a given word possessing its particular meaning: is such a fact normative 'all the way down'? Or is it ultimately and entirely constituted by the word's naturalistic properties—including, perhaps, that the word tends to be used in conformity to a certain distinctive, naturalistically characterized regularity?

My discussion of these related questions will focus on Allan Gibbard's answers to them—answers he began to develop in the early 1980s and which he elaborates with characteristic insight, subtlety, and power in his *Meaning and Normativity*.[1] I'm afraid I won't be able to do justice here to that searching treatment of the topic. I'll have to confine myself to some comments on the following sequence of ideas, which I take to form his central line of thought.

1 Oxford University Press, 2012. All page references are to this book unless noted otherwise.

2. The Core of Gibbard's Position

G1. The word, "means", is normative.[2] This explains (and may therefore be inferred from) the fact that any explicit attribution of meaning—for example, "The German term, 'oder', means OR"—is normative.

G2. The clearest evidence of this is that meaning-attributions *analytically entail* OUGHT-propositions.[3] For example, the above attribution entails, in virtue of *its* meaning, that German speakers *ought* never to accept a certain sentence, "p", yet at the same time reject the result, "p oder q", of using the term to connect that sentence with another one, "q". This entailment manifests the normative character of the meaning-attribution (and hence of the word "means"). That's because, as Hume observed, "You can't get an *ought* from an *is*"—no application of a *non*-normative concept to a person, S, can analytically entail that S ought (or that S ought not) to do A—where "A" is uncontroversially non-normative.[4]

G3. A word's meaning what it does is *conceptually constituted* by use-norms. That is to say, for each word-meaning there's an OUGHT-proposition that not

---

2 If a *word* is normative, that's in virtue of what it means: the question of whether a given *term* is normative is equivalent to the question whether its *meaning* is normative. And since, like Gibbard, I identify *concepts* with *meanings*, I'd say that another equivalent question is whether the term expresses a normative *concept*.

Also like Gibbard, I'm using capitalized expressions as the names of the meanings of (= the concepts expressed by) the lowercase versions of those expressions. Thus, "OR" is toname whatever happens to be the meaning of the English word "or". Similarly, "MEANS" is to name the meaning of "means" (= the concept expressed by that term).

Presumably, we can define "the meaning of w" as "what w means" (or vice versa); so any considerations relevant to whether MEANS is a normative concept and "means" a normative term will also apply to MEANING and "meaning" (and vice versa).

3 Gibbard's reason for inserting the word "analytically" before "entails" in his Hume-inspired test for whether the proposition *that p is* a *normative* one is to make clear that the test won't be passed merely because there's an OUGHT-proposition that's true in all the *metaphysically* possible worlds in which the tested proposition is true (even when this necessitation is knowable *a priori*). What's needed (see the following sections 3–6 for detailed discussion) is that it be *nonsense*, that is, *unintelligible*, that is, *incoherent*, that is, *inconsistent with the meaning of "p"*, to believe that proposition while denying the OUGHT-proposition. (In what follows, "x entails y" will abbreviate "x *analytically* entails y").

4 In order to protect his Humean test from trivial counterexamples, Gibbard (p. 11) requires that the implied OUGHT-propositions not be "*degenerate*." For even a plainly *non*-normative proposition about a word will analytically entail, for example, *that the word ought to be used as it ought to be used*, and will also entail any disjunction of itself with a normative proposition.

It remains unclear (at least to me) how to devise, in light of such examples, a general necessary and sufficient conditions for being a 'non-degenerate' OUGHT-proposition. But, in the meanwhile, we can perhaps make do with the plausible *sufficient* condition deployed above: namely, that an OUGHT-proposition is non-degenerate if it takes the form, <S ought to A>, where A can be a matter of *not* doing things, and is characterized in uncontroversially non-normative terms.

only *follows* analytically from an attribution of that meaning (as in G2) but that—going in the opposite direction—analytically *dictates* the attribution.

G4. Although meaning-*attributions* are normative, the *properties* they attribute are not. For the properties that are designated by *normative* predicates (and concepts) are also designated—but at more fundamental explanatory levels—by *naturalistic* predicates (and concepts). Thus, it's *ultimately* in virtue of *non*-normative facts that "oder" means what it does.[5]

G5. More specifically, it's in virtue of how a word *tends* to be used – i.e. the characteristic basic dispositions regarding its use—that it *ought* to be used in certain distinctive ways and hence means what it does. For example, although the sound of the German "oder" is also used in England (where it means SMELL), its *norms* of use are quite different in the two places. And that's because the sound-type's *actual* use is different in the two places. A plausible candidate for the general dependency relation of a word's 'obligations of use'[6] on its 'actual use' is that *person S ought to use w in conformity with regularity R(w) because the fundamental fact governing S's actual overall use of w is S's disposition to conform with R(w).*[7]

---

5 This formulation of G4 accords with Gibbard's preference for deploying a *coarse*-grained sense of "property," whereby conceptually nonequivalent predicates at distinct explanatory levels (e.g., "water" and "H_2O") may designate one and the same property. In that terminology, the second of the issues mentioned at the outset (which I, preferring a *fine*-grained notion of property, expressed as the question of whether the meaning-properties of words are ultimately constituted by naturalistic properties) becomes the question of whether the properties designated by predicates of the form, "w means such-and-such", are designated at all more basic explanatory levels by naturalistic predicates. (Some reasons for my own terminological preference, together with the suggestion that it doesn't much matter, are sketched in footnote 15 below).

6 I use this expression (here and in what follows) for terminological convenience. The rather awkward, "ought of use", would be more accurate because it wouldn't suggest, as "obligation" does, that it's for the sake of *others* that one must meet the requirements at issue.

7 The word "meaning" has several meanings (even when the topic is confined to language). It—and the correlated transitive verb—can legitimately be used to specify *reference* (as in, "By 'Big Ben' the British mean the clock above their Parliament building"). Also, "meaning" can be used for the *propositional constituent* that a particular token of some word expresses (e.g. when I say, "I'm hungry", I don't *mean*, in this sense, what *you* would mean by "I"; but nor—for Fregean reasons—do I always mean, "PH is hungry"). In addition, it can be used for the *pragmatic intention* of the speaker in selecting that word (e.g., the Australian use of "Pom" as a mildly derogatory term for English people).

But I take Gibbard's claim *that "meaning" is a normative term* to be focused on a further notion of meaning—one that can plausibly be regarded as *fundamental* in relation to the others (i.e., to be part of the basis for our understanding them). This is the notion someone is deploying when she says, for example, "'Wasser' in Kurt's language has the same meaning that 'water' has in mine, but not the same meaning as my 'H_2O'"; or "In virtue of the meaning of 'nothing' in Jane's language, she ought not accept both 'Snow is white' and 'Nothing is white'"; or "Billy has just learned the English meaning of the word 'exaggeration.'" The present inquiry into whether or not "meaning" is normative and whether or not *meaning*-properties have *normative* underlying natures will be restricted to this notion of meaning: *the semantic meaning of an expression-type in a language*.

G6. Adoption of a normative conception of meaning is motivated, not only by G2's 'means-entails-ought' thesis, but also by the fact that such a conception will enable us to escape the many implausible attributions of *meaning-indeterminacy* that inevitably issue from fully naturalistic theories.

To avoid getting lost in the many details to follow, it's worth keeping in mind the main resemblances and differences between Gibbard's norm-laden picture and a certain fully naturalistic alternative: namely, that (i) "means" is a wholly *naturalistic* term, and (ii) *S's meaning what she does by a word is synthetically reducible to S's naturalistic tendency to use it in conformity with a certain regularity*.

Despite Gibbard's endorsement of point (ii)—and even his inclination to agree with the naturalist on exactly what, for any given word, its distinctive meaning-constituting use-tendency will be—we can see in G1–G6 that he diverges from *thoroughgoing* naturalism in three crucial respects, which result in his denial of (i). First, he holds that the reduction of a meaning-property to a use-*tendency* is mediated by its *initial* reduction to a use-*norm*:—*S's meaning what she does by w* is explained to begin with as *S's basic obligation to use w in conformity with a certain regularity*. Second, this initial reduction is held to be a matter of *conceptual analysis*. Thus, "S's word w means ODER"—insofar as it's *analyzable* as "S is *obliged* to use w in conformity with such-and-such regularity"—is held to be normative, contrary to (i). And third, he takes the grounding of such use-norms in use-tendencies to be *non-analytic*, but nonetheless *a priori*; so that the two-stage constitutive relation between use-tendencies and meaning-properties is a discoverable *a priori*. That contrasts with what one might well think is the most promising *naturalistic* view of the matter—which is that this relationship (like that between H_2O and water) can only be discovered by means of an *empirical* investigation.

3. What Is It for a Concept to be Normative?

Let's begin a more careful examination of Gibbard's position by focusing on what he intends in describing a concept (or meaning, or associated linguistic expression) as "normative". As I indicated at the very outset, the rough idea is that the terms so described have a function similar to that of words such as "should", "justified", "vicious" and "cool": we use them to express 'value judgments'. And this characterization does succeed in pointing us in the right direction. But we'll need a criterion of 'being normative' whose satisfaction is easier to

---

Assuming the fundamentality of that sense in relation to the others, if *it* is normative then the others are very likely to be too. In contrast, if that fundamental sense is *not* normative it remains quite possible that some of the further notions of meaning constructed in terms of it might nonetheless be normative. For one of the additional concepts needed in the construction of some of those further notions may be normative. (I won't have space in this essay to investigate that possibility.)

objectively establish if we are to be in a position to determine, for any given term (such as "meaning"), whether it does or does not belong in this category.

Gibbard focuses on two tempting proposals. Both of them depend on the view that "ought" (in its *subjective* sense) is the fundamental normative term and that all other normative expressions owe their normativity to their intimate conceptual relationship to the subjective "ought."[8]

One way to implement this idea—a natural initial thought—is to say that a term is normative just in case it has an accurate *definition* that invokes "ought". Thus "x is good" will qualify as normative, by this first proposed criterion, if its meaning is supplied by "x *ought* to be desired". Similarly, "the meaning of w" will be normative if it's definable as "the way in which w *ought* to be used".

But, since very few terms are explicitly definable, this notion of "the normative terms" is an undesirably strict and narrow one. Many words that we may intuitively feel should be categorized with "ought" (e.g., "delicious", "brave" and "cool") are likely to be excluded—perhaps even "good", since the just-mentioned possible definition of it may well be incorrect.

For this reason, Gibbard relies, as we've seen, on a second and less demanding general test for normativity—one that's based on Hume 'no-*ought*-from-an-*is*' principle. This second criterion is that a term will qualify as *normative* if and only if simple ascriptions of it analytically entail propositions either of the form, <x ought to be done>, or of the form, <x ought not to be done>. Thus "courageous" will pass this test if <Act x was courageous> entails <x *ought* to be admired>—even if the word isn't *definable* in terms of "ought". And "meaning" will be normative (in at least the more relaxed sense) as long as Gibbard is right that meaning-attributions entail OUGHT-propositions—for example, that <S's word "#" means SOMETHING> entails <S *ought* not accept her *name*-predicate sentence, "k is f", while rejecting the *quantifier*-predicate sentence, "# is f">.[9] (As indicated in G3, he holds that attributions of word-meaning, *unlike most other normative propositions*, not only *entail*

8 Gibbard (2012, 14–15) attributes to A. C. Ewing ("A Suggested Non-Naturalistic Analysis of Good," *Mind* 48 (1939): 1–22) the idea that all normative concepts owe their normativity to their relationship to the subjective OUGHT.

This concept differs in various interconnected respects from the so-called *objective* OUGHT First, that a person ought *objectively* to do a given thing typically depends on facts of which she is unaware—indeed, she may even have reason to deny those facts. (E.g., "I ought to have bet on that horse," said merely as an expression of regret for not having done what would have turned out to be best). Second, the *subjective* ought is epistemologically more fundamental: conclusions about what ought *objectively* to be done (or believed) rest on prior conclusions to the effect that certain things ought *subjectively* to be done (or believed). Third (and more specifically), we might with some plausibility follow Gibbard ("Truth and Correct Belief," *Philosophical Issues* 15, 1, 338–50, reprinted in his 2012) in defining "what someone ought *objectively* to do" as "what she ought *subjectively* to do when in possession of all relevant facts." And fourth, someone can reasonably be *criticized* for having failed to do what he ought *subjectively* to have done, but not so easily for having failed to do what he ought *objectively* to have done. (In what follows, when I speak of "the concept OUGHT," I'll mean the *subjective* concept.)

9 "<p>" is shorthand for "the proposition that p".

non-degenerate OUGHT-propositions, but are each analytically *equivalent* to some conjunction of such propositions).

Thus, what Gibbard takes to show that meaning-attributions are normative is evidently *not* the weak uncontroversial fact that there are true *material* conditionals whose antecedents are such attributions and whose consequents are normative. Nor is it even the fact that some such material conditionals are *necessarily* true and knowable *a priori*. For substantive norms (e.g., "Torture is wrong") can typically be reformulated as generalized conditionals whose antecedents characterize one or another kind of phenomenon (or situation, or action) in *non*-normative terms and whose consequents proceed to evaluate these phenomena with the use of normative expressions (e.g. "[x][x is an instance of torture → x is wrong]"). Moreover, *fundamental* substantive norms are typically both necessary and *a priori*. So if Gibbard's normativity test isn't going be absurdly liberal—if it's going to exclude terms that are obviously non-normative—then it must require that the relevant MEANS-TO-OUGHT conditionals are not merely true and necessary and *a priori* but are *analytic entailments*, that is, conditionals *it wouldn't make sense to doubt*, that is, conditionals whose acceptance is mandated by what they mean.[10]

Of course, there are yet more senses that *can* be given to "w is normative"—senses that aren't at all close to the one that Gibbard is trying to capture, because they won't amount, even roughly speaking, to the claim that w is a word whose function is at least partly evaluative or prescriptive.

---

10 A *third* indication of a term's being normative, which Gibbard mentions (pp.10-11), is that the concept it expresses be *philosophically puzzling* in just the way that concepts such as OUGHT and GOOD are. This approach might eventually deliver a test that would be deeper and more general than the both of those discussed above—a test that would perhaps supply the basis for *any* concept, including OUGHT itself, to qualify as normative. But the idea, as it now stands, is too vague to enable us to settle whether MEANING is normative. Something sharper is needed.

A *fourth* test (one proposed by Paul Boghossian in "The Normativity of Content," *Philosophical Issues* 13, no. 1 (2003): 31–45) is that "content attributions are normative just in case . . . it's a condition on understanding a content attribution that one understand that if it is true then [some] ought claim is also true." This, as Boghossian intends, is more restrictive that Gibbard's "analytic entailment" criterion (which allows that meanings can be correctly attributed to the words of someone who *can't* grasp the OUGHT-propositions that the meaning-attributions entail).

However, we might feel that Gibbard's requirement that OUGHT-propositions be analytically entailed was *already too restrictive*. So a further test suggests itself, in which that requirement is dropped, but something very like Boghossian's additional requirement is retained. According to this *fifth* test, a term qualifies as normative if a person cannot fully grasp its meaning unless she has a prior understanding of "ought" (presumably, because a condition of that full grasp is that her use of the term bear a certain specific relation to her use of "ought"). In one respect, this is more demanding than Gibbard's test, which doesn't require that a normative term can be fully understood only if "ought" is already understood. But in another respect its test is *less* demanding, since it doesn't require that simple ascriptions of a normative term must analytically entail OUGHT-propositions.

It's tempting to wonder which of these alternative tests for normativity is the "right" one—the one that accurately specifies the scope of that concept. However, there may very well be no such thing. We should perhaps recognize a variety of legitimate normativity notions corresponding to the variety of criteria just surveyed.

In particular, one *might* decide to say that a certain term is "normative", having in mind merely that the phenomenon it stands for has some normative *import*—i.e. that the designated phenomenon has normative properties. Thus, despite its purely naturalistic definition, "torture" will be "normative" *in this very different and currently irrelevant sense*. Similarly, the word "umbrella" will be "normative" simply because it's *good* to have an umbrella in the rain. And it can't plausibly be denied that "meaning" is "normative" according to that criterion. After all, no competent English speaker would deny the principle, "Given what our word 'dog' means, we *ought* make some effort to avoid applying it to cats."[11]

4. Has Gibbard Made It Plausible That "Meaning" Is Normative?

Granting his favored criterion, we are led to the question of whether conditionals of the sort Gibbard uses to illustrate his position really are cases in which meaning-attributions *analytically entail* subjective OUGHT-propositions. Is it plausible, for instance, that ascriptions of "S's word, w, means TRUE" *entail* propositions of roughly the following form: <S ought not reject her sentence, u, while accepting the sentence whose subject is "the proposition expressed by u" and whose predicate is w>?[12]

---

11 My thanks to David Plunkett for stressing to me that further questions should be raised about the conception of "ought" that's operative in Gibbard's theory: (1) Is it the notion deployed in *moral* claims, or else the sense of "ought" that's deployed in speaking of *epistemological justification*, or else the OUGHT of *decision* theory, or of *etiquette*, or of some other domain? (2) Are there circumstances in which a meaning-constituting use-obligation will *clash* with an obligation of some other kind—and, if so, which takes precedence? And (3) are meaning-constituting use-obligations *compulsory* and *authoritative* (as moral norms appear to be), or—on the contrary—can a person somehow opt out of them and be legitimately unmoved by them, as one arguably can in the case of *legal* norms (if what they really state are the punitive consequences that are likely to issue from acting in certain specified ways), and in the case of the norms of a specific game, such as chess (which may be nothing more than the player-regularities that define the game)?

I'm afraid that there's no space here for a decent discussion of Gibbard's take on these important issues. But my inclination is to think: (1*) that he rejects the view that we deploy different senses (or 'flavors') of subjective "ought", and supposes, rather, that different act-types and state-types (such as beliefs, desires, practical choices, feelings of shame, etc.) are governed by distinctive norms that are all couched in terms of a *single* notion of OUGHT. One such domain is that of *sentence-acceptance*, and here only *epistemological* considerations are relevant. So (2*) there can be no conflict between our basic norms of word-use and, for example, our moral obligations. And (3*) that the norms of word-use that hold within the community of L-speakers are robustly authoritative. Anyone interacting verbally with those people is genuinely obliged to obey these norms.

For further discussion, see D. Plunkett and T. McPherson, "The Nature and Explanatory Ambitions of Metaethics." In *The Routledge Handbook of Metaethics* (2017), edited by T. McPherson and D. Plunkett, 1–28. New York: Routledge.

12 What about the objection that S might on some occasion have certain strong *moral* obligations, or be subject to powerful *pragmatic* pressures, dictating that, despite what she means by "true," she ought not on that occasion use it as prescribed by Gibbard's conditional? As indicated in footnote 11 above, I take his reply (see pp. 12–16) to be that *sentence-acceptance* is governed by a special-purpose set of subjective-ought norms—"epistemological norms"—and that neither pragmatic nor moral considerations can be relevant

An initial misgiving about this claim arises when we replace "S's word w" with "*Our word 'true'*". For "Our word 'true' means TRUE" is completely uninformative. Given the capitalization convention for naming concepts (see footnote 2), it reports merely that our word "true" expresses the concept it in fact does; and this surely can't convey anything nontrivial about what that concept is and how it should or shouldn't be deployed. And the same goes for any term. To put the point another way, for us English speakers it goes without saying that our words "dog", "bachelor" and "true" mean (respectively) DOG, BACHELOR, and TRUE; but how those words ought to be used may remain debatable matters.[13]

Moreover, a similar doubt concerns attributions of meaning to *foreign* terms. Granted, it *is* informative to say, "Pierre's word 'vrai' means TRUE".—We are asserting that Pierre's "vrai" and our "true" voice the same concept. But we convey *nothing* about the *identifying nature* of that common concept—about how, in detail, it tends to be deployed or ought to be deployed. So knowledge of such detailed matters of usage can't simply be extracted from the meaning-attribution, but would call for an investigation based on additional information. This is not to *deny* that

(T) If S's w means TRUE, then S ought accept instances of "w(<p>) ↔ p".

It's merely to suggest that the correctness of that conditional doesn't derive from its being an *analytic entailment*.

5. An Easier Route to the Normativity of "Meaning"?

The concerns just aired were focused on what Gibbard takes to be the distinctive, *detailed*, normative implications of different meaning attributions: for example, the implications, roughly spelled out above, of meaning OR, or SOMETHING, or TRUE—the particular

to their authority, since those other considerations concern, not what we ought *to accept* but (respectively) what we ought *to be ashamed* for accepting and what we ought *to hope* to accept.

And I take it that he'd give a similar response to the further objection: that nothing of the form, <*S means F by w*> *entails* <*S ought to conform with regularity R(w)*> could be right, since its consequent will be false when, although S *does* happen to mean F by w, she *ought* not to. Gibbard would say (I think) that this is a mistake. For whether S should, or should not, mean F by a given word is not an epistemological matter but a pragmatic matter. And the pragmatic undesirability of S's meaning a certain thing by w (arising perhaps because of the confusion that would result from w's already having a further meaning) has no bearing on the fact that certain epistemic obligations are incurred by S's nonetheless giving w that meaning. Similarly, even if it was unwise for S to make a certain promise, she's nonetheless obliged to keep it.

My thanks to Hannah Ginsborg for pressing me on this pair of issues.

13 I am assuming here that the words "dog", "bachelor" and "true" are *understood*. Granted, someone with no understanding of (say) "tachyon" might nonetheless come to know, merely via the convention for naming meanings, that "tachyon" means TACHYON. But such a person would be in an *even worse* position to say how that term is used or ought to be used.

regularities of use to which a speaker *ought* to conform in her use of the words that express these concepts. But perhaps such misgivings can be mitigated, or even eliminated, by observing that those MEANS-to-OUGHT conditionals all depend on a much simpler and more immediately compelling one: roughly speaking that

(Q) If words w and v have the same meaning then they ought to be used in the same way.[14]

This functions as follows in explanations of Gibbard's detailed conditionals:

w means F → w means the same as our "f" *[convention for naming meanings]*
w means the same as our "f" → w ought to be used in the same way as our "f" ought to be used *[instance of (Q)]*
∴ w means F → w ought to be used in the same way as our "f" ought to be used *[transitivity]*
Our "f" ought to be used in accord with regularity R *[established independently]*
∴ w means F → w ought to be used in accord with R

So we might well suspect that a simpler and less controversial route to Gibbard's normativity-of-"meaning" conclusion—avoiding the concerns raised in section 4—might instead be based solely on principle (Q).

But, on reflection, such hopes can't be sustained. Granted, (Q) is more obviously correct than either the principle (T) concerning "true" or any of the other detailed MEANS-to-OUGHT conditionals that Gibbard offers us. And granted, it's harder than it is in those detailed cases to see how (Q) could coherently be denied. So, we might well suspect that it really is an analytic entailment.

Still, there's a special problem with relying on it to reveal the normativity of "meaning". The problem is that even if (Q) *is* an analytic entailment, it exhibits a form of *degeneracy* (see footnote 4 above) in light of which we can see that it *doesn't* show its antecedent to be normative.

Consider, for example, "If Cicero was Tully, then Cicero should have done whatever Tully should have done". That's plausibly analytic. But it obviously doesn't reveal the sentence, "Cicero was Tully" to be normative, because the conditional is much better explained as a mere instance of Leibniz' Law—$(x)(y)(\psi)[x=y \rightarrow \psi(x)\leftrightarrow\psi(y)]$. Similarly, if the concept BACHELOR is the same as the concept UNMARRIED MAN, then it follows analytically (given Leibniz' Law) that S ought to *blah-blah* a bachelor ↔ S ought to *blah-blah* an unmarried man (no matter what's substituted for "*blah-blah*"). And, the

14 To be more precise: the consequent, "S ought to use w and v in the same way", means, in the present context, "S ought not to accept a sentence containing one of those words whilst rejecting the sentence that results from replacing it with the other".

antecedent here is analytically equivalent to the supposition that "bachelor" means the same as "unmarried man".

So, we're back to square one. The case for "meaning" being normative must after all rest (as Gibbard proposes) on our being able to make it cogent and plausible (along the lines illustrated above for "true", "or" and "something") that, for each particular word w, there's a distinctive acceptance-regularity, R(w), such that w's meaning what it does analytically requires that it ought to be used in accordance with R(w).

Yet there remain the considerations of section 4, to the effect that this couldn't possibly be so——that there simply isn't enough substance in a meaning-attribution (especially if it takes the form, "My word "f" means F") to preclude perfectly coherent differences of opinion about which particular tendencies or obligations of use characterize the concept that's attributed.

6. The Nature of Incoherence

In response to such concerns, it might well be protested that they stem from wrongly deploying a conception of *incoherence* (i.e. *unintelligibility*, i.e. *nonsense*)—and hence of *analytic entailment*—that's simply not the one Gibbard has in mind. We should appreciate that, as he's using the term, for it to be *incoherent* to doubt "p→q" (hence, for this conditional to be an *analytic entailment*), it suffices that such a doubt would violate one or another requirement (identifiable via the methodology of conceptual analysis) for "p→q" to mean what it does. But since the needed conceptual investigation might turn out to be extremely difficult, and its results highly contentious, someone oblivious of the right answers may well qualify as "coherent", "intelligible" and "making sense", in some *ordinary* sense of these terms, even though she's in fact violating conditions on her "p→q" meaning what it does. That's what's illustrated in section 4's misgivings. But given Gibbard's more demanding sense of those terms, such violations, however forgivable, nevertheless qualify as incoherent, unintelligible nonsense.

Consider, for example, certain philosophers in the 1960s who, on first hearing Gettier's alleged counterexamples to the "justified true belief" analysis of "knowledge," were inclined to stick with the traditional view that S would *know* a certain thing even if she were in the special circumstances to which Gettier drew out attention. These philosophers doubted what are in fact correct (indeed *analytic*) cases of "If S is in C(p), then S doesn't know *that p*". Their attitude, although mistaken, may nonetheless be described in ordinary language as "making sense"—but not in Gibbard's perfectly valid alternative terminology.

So far so good. But Gibbard appears to be suggesting that the very methodology that philosophers have deployed in attempts to analyze "knowledge", "causation", "truth", "free choice" and so forth (and in their attempts to *criticize* proposed analyses), may also be used to identify analytically necessary conditions for our words to mean what they do. And he seems to be contending that, for each word, one such condition will turn out to be an obligation to use it in a distinctive way.

However, although it may be granted that conceptual analysis can take us from the meaning attribution

w means F

down to

The meaning of w = the meaning of our word, "f,"

it isn't prima facie plausible that mere conceptual analysis would suffice to show that "*The meaning of* ___" does **not** reduce to "*The basic acceptance-tendency governing* ___" but reduces rather to "*The basic acceptance-obligation governing* ___"—which would boil the above meaning-attribution down to

The basic acceptance-*obligation* governing w = the basic acceptance-*obligation* governing our "f"

And it's even less plausible that *conceptual analysis* is the appropriate method for identifying our precise basic acceptance-obligation regarding word the "f,"—knowledge of which would enable us to get down to a specific instance of

The basic acceptance-obligation governing w = the basic obligation to conform to regularity, R(w)

It's hard to think of any case in which the as-yet-unidentified item that satisfies a given definite description is finally identified by *conceptual analysis*![15]

7. A New Approach

Thus, Gibbard's claims about what are "analytically entailed" by meaning-attributions stand in need of further explication and defense. In order to provide a neutral setting in which this *might* be done, I propose to make a fresh start.

Our ultimate aims will still be to determine: (1) what *kind* of underlying property of a word engenders its meaning, (2) how we can proceed to identify what *particular* property of that kind engenders the meaning of a particular word, and (3) whether this "engenderer" might emerge from *conceptual analysis*. But we first need to devote some attention to the meta-issue of how such general questions are to be approached.

15 Unless the description happens to take the form, "the conceptual analysis of expression, e."

To begin with, I'll address a couple of considerations that might seem to stand in the way of ever being able to answer them. This pair of skeptical concerns will call for a brief methodological discussion of how the issues *can* in fact be settled. And with the conclusions of that discussion in mind, I'll compare and contrast two specific theories as to what the right answers are: Gibbard's norm-infused account of meaning-properties and my own predominantly *non*-normative proposal. Despite that glaring difference between them, these theories are strikingly parallel to one another. So, placing them side by side will enable us to see, with peculiar sharpness, whether normativity should be brought into the picture—and, if so, where and how.

8. A Couple of Pseudo-Problems

A first apparent obstacle to providing a satisfactory account of how facts of the form, ***w means x***, are grounded is that we aren't able to say—not even in *superficial common-sense terms*—what sort of 'things' their ingredient meaning-entities, **x**, are: that is, what *sort* of object satisfies, for example, <x = the meaning of "dog"> (i.e. <x = the concept expressed by "dog">). We can't even say whether it would have to be a sort of *physical* object, or a sort of *mental* object, or a sort of *abstract* object—let alone, within one of those broad categories, what *specific sort* or *kind* of entity it is, and let alone which particular instance of that kind is the meaning in question. Granted, one can *describe* a meaning as, say, "the meaning of my word 'dog'"—just as one can describe a color as "the color of my kitchen wall" or a number as "the number of Jupiter's moons", But whereas particular colors and numbers are easily designated *directly* (as, e.g. "green" and "sixty-seven"), deploying modes of identification that don't depend on any contentious scientific or philosophical assumptions, in the case of *meanings* no such direct uncontroversial articulations are available.[16] Claims to the effect that meanings *boil down* to one sort of thing or another—for example, mental images, or functions from possible worlds to sets of objects, or rules, or bunches of epistemic obligations, or use-dispositions—would appear to be controversial *reductive theories*. And one might well wonder how any of them could be justified if we can't even identify the superficial phenomena they allegedly reduce.

On reflection, however, there is no genuine puzzle here, but simply confusion. This can be appreciated by remembering that, although superficial predications of meaning (using, e.g. "w means DOG") don't tell us anything substantive about which meaning (i.e. concept) is ascribed, we do nonetheless have a rich sense of the empirical import and explanatory power of the meaning-properties they designate. For example, the fact that my word "dog"

16 Pretty clearly, this problem is not mitigated by the introduction and deployment of our special terminology for meanings—our stipulation that an expression in capital letters (e.g., "DOG") is to name the meaning of the lowercase word ("dog"). For the meaning-predicates we can then formulate (such as "w means DOG") simply rearticulate the above-mentioned merely *descriptive* and *indirect* characterizations of meanings (e.g., "The meaning of w = the actual meaning of our word, 'dog'").

means what it does—i.e. that means DOG—helps account for my overall use of it, including my acceptance of certain sentences containing "dog" and my rejection of others. So—just as in the case of the properties invoked within physics, chemistry, and other areas—we may reasonably expect to have information about *meaning*-properties (of the form, ***w means F***) that will be sufficient to enable us to discover their underlying natures. And—as will be discussed in section 17 below—the reductive ground of any such relational property can be expected to contain a factor that constitutes (or simply *is*) the concept F.[17]

A second anxiety about our project might derive from the sense that it's afflicted with epistemological circularity. For, on the one hand, it would seem that we won't be able to say what it is about an arbitrary word, w, that engenders its *meaning* until we have settled on the meaning of the word "meaning" itself. But, on the other hand, how can we hope to settle that prior question independently of a plausible *general* view of how the meanings of expressions are engendered?

This 'difficulty' is also an illusion. It's simply not true that in order to take a reasonable view of how to arrive at the grounds of the meaning-property of *any* given word, we would first need to have figured out how the specific term "meaning" comes to mean what it does. This is not to deny that we must already *understand* that term. We must (as we would ordinarily put it) "know what it means". But we *do* understand it; we already have a variety of more or less strong convictions about meaning that we articulate with the word's help. And a reasonable general view as to what should replace "#" in a theory of the form, <w means what it does in virtue of the property of w that satisfies the condition that #(w)>, can be arrived at by considering which choice of "#" would best accommodate these pretheoretical convictions. Once that selection of "#" has been made, we can then apply the

17 I'm continuing to deploy a *fine-grained* notion of *property*, whereby necessarily co-extensional predicates with different meanings (e.g. "water" and "liquid collection of H_2O molecules") stand for *different* properties. This enables me to respect the common and useful way of speaking whereby an asymmetric relation of *constitution* (or *reduction*) can be said to hold between distinct properties. (So I can say, "Water is constituted by H_2O, not the other way round.") From this point of view, it's natural to suppose that a given property will qualify as *normative* if and only if any predicate that expresses the property includes (nonredundantly) a normative term. And we can then ask whether or not a normative property might be *constituted* by (i.e., reducible to) a nonnormative property.

As we've seen, an alternative way of articulating these matters—long advocated by Gibbard—is to sharply distinguish predicative *meanings* (i.e., predicative *concepts*) from the *properties* they designate, and to suppose that there can easily be different concepts of one and the same property. For example, "water" and "H_2O" are said to express different concepts but to designate a single (*coarse-grained*) property. Similarly, Gibbard would say that the property designated by a normative meaning-attribution-concept is also designated at a more fundamental conceptual/explanatory level by a naturalistic concept.

I suspect that the general issue here is not substantive but merely terminological. After all, the word "property" as used by philosophers is a refined, technical term; and we shouldn't be surprised to find that more than one reasonable refinement is deployed. Still, some notations can be more perspicuous than others.

resulting general theory to the particular word "meaning", thereby discovering constitutor of *its* meaning-property.[18]

9. Constraints on a Theory of Meaning

Thus, the right methodology for finding out how meaning-properties are constituted is not hard to discern. It's exactly the same as used elsewhere when the goal is reductive analysis. We begin with uncontroversial pre-theoretical views about the phenomenon at issue. We can regard our conjectures about its constitution as plausible to the extent that they cohere with those initial ideas. And, at the same time, we must appreciate that these ideas are not sacrosanct. We should leave open the possibility that, for the sake of a sufficiently simple reductive analysis, we might well put up with slight revisions in some of our original opinions.[19]

But what are the pre-theoretical views we have about meaning that constitute the real epistemic basis of our enquiry? For most of us, they will include many of the following claims:

- Two words (e.g., "Hesperus" and "Phosphorus") may refer to the same thing (or be true of the same collection of things) yet not have the same meaning as one another. (FREGE)
- What speaker S *personally* means by his word, w (i.e. w's meaning in S's *idiolect*) enters into explanations of S's acceptance (both internal and overt) of sentences containing w. (EXPLANATORY POWER)
- The members of a linguistic community will typically vary with respect to one another as to how fully they grasp the public *communal* meaning of a given word—that is, they

18 Gibbard mentions a similar threat of circularity: "What does the word 'means' mean? We can't say what this question means until we have its answer, until we know what its last word means" (7). But again the response is that we *do* know what its last word, "means," means; and so we know what the whole question means. We understand these expressions. Of course, being able to *say* what they mean (in a nontrivial way) is a further matter. But this problem is far from insoluble. We certainly wouldn't need to possess its solution before we could understand and address the problem. Our understanding of meaning-attributions, and hence of questions as to how they are grounded, puts us in a position to assess alternative general proposals about which property of a given word constitutes its meaning; and the best such proposal can then be applied to the word "meaning" itself.

19 I'm not suggesting that the pre-theoretical constraints on a good account of something could include some opinion as to what the correct account is. After all, different people will begin with different hunches about what the correct account will turn out to be. So, none of these could be regarded as 'widely shared pre-theoretical intuitions'. The task is to discover which of them is right.

Moreover, I'm not denying that an ordinary language term may be used in a scientific theory, but with a substantially different (technical) meaning. In such a case, it obviously wouldn't be a desideratum that we respect the common-sense convictions articulated with the help of the word in its ordinary sense.

will vary in how similar what they *personally* mean by it is to what it means *in the communal language (e.g., English, French, Arabic, . . .)*.[20] (COMMUNAL LANGUAGE)

- The degree to which member, S, of a linguistic community understands a word, w, is normally assessed by observation of the degree to which the w-sentences that S accepts match those that the majority (or most of the relevant experts) within S's community would accept in similar circumstances. (EPISTEMOLOGY)
- Meaning has various kinds of *normative import*, relating to (1) the value of true belief, (2) epistemic justification, (3) good communication, and (4) rules of the language. More specifically:
    1. If w's meaning within a certain linguistic community is the same as the meaning of our (English) predicate "f", then its members *ought* to want *that they apply w to x only if x is f*.
    2. The correct norms of belief, in conjunction with what S means by her sentences, yield the circumstances in which S *justifiably* accepts those sentences.
    3. It's *desirable* (for the sake of smooth communication) that the *personal* meaning given to a word by a member of a linguistic community coincides with the word's public meaning. (And if it doesn't exactly coincide, then the closer it comes to coinciding the better.)[21]
    4. What's meant by a word, w, in a language, L, is associated with a rule dictating how, in that language, the word *ought* to be used.[22] (NORMATIVE IMPLICATIONS)

---

20 The communal meaning of everyday English words (like "Mars", "blue" and "true") may be identified with the idiolectal meanings of them that are roughly shared among most members of the community—and, in the case of technical terms (such as "arthritis", "felon" and "carburetor"), with the idiolectal meanings that are given to them by most of the relevant *experts*—that is, those in the community to whom the rest of us tend to defer (e.g., the doctors, lawyers, and auto-mechanics). For detailed discussion of the relationship between a word's communal meaning and its personal meaning, see my "Languages and Idiolects." In *Language and Reality from a Naturalistic Perspective—Themes from Michael Devitt*, edited by A. Bianchi, Springer, 2018.

21 I don't mean to deny that someone might sometimes have good reasons (perhaps moral, perhaps epistemological, perhaps pragmatic) to knowingly use a word in a way that deeply diverges from its current communal tendency of use (with the hope perhaps of bringing about a change in that word's communal use).

22 Point (4) is akin to a tempting constraint on theories of meaning that's articulated by Saul Kripke (in his *Wittgenstein on Rules and Private Language*, Cambridge, MA: Harvard University Press, 1982, 7–14): namely, that a person's way of understanding a given word must incorporate *directions* that guide and justify her use of it. Now, despite its superficial plausibility, such a constraint on a theory of *understanding* appears to be viciously regressive (since the alleged 'directions' would already need to be understood, as would the further directions that would constitute that prior understanding, and so on). Indeed, Kripke himself rejects it on those grounds. His own solution to his 'skeptical paradox of meaning' involves coming to recognize that this particular guidance/justification requirement is misconceived (see p. 87). However, there remains a sense in which Kripke's 'guidance constraint' might be perfectly right and important:

Thus, our goal is to discover which particular criterion of the general form, "w means what it does in virtue of the property of w that satisfies condition, #(w)," is the one that optimally coheres with this collection of principles. In other words, which choice of "#" provides the criterion that best coheres with our pre-theoretical view of meaning.

Calling this pre-theory "T*{M(w)}" (where "M(w)" is short for "the meaning-property of w"), we may conclude that if there's a unique condition, #(w), such that T*{#(w)}, then any word's meaning-property is engendered by whichever property of the word satisfies that condition.

As for questions of whether such a method of investigation is outright *empirical*, or is purely *a priori*, or might even qualify as *conceptual analysis*, I'm postponing discussion of those matters until section 16 below.

10. The Case against *Normatively* Grounded Meaning-Properties

With the just-sketched methodology in mind, let's compare a *prima facie* plausible purely naturalistic proposal for how meaning-properties are constituted with a *prima facie* plausible normative proposal:—on the one hand, the view that our word w means what it does in virtue of some fact of the form, <Our use of w *tends* to conform with regularity R(w)>; and, on the other hand, that it's in virtue of some fact of the form, <Our use of w *ought* to conform with R(w)>. For example, the naturalist might claim that "true" means what it does to us because w *tend* to accept instances of "<p> is true ≡ p"; and the normativist might claim instead that it's because we *ought* to accept such instances.[23]

---

namely, that we are *implicitly* guided by the rules of our language (including, those rules whose followings constitute our meaning what we do). For further discussion, see section 12 below—"Rule following".

23 Here are some further examples of naturalistic hypotheses that have some plausibility:

- Our word "red" means RED in virtue of the fact that we tend to accept "That's *red*" in response to red light.
- Our word "ought" means OUGHT in virtue of the fact that we tend to accept "I *ought* to do A" only when we feel some inclination to do A.
- Our word "not" means NOT in virtue of the fact that we tend to accept "*not* p" if and only if we reject "p".
- Our word "neutrino" means NEUTRINO in virtue of the fact that we tend to accept "If there's a single kind of thing of which '$T^{\$}(\_)$' is true, then things of that kind are neutrinos" (where "$T^{\$}$ (neutrino)" is the standard formulation of neutrino theory).

Note that these formulations, which invoke "tendencies," could have instead alluded to "propensities" or "dispositions," or "*ceteris paribus* laws." Although, for expository convenience, I've often used "tends", that term can easily be read as in a way that's much too weak. So I think the stronger notion of "*ceteris paribus* law" or "nomological requirement" is more accurate. As Christine Korsgaard once pointed out to me, someone doesn't mean what we do by "bachelor" just because he is *very often* prepared to interchange that word with "unmarried man"! (For further discussion, see "Regularities, Rules, Meanings, and Epistemic Norms", chapter 7 of my *Truth-Meaning-Reality*, Oxford: Oxford University Press, 2010).

To obtain the corresponding normatized proposals, simply replace "tend" (or "are nomologically required") with "ought".

On the face of it, the singular virtue of supposing that meaning-properties are constituted *naturalistically* is that this is the view of them that best squares with our central pre-theoretical convictions about what they are supposed to explain and about how their attribution is justified. For, on the face of it, the view of them as constituted *naturalistically* is the only way for these constraints to be accommodated. As we have seen, it's hard to deny (a) that when we accept and utter a sentence, that is explained in part by what we mean by it (and therefore by the meanings of the *words* in it); and (b) that we discover (via inference to the best explanation) that what others means by a certain term of theirs is just what we mean by a given word of our own, by observing that they tend to use theirs in the same basic way that we tend to use ours.

Thus, we should expect the meaning-property of our word w to be the common factor in all of the thousands of explanations of the thousands of particular events that consist in one or another sentence containing w being accepted. And the conjecture that such properties reduce to *tendencies* (i.e., ceteris paribus *laws*) enables us to see how this might be so, and one can thereby understand why the epistemological bases for meaning-attributions are what they are. However, on the face of it, a *normative* property of the word won't do, since—on the face of it—how a word *ought* to be used cannot help to explain how it actually *is* used.[24,25]

11. A Normativist Response

But from Gibbard's perspective this case in favor of taking a fully naturalistic view of meaning-properties is not as strong one might think.

First, he recognizes that how a word-sound *ought* to be used within a linguistic community (which may of course typically differ from how that same word-sound ought to be used somewhere else) must be explained by something about how the word is *actually* used in that community. For example, he suggests that it's because in England we *actually* have

24 Advocates of Davidsonian truth-conditional semantics, or of Montagovian possible-world semantics, may find it obvious from the start that meanings are clearly *not* typically normative. For the referents (taken to be meanings) of words like "Mars", "red" and "or" are obviously not normative. And functions from possible worlds to objects (or to sets of objects) are obviously not normative either. But these correct observations are beside the present point. For our present question is not whether the *things meant* by words are constituted normatively but whether a word's *meaning what is does* (i.e., its meaning-*property*) is constituted normatively. (For further discussion of this distinction, see section 17 below). From the perspective of truth-theoretic semantics, this becomes the question of whether properties like '*w refers to Mars*' and '*w means the function taking each possible world into the set of things in that world that are red*' are constituted normatively. And the answer to *these* questions is *not*, "*obviously* no."

25 The following sections will be concerned exclusively with how the meaning-properties of *words* are constituted. As for *complex* expressions, including sentences, there won't be space here to address the contentious matter of how their meanings depend on the meanings of the words they contain. (This issue is addressed in "Deflating Compositionality," chapter 8 of my *Reflections on Meaning*, Oxford: Oxford University Press, 2005). But, presumably, a necessary and sufficient condition for the meaning-properties of complexes to be grounded normatively is that the meaning-properties of words be so grounded.

a *tendency* to infer "S is a bachelor" from "S is an unmarried man", and vice versa, that we *ought* to sanction *all* such inferences.

Second, he thinks of this explanatory relation as one of *grounding*, or *constitution*—that is, as akin to the relations of property-constitution discovered in science (e.g., that a gas having a certain temperature is a matter of its molecules having a certain mean kinetic energy).

And on the basis of this pair of assumptions he's in a position to argue that insofar as a word's basic *tendency* of use helps explain its overall usage then the constituted *norm of use*—which in turn constitutes the word's meaning-property—also explains that usage. So, he thinks his theory *is* able, after all, to accommodate the pre-theoretically recognized explanatory and epistemological relations between meaning and actual use.

However, Gibbard's second premise—his assumption that a word's actual use not only *explains* how it ought to be used but *constitutes* that normative characteristic—might well strike one as *ad hoc* and implausible.

After all, that assumption is made *simply* to ward off a decisive objection to the normativist proposal: namely, that it would fail to accommodate the fact that our actual uses of words are caused in part by how we understand them, by what they mean.[26] Gibbard's idea is that since the naturalistic *constitutors* of a word's use-norms can provide the needed causal explanations, we can also suppose that *these norms themselves* are providing them. But notice that no explanatory relation weaker than *constitution* would suffice for this. For, just because the naturalistic source of the norms governing a word's use *explains* the word's overall use it doesn't follow that *the norms themselves* explain that use. From the fact that X explains both Y and Z, it obviously doesn't generally follow that Y explains Z. In order for that to follow, it would have to be supposed that the explanatory relation between X and Y is *constitutive*. For example, since being H_2O constitutes being water, and since we can see through H_2O, we can say that we can see through the liquid in the glass because it's water. Similarly, Gibbard needs to suppose that his meaning-giving oughts are *constituted* naturalistically if he is going to maintain that the overall use facts can be explained in terms of meaning.

---

26 In Gibbard's defense, it might be protested that his second premise isn't at all *ad hoc*, since it derives from a *general naturalism*—from the metaphysical view that *nothing* exists beyond the vast network of objects and properties that bear causal and spatiotemporal relations to one another.

But this general doctrine merely compounds the need for *ad hoc* maneuvers. Insofar as languages are instruments that have evolved to cater to interests of ours that extend well beyond the desire to predict and explain natural phenomena, we should not expect that everything we find it useful to talk about—including numbers, possibilities, logical relations, and values—will be locatable within the naturalistic network. And so we shouldn't be surprised at the implausible contortions into which hardline naturalists are invariably driven in their efforts to defend their doctrine. These include (a) blatantly *ad hoc* analyses (e.g., "Numbers are numerals"); (b) *ad hoc* skepticisms ("There are no moral truths"); and (c) *ad hoc* decisions about what a *naturalistic* reductive base can include (e.g., "Grounding by *modal* facts is naturalistic, but grounding by *normative* facts isn't"). For further discussion, see my "Naturalism, Deflationism, and the Relative Priority of Language and Metaphysics" in *Expressivism, Pragmatism and Representationalism*, edited by H. Price, Cambridge, Cambridge University Press, 2013, 112–27.

But *ad hoc* assumptions (which are made merely to save pet doctrines from falsification) tend to be implausible, for they are made will little concern for their own credibility. And this one is no exception. On the face of it, the fundamental normative conditionals that specify the OUGHT implications of non-normatively characterized states of affairs (e.g., "Killing is *prima facie* wrong") are *not* expressions of constitution. They are *not* akin to *gas temperature* is constituted by *energy of the molecules*.

This isn't merely a matter of intuition. It's shown by the very different ways in which the two kinds of explanatory claim are established. We conclude that a given molecular energy *constitutes* a given temperature, not only because of the constant correlation between these things but crucially because we are able to explain the symptoms of a gas having that temperature in terms of the associated molecular energy. Similarly, a given brain state might be shown to *constitute* pain, but only if it could be shown that this brain state would give rise to the familiar behavioral consequences of being in pain.

But no such supporting arguments are expected from someone claiming that pleasure is the sole and fundamental source of goodness, or that an act that maximizes expected utility ought to be performed, or that if a word tends to be used in a certain way then it ought to be used in that way.

If these doubts are right, then Gibbard's normative view of meaning won't be able to accommodate the explanatory relation between a word's meaning-property and its overall use. And I can't see how any other view in which meaning-properties are constituted normatively could fare any better on this score.

12. Rule Following

However, there's a third common idea for how meaning-properties are constituted that's worthy of serious consideration—namely, that they are a matter of *following rules of use*. This seems to hover in between the view of meaning-properties as *tendencies* or *laws* of use and the view of them as *norms* of use. For the statement, <S follows the rule, *'Conform to regularity R!'*> doesn't *analytically entail* <S *ought* to conform with R>[27]—so this proposal can't be assimilated to Gibbard's. But the rule-following attribution *does* analytically entail something like <S *thinks*—i.e. *takes it*—that she ought to conform with R>; and that does still smack of normativity.

Of course, the claim cannot be (on pain of circularity) that meaning is a matter of *explicit* rule following, or that the entailed normative thoughts are *explicitly* entertained. The rule following (and whatever normative thinking is involved) must be *implicit* (see footnote 22 above). It must be that speakers act, in certain respects, *as though* they were explicitly following rules and having explicit normative thoughts. And so the suspicion arises that,

27 I'm assuming that although it's true that S's wanting to conform with R(w) provides *reason* for him to do so, this truth is not analytic.

once we spell out exactly what the phenomena are in which the rule-following and normative thoughts are implicit, the impression of having here a form of *normative* meaning-constitution will evaporate.

This suspicion is confirmed by consideration of various specific forms of the proposal at issue. I take Wittgenstein's account of a person's implicitly following rule, R! to be that it's a matter of her conforming to the *ceteris paribus* law-like regularity, R!, plus her tendency to occasionally react with dissatisfaction and correction to what she has just done.[28] Robert Brandom's account (roughly speaking) is that it's matter of conformity with such a regularity when that is sustained by *communal* normative attitudes that are manifested in reinforcement and correction by members of the linguistic community.[29] And Hannah Ginsborg has suggested that implicitly following a rule of word-use is constituted by a *conjunctive* disposition:—to be disposed both to apply the word in accordance with a certain regularity and to take those applications to be correct/appropriate.[30]

My own view is that, to a first approximation, we can suppose the naturalistic ground of a meaning-property is merely a certain explanatorily fundamental use-regularity. But in order to respect the compelling idea that languages, like games, are constituted by their rules, it's better, I think, to embrace the idea that meaning is constituted by *rule following*—in which case the more basic ground of a word's meaning-property will involve, not merely the use-regularity, but also the word's use being accompanied by feelings of satisfaction or dissatisfaction, or by self-correction behavior, or by some other naturalistic candidate for what constitutes a person's implicitly supposing that his applications of the word are appropriate (and sometimes, afterward, that they were not). In virtue of such motivating attitudes,

28 Note Wittgenstein's *Philosophical Investigations*, section 54. "Or a rule is employed neither in the teaching nor of the game itself; nor is it set down in a list of rules. One learns the game by watching how others play. But we say that it is played according to such-and-such rules because an observer can read these rules off from the practice of the game—like a natural law governing the play. But how does the observer distinguish in this case between players' mistakes and correct play? There are characteristic signs of it in the players' behavior. Think of the behavior characteristic of correcting a slip of the tongue. It would be possible to recognize that someone was doing so even without knowing his language." (For further discussion, see my *Wittgenstein's Metaphilosophy*, chapter 2.)

29 See R. Brandom's *Making It Explicit*, Cambridge: Harvard University Press, 1984, 18–66.

30 See Ginsborg's "Primitive Normativity and Skepticism about Rules," *Journal of Philosophy* 108, no. 5 (2011): 227–54). She doesn't specify the nature of these "normative takings." But, for the reason just given, she's surely not regarding them as *explicit* commitments. So, what account of them might be supplied? They are supposed to be as similar as they can be to *explicit* commitments—but without *actually* being explicit. It's *as if* they were explicit—although they aren't really. Thus, S behaves on each occasion of predication *as though* he were *explicitly* taking his predication to be correct/appropriate. What is this behavior? Quite plausibly, it's that he doesn't correct himself, or express subsequent dissatisfaction with his predication in some other way. One might put this by saying that *S's taking his predication to be appropriate* is *implicit* in these facts. Without textual evidence I can't confidently pin this account on Ginsborg. But nor do I see any other way of elaborating her position.

anyone whose rule is to conform to a certain regularity will thereby have good pragmatic reason (i.e., ought *pro tem*) to conform.

13. The Normative Constraints on Meaning-Constitution

It was argued in sections 10 and 11 that the view of meaning-properties as purely normative cannot accommodate the strong sense we have that the meaning of any word helps explain the circumstances in which different sentences containing it are accepted. But we mustn't forget that our pre-theoretical convictions about meaning-properties also include some that concern their distinctive forms of *normative* import (see section 9 above). And it remains to be seen whether a purely naturalistic constitution theory can do justice to those constraints.

Consider, to begin with, the conditional,

(CORRECT) If S's foreign sentence u means *that p*, then S's acceptance of u is correct iff p

A plausible view of this principle is that it's merely an extension of the *home*-language, non-analytic, substantive norm,

(CORRECT*) Our acceptance of our sentence, "p", is correct iff p

which makes no mention of meaning. The extended norm—equally synthetic—adds that foreign sentences are correctly accepted whenever, according to (CORRECT*), their translations into our language would be correctly accepted. And the naturalist view is that such translations derive from the matching of words whose uses are explained by the same fundamental tendencies of use.[31]

Similarly, conditionals of the form,

(EP-JUST) If a foreign-language sentence, u, means that p, then its speakers are epistemologically justified in accepting u in circumstances C

deploys "means" merely in order to extend the substantive home-language norm

(EP-JUST*) We are epistemologically justified in accepting our sentence "p" in C*

31 By analogy, consider an extension of the moral norm, "*People* should be treated with respect." to "*Creatures with mental states like those of people* should be treated with respect." This implies that X HAS MENTAL STATES LIKE THOSE OF Y has *normative import* (in the sense discussed in section 3 above)—but doesn't imply (or provide the slightest reason to think) that it's itself a normative concept.

where C* is a combination of external facts and internal evidential commitments, and C combines the same facts with translations of our evidential commitments into the foreign language.

And the same goes for those epistemic norms of acceptance that issue immediately from the meanings of words. We have,

(SEM-OBLIG) If L-speakers mean F by w, then they ought to use w in conformity with regularity R(w)

Such a principle deploys the concept, MEANING merely in order to extend the *home*-language norm

(SEM-OBLIG*) We ought to use our "f" in accordance with R("f")

For if w translates into our "f" then the basic rules for their use are the same. So—given section 12's analysis of rule-following—the speakers' implicit desires regarding their use will be the same. Therefore, their w and our "f" *should* (for the sake of those desires, and *pro tem*) be used alike.[32]

Thus, each of the various ways in which word-meaning has normative import can perfectly well be explained in terms of: (i) a home-language normative principle that has nothing to do with meaning, and (ii) the supposition that whatever norms govern a given expression of ours will also govern any accurate translation of that expression.

Moreover, these explanations put no constraint on the nature of *correct translation* (i.e., on the nature of *e having the same meaning as f*). In particular, they aren't in the slightest undermined by the assumption that meaning-properties—hence the relation of *sameness of meaning*—are constituted naturalistically.

So, we can conclude that a purely naturalistic account of meaning-facts can accommodate *all* the intuitive constraints on an adequate account of meaning—including the need to explain its normative import.

But what about the Gibbardian insistence that at least *some* of the principles that articulate the normative implications of a word's meaning are *analytic entailments*. Faced with the above explanations of those conditionals, a Gibbardian might bite the bullet and maintain

32 In addition to the forms of normative import possessed by meaning that have just been examined, there are the norms of 'good *idiolectal* meaning' that derive from the value of accurate communication:

> (GOOD MEANING) If S's w *communally*-means F, then it's desirable for the meaning of w in S's idiolect to be 'close' to F.
>
> A naturalistic account of this might be that (i) a word's communal meaning consists in the use-rule followed by nearly all members of the community, and (ii) valuable communication is fostered when each member's use-rule for a given word is similar to the rule for it that's followed by the other members.

that the *home*-language normative principles were *already* trivially analytic. However, this is clearly a desperate move. In the first place, it's very hard to identify which terms within, for example,

> You really *shouldn't* have accepted "Lying is wrong" whilst rejecting "It's true that lying is wrong".

are those whose meanings could conceivably be responsible for making this statement analytic. And, in the second place, it's pretty obvious that the extensions to foreign languages of home language principles such as that one preserve their real *normative force*. They are deployed in criticism. But that force wouldn't exist if they were analytic entailments.

Consider, by analogy, "Women are treated unjustly". If the word "woman" is defined in purely non-normative terms (e.g. via biological and/or sociological features), then this statement has normative force. It conveys disapproval of how the people with those features are treated. But if the word is defined in part as "treated unjustly"—so that <x is a woman> *analytically entails* <x is treated unjustly>—the statement becomes an empty tautology, devoid of normative force.[33]

14. Gupta's Problem

Here's a general objection to the idea that meaning-properties are constituted in one of the naturalistic or normative or regulative ways we've been entertaining so far. Consider, the idea that

> (T) "true" means what it does in virtue of our tendency to accept the instances of "<p> is true ≡ p"

Anil Gupta has shown that this cannot be quite correct.[34] For, in general, if w's meaning is constituted by its having property, U, then any word with that property must mean the same w. Therefore, an implication of (T)'s being correct is that

> *For <u>any</u> word "f", if we tend to accept the instances of "<p> is f ≡ p", then "f" will have the same meaning-property as "true".*

---

33 See Sally Haslanger's "Gender and Race. (What) Are They? (What) Do We Want Them to Be?" (*Nous* 34, no. 1 (2000): 31–55) for a proposed redefinition of "woman" along these lines. Note that I am not arguing here that this move couldn't somehow have beneficial social consequences. My point is merely that it would strip the sentence, "Women are treated unjustly", of its previous normative force.

34 This objection was put to me by Gupta in October 1992 and appears in note 17 of his "Deflationism, the Problem of Representation, and Horwich's Use Theory of Meaning", *Philosophy and Phenomenological Research* 67 (Nov. 2003): 654–66.

But this implied generalization is clearly false. To see this, consider the term, "tanr", introduced to mean "true and not red". Since it's obvious to us that no proposition can be red, we will tend to accept the instances of "<p> is tanr ≡ p" (inferring them from the truth-schema and the definition of "tanr"). Nonetheless, "tanr" doesn't mean the same as "true". So, the initial idea—(T)—about the meaning of "true" can't be right. And this problem arises across the board. No word can mean what it does merely in virtue of the tendency to use it in accord with some distinctive regularity.

Can we accommodate Gupta's objection while preserving the idea that the meaning-property of a word derives *somehow* from its use? Yes, we can. We'll need to suppose the relevant 'use' is *not* simply that the word tends to be applied in accord with such-and-such regularity, but that, in addition, this use-tendency is *basic*; it's what accounts for the word's *overall* use; it's the common core of explanations of all other acceptances of w-sentences. Once our initial formulation of the naturalistic proposal has been modified in this way, we'll no longer get the incorrect result that "true" and "tanr" must have the same meaning. That's because, whereas our commitment to the truth-schema *is* plausibly the explanatorily basic fact about our use of "true", the explanatorily basic fact about our use of "tanr" is clearly *not* our endorsement of the parallel schema. For *that* endorsement wouldn't explain why we accept, "Nothing tanr is red". What *would* explain it is the stipulation with which the word was introduced—namely, "Something is tanr just in case it's true and not red." And, as for our endorsement of "<p> is tanr ≡ p", that's explained by combining the fundamentality of the introducing stipulation with the fundamentality of our endorsement of the truth-schema. Or to put it more superficially, it's explained by our understanding of "true" and our understanding of "tanr".

Unsurprisingly, these considerations bear equally on the *normativist* view according to which the meaning-property of a word derives from some regularity in its use that *ought* to be displayed. For whichever such norm might be regarded as the one that engenders the meaning of w, we'll be able to find another word, v, to which that very norm also applies, but which doesn't have the same meaning as w. However, the obvious concessive response is to modify the initial normativist proposal in just the way that we modified the parallel naturalist proposal. The improved normatized account of meaning-properties will be that their grounds take the form: *w's explanatory basic norm of use is that one ought to conform with regularity R(w)*.

Similarly, the rule-theoretic proposal cannot be that our word w means F simply in virtue of our following the rule, <Conform with acceptance-regularity R(w)!>. The engendering fact must instead be this is the *basic* rule we follow in our acceptance practice with w.

15. Which Are the Relevant Regularities?

What is the specific regularity, R(w), that, for a given word, w, figures—in one way or another—in each of these three candidates for how w's meaning property is engendered?

Gibbard endorses something close to the well-known Russell/Carnap/Ramsey/Lewis approach to defining terms. Their idea was that we are to take the basic theory, <Tf>, of any phenomenon, f-ness, to be a conjunction of two independent commitments: first, a substantive one, <(∃!ϕ)Tϕ>, to the effect that there exists a single phenomenon with the characteristics specified by <T_>; and second, a purely linguistic component—an *a priori* and underived acceptance of the conditional, "(∃!ϕ)(Tϕ) → Tf", amounting to a conditional decision that, if such a unique phenomenon really does exist, then it's to be called "f".

Gibbard sympathizes with this approach (which, as it stands, is naturalistic) but introduces a normative element (pp.114). He allows that the meaning-fact may indeed *eventually* be constituted by our basic *disposition* to accept that conditional. But he maintains that the *immediate* constitutor of the term's meaning what it does is *not* that disposition, but is rather the fact that we **ought** to accept the conditional and to do so *fundamentally* (i.e., no matter what evidence we might acquire or have acquired, and no matter what we might intelligibly suppose, consistent with the meanings of terms in "T_").

And the *rule*-theoretic version will clearly be that "f"'s meaning is engendered by the fact that the *basic* rule we follow for its use is to accept the conditional, "(∃!ϕ)(Tϕ) → Tf".

16. Are Analyses of Meaning-Properties *a Priori*, or Are They *a Posteriori*?

We have seen that the meaning-constituting properties of words can be discovered only via the following sequence of assumptions and steps:

1. *The meaning-properties of words are engendered by underlying properties of the sort that can best accommodate our pre-theoretic convictions about meaning.*
   This is an instance of a general view, presumably a priori, as to how to identify the grounds of properties.
2. *The conjunction of these general pre-theoretical convictions is formulated by* "T* (meaning-property of w)"—which includes the items listed in section 9.
   This conjunction is arrived at by introspection, which is typically (but not invariably)categorized as a form of a posteriori access.
3. *The constituting property that best coheres with T*(meaning-property of w) is w's basic acceptance* ***tendency*** *(alternatively: w's basic acceptance* ***obligation****, or else w's basic acceptance* ***rule****). Thus, any particular meaning-property consists—for some specific acceptance-regularity R(w)—in the basic tendency to conform to R(w) (alternatively, the basic obligation to conform to R(w), or the basic following of the rule dictating conformity to R(w)).*
   This looks a priori.
4. *The regularity that figures in each of these three alternatives is, <Acceptance of the conditional whose consequent formulates our fundamental theory of the phenomenon designated by w, and whose antecedent is the Ramsey sentence of*

that formulation (obtained by existentially quantifying into the position of w)>. So the alternative constitutors of w's meaning are either the basic tendency to conform to that regularity, or the basic obligation to conform, or the basic rule dictating conformity.

This also looks a priori.

5. *Our fundamental theory of the particular phenomenon, f-ness, is formulated* by "T(f)".

   This identification requires introspective access to what our beliefs about f-ness are. But in addition, in order to determine which of them are epistemically *fundamental* (and hence which, if any, are *a posteriori*), we must bring to bear assumptions of various kinds—assumptions needed in our derivations of *some* of those beliefs—that will enable us to tell which of the beliefs are *not* fundamental. These assumptions are likely to concern (i) observed facts, (ii) principles of cognition, and (iii) the meaning-constituting properties of other words. And any assumptions of type (i) and (ii) would be a posteriori—in which case, so must be the assumptions of type (iii).[35]

6. *Thus "f"s meaning what it does is constituted by our basic tendency (or obligation, or rule) to accept "$(\exists!\phi)(T\phi) \rightarrow Tf$."*

   Given the dependence of this thesis on the previous steps, we can conclude that its epistemic status—as *a posteriori* or *a priori*—is not completely clear. But the former looks like the better bet.[36]

---

35 It's worth emphasizing that the naturalist's problem of identifying those ways in which "f" *tends* to be used that will help explain all acceptances of sentences containing "f" *is exactly the same as* Gibbard's problem of identifying those ways that "f"' *ought* to be used that will help explain all obligations to accept sentences containing "f." This shows:

1. that his method for identifying the meaning-constituting obligations of word-use is not really *conceptual analysis*—not the method deployed by philosophers in attempts to define TRUTH, KNOWLEDGE, CAUSE, FREE-WILL, and so forth.
2. that we therefore have little reason to think—even if **w's meaning F** does consist in **the basic obligation to use w in conformity to R**—that these predicates of w express the same concept.
3. that Gibbbard's test for a concept's being *normative* can cohere with his overall view only if an "analytic entailment" is explicated as a "conditional such that our obligation to accept it is dictated by the normative constitutor of its meaning" (recognizable as such by its explanatory fundamentality of that norm).
4. that since the acceptance-regularity, R, that, according to the naturalist, enters into the naturalistic ground of w's meaning *is exactly the same as* the regularity that, according to Gibbard, enters into the normative ground of the word's meaning, and since the method by which both parties discover that regularity is essentially the same, then if the best naturalistic method is *a posteriori*, we'd have to conclude that Gibbard's normative 'analyses' of meaning-properties must also be *a posteriori*.

36 In earlier expositions of the naturalistic perspective on meaning, I have taken the view that its hypotheses about the specific grounds of different meaning-properties are *a posteriori* and part of *empirical* linguistics.

17. The Two-Factor Problem

A word's meaning what it does *appears* to have *three* components: (i) the word itself (a mere sound, or mark, or gesture, or mental symbol), (ii) the thing that's meant (a meaning-*entity*, i.e., a *concept*), and (iii) the *relation* of meaning that holds between (i) and (ii). And *if* that's right—*if* meaning-facts are indeed relational in that way—then the problem of how they are constituted goes hand in hand with two further problems: What is the meaning-*relation*? And what sort of thing is a *concept*?[37] We might well expect to be able to factor any proposed constitutor of a word's entire meaning-*property* into a part that supposedly constitutes the meaning-*relation* and a part that supposedly constitutes the meaning-*entity* (i.e., the concept).

Granted, these first impressions may prove to be incorrect. The superficial *relational* form of meaning-attributions may turn out *not* to be preserved in their underlying logical form, or in yet more fundamental articulations of the facts they are used to state. But it's reasonable for our initial working hypothesis to be that the appearances are *not* misleading, and to abandon that assumption only in light of evidence to the contrary.

And in that case we must confront what I'm calling "the two-factor problem". Given any particular proposal about the way some specific meaning-property *as a whole* is constituted, how should that constituting property be factored into a part that is (or constitutes) the *concept* (—this will of course vary from one meaning-property to another—), and the remaining part, which constitutes the meaning-*relation* (and which will occur within the grounds of *all* word-meaning-properties)? How, for instance, should we divide up the property, ***w's basic use-explainer is the tendency for its use to conform to regularity R***?

One conceivable answer to this last question is to identify the *concept* with that *entire* property and to suppose that the meaning-*relation*, ***w means m***, reduces to ***w exemplifies m***.

(See, e.g., "A New Framework for Semantics", *Philosophical Perspectives* 22, no. 1 (2009): 233–40; and "Semantics: What's Truth Got to Do with It?", chapter 8 of *Truth-Meaning-Reality*, Oxford: Oxford University Press, 2010). Gibbard, in contrast, has claimed, and continues to claim, that his normatized versions of these hypotheses are *a priori*. I have recently come to an overdue recognition of the attractions of his position. But, for the reasons given above, I'm still not convinced. Perhaps, this tension can be attributed to the fact that the data in this domain (whether they be naturalistic or normative) are often difficult to place definitively on one side or the other of an intuitive line between the *a priori* and the *a posteriori*. For the most salient naturalistic data concern our own mental states—more specifically, *what we are inclined to say*, or *what we are inclined to feel*, or *what we are planning to do*—which we know by introspection. But facts known by introspection are notoriously hard to categorize. Indeed, as argued by Timothy Williamson, this difficulty may well be indicative of the superficiality and theoretical unimportance of the distinction between *a priori* and *a posteriori* judgment. See his "How Deep Is the Distinction between A Priori and A Posteriori Knowledge?" in ***The A Priori in Philosophy***, edited by A. Casullo and J. C. Thurow, Oxford: Oxford University Press, 2013.

37 I'm taking it that the nature of a word-sound is relatively unproblematic.

An alternative (at the opposite extreme) would be to identify the meaning-*entity*, **m** (the concept), with the regularity, ***R***, and to suppose that the meaning-*relation* reduces to ***w's basic use-explainer is the tendency for its use to conform to m***.

A third possibility would be to reduce ***w means m*** to the relation, ***w's basic use-explainer is m***—so the meaning-entity would be **the tendency to conform to R**.

And there might well be further possibilities. But how are we to decide among them? (And the same question would arise were we to begin with the assumption that meaning-properties as a whole reduce to analogous properties, but of either a *normative form or a rule-following* form).

A reasonable-looking initial idea for how to settle these matters is to assume that the largest *common* part of every total constitutor of every meaning-property (no matter which word exemplifies it) should be taken to ground the meaning-*relation*; and that the remaining part of each such constitutor—the part that varies from one to another—should be taken to be (or to ground) the concept meant. But then we'd have to embrace the *second* of the above options: we'd have to say that the part of ***w's basic use-explainer is the tendency for its use to conform to regularity R*** that constitutes ***w means m*** is ***w's basic use-explainer is the tendency for its use to conform to m***, and the part that constitutes a specific concept is ***the regularity R***. But this result is counterintuitive. It would be peculiar, for example, to identify the concept DOG with a word-regularity to which our use of "dog" does not conform (but only *approximately* conforms).

So, we could move all the way in the other direction, embracing the first of the above options and supposing that the ***means***-*relation* reduces to the *minimal* common element within the grounds of entire meaning-properties—which is the relation of ***exemplification***. But surely that can't be right either. After all, each word exemplifies innumerable properties. And obviously they can't *all* be identified with concepts that are meant by the word.

We are left (pending the suggestion of further options) with our second proposal—a middle ground that, as far as I can now see, is unobjectionable. ***w means m*** will be reduced to **w's *basic use-explainer is m***. And different concepts will be identified with potential tendencies to conform to different use-regularities.[38]

---

38 For parallel reasons the alleged *normative* constitutor of an entire meaning-property—namely, ***w's basic use-norm is the obligation to conform to regularity R***—should perhaps be factored into ***w's basic use-norm is m*** (for the meaning-relation) and ***the obligation conform to R*** (for the concept). Similarly, relative to our rule-following analysis, a decent factorization might well be ***w's basic rule being followed is m*** (for the meaning-relation) and **conform to R!** (for the concept).

In light of these issues, the nice question arises for Gibbard as to whether to take the normative grounding of meaning-properties to stem from the meaning-*relation*, or the concepts *meant*, or *both*. Although the just-suggested factorization of Gibbardian constitutors of meaning-properties implies that *both* the meaning-relation *and* the concepts are constituted normatively, neither of the other two alternative factorizations considered above (when normatized) would have this implication.

18. Indeterminacies of Meaning

We have so far examined two of Gibbard's reasons for thinking that disputes about what words mean are *normative* disputes. First, there's his argument that meaning-attributions must be normative since they analytically entail non-degenerate OUGHT-propositions. And second, there's his defensive argument—based on his claim that normative properties are *constituted* naturalistically—that the view of meaning-properties as constituted normatively can perfectly well satisfy the constraints on an adequate theory (since it *will* be able to accommodate the uncontroversial fact that our acceptance of the sentences we accept is due, in part, to what we mean by their component words). But we've found reason to question the premises of both these arguments: to doubt both that meaning-attributions analytically entail non-degenerate OUGHT-propositions, and that normative properties are constituted naturalistically.

However, there's a third consideration to which Gibbard also gives substantial weight. And it remains for us to address it. He maintains that purely naturalistic conceptions of meaning force us to say, often contrary to our intuitive sense of its being perfectly definite what a certain word means, that the word's meaning is actually *indeterminate*. And he proceeds to argue that these implausible results can be avoided by supposing that questions about what words mean *aren't* naturalistic questions, but are normative.

For example, Quine's naturalistic view that meanings consist in assent/dissent dispositions leads him to the counterintuitive conclusion that there's no determinate fact of the matter as to whether "gavagai" (uttered by a foreign speaker when and only when a rabbit is in view) means RABBIT or means UNDETACHED RABBIT-PART. And Kripke argues that *no* identification of *our meaning what we do by "plus"* with some naturalistic fact about the word could explain why it has the particular extension that it has (especially since that extension includes triples of numbers that are too big for us to entertain). Gibbard says he was inspired by Kripke's diagnosis: that the relation of meaning to use is not dispositional, but normative.[39]

However (and as Gibbard appreciates), Quine and Kripke can be criticized for underestimating the naturalistic resources for resisting their arguments.[40] So the question arises

39 See Quine's *Word & Object* (1960), Chapter 2, and Kripke's *Wittgenstein on Rules and Private Language* (1982), 7ff.

40 In the case of Quine, one might well complain about his unmotivated assumption that the *only* naturalistic facts relevant to fixing the meanings of S's expressions are the circumstances in which she is disposed to accept (or reject) the sentences in her language. (See my "Quelling Quine's Qualms," chapter 9 of *Meaning*, Oxford: Oxford University Press, 1998).

Regarding Kripke, the essential idea behind the resistance I would advocate is to embrace the deflationary view that the explanatory route from a word's meaning-constituting property to the word's extension cannot be *direct*, but must pass through the words meaning. Consider, for example, the case of "dog." One of the explanatory premises must be, <*"dog" means DOG in virtue the fact that it possesses basic use-property* U_{57} *(say)*>. This will be established by showing that <U_{57} *("dog")*> explains the word's overall use. The other

as to why Gibbard thinks that implausible indeterminacies will also issue from the form of naturalistic approach that's been defended here.

While addressing this question, it's important to keep in mind the various different things that can be meant in maintaining, "The meaning of word, w, is *indeterminate*".

One is that, although w definitely has a certain meaning (perhaps its meaning is definitely GREEN), that meaning—hence any word expressing it—is *vague* (some things are neither *determinately green* nor *determinately not green*).

Second, the claim might instead be that it's indeterminate which particular underlying property of w *constitutes* w's having the meaning it does. The indefiniteness here is *not* in which things the word truly applies to, but in whether the basic use-tendency (or use-obligation, or use-rule) that constitutes its meaning is *to conform to acceptance-regularity* $\boldsymbol{R}^j$ (rather than to regularity $\boldsymbol{R}^k$).

Third—and intimately related to this second form of indeterminacy thesis—is the idea that, for a certain concept of ours, F, it's indeterminate whether or not a given foreign word, w, means F. This will be so if (a) it's indeterminate which, among some collection of underlying properties of our "f", is the one that constitutes its meaning-property, and (b) it's also indeterminate which among some *overlapping* collection of underlying properties is the one that constitutes w's meaning-property. In virtue of the overlap there's some possibility that w and our "f" have the same meaning-ground and hence the same meaning—but its indeterminate whether that possibility is realized.

In order to assess Gibbard's contention—namely, that purely naturalist accounts of meaning foster implausible attributions of indeterminacy—let's consider, relative to each of these ways of construing such attributions, how that contention might be made plausible.

To begin with, should we agree that, from the particular naturalistic perspective that advocates *dispositional* meaning-grounds, there will typically be no determinate fact as to which such property is responsible for my "f"'s meaning what it does? Will it typically be the case that if my acceptances of "f"-sentences would be perfectly well explained by supposing that my basic disposition for "f"'s use is to conform with regularity R, then there will be a different regularity R* such that my having a basic disposition to conform to *that* one would explain the phenomena equally well?

It's tempting to think that there are indeed such cases. Consider, for example, the theorems of classical propositional logic—all of which, let's assume, we are disposed to accept. It's well known that there's no such determinate thing as *the* unique set of basic axioms, rules, and definitions that ground all these logical facts. There are various more-or-less equally simple alternative systematizations that do the job perfectly well. In other words, there will

---

explanatory premise will be an instance of the deflationary schema, <*w means F → (w is true of x ↔ fx)*>—specifically, <*"dog" means DOG → ("dog" is true of x ↔ x is a dog)*>. And, putting these together, we derive, <U_{57} *("dog") → "dog" is true of the dogs and only the dogs*>. (For details, see "Kripke's Paradox of Meaning," chapter 6 of my *Truth-Meaning-Reality*, Oxford: Oxford University Press, 2010.)

be several dispositions, one to accept logical principles X, another to accept principles Y, and so on, such that no matter which *principles* are regarded as fundamental, all the others can be derived. Similarly, no matter which acceptance *disposition* is taken to be fundamental, the existence of the other dispositions can be explained. So which one of them fixes what we mean by "and", "not", "or" and "if"? How can the right answer be anything other than, "It's indeterminate"?

But is this not exactly the sort of thing Gibbard is complaining about? For isn't it wildly implausible that the meaning of "and" is indeterminate? Isn't it absurd to suggest that the French "et" doesn't determinately mean AND?

I'd have to say "yes" to these last three questions. Still, I'll now argue, first, that Gibbard's normative account faces exactly the same objection; second, that there's a decent way for the dispositionalist to respond to it; and third, that Gibbard is not as well-placed to do so.

His view, as we've seen, is that the meanings of S's logical terms are constituted, not by S's basic *dispositions* to conform to some particular acceptance-regularity but by S's basic *obligation* to conform. And this obligation qualifies as basic in virtue of the fact that it explains which other logical sentences S ought to accept. But if a given body of logical facts will explain *all* the logical facts, then a *tendency to accept* the collection of sentences that articulate that body of facts will explain the tendency to accept the sentences that articulate all the derived logical facts. Similarly, S's *obligation* to accept the sentences in that collection will explain why S *ought* to accept the further logical truths. So if it's indeterminate whether or not S's disposition to accept logical sentences, S1, S2, . . . , Sk, is explanatorily fundamental in relation to her other logical acceptance-dispositions, then it's bound to be equally indeterminate whether or not S's obligation to accept logical sentences, S1, S2, . . . , Sk, is explanatorily fundamental in relation to her other logical acceptance-obligations.

A reasonable dispositionalist response to the problem, it seems to me, is to begin by distinguishing between, on the one hand, explanatory relations between *logical facts* and, on the other hand, explanatory relations among *a person's dispositions to believe such facts*. We might well concede that it's indeterminate which of various collections of logical facts is the fundamental one. But that would provide no reason to think that there's no determinate fact of the matter, concerning a given person S, as to which subset of S's logical beliefs-dispositions s/he takes to be basic—as immediately obviously correct—as opposed to the rest, which s/he arrives at via inference. Granted, there may be a degree of redundancy in S's body of initial commitments—perhaps some weren't really needed for the sake of complete explanatory adequacy. Nevertheless, that body of dispositions—with all its redundancy—determinately constitutes the idiolectal meanings of S's logical terms.

This response hinges on the fact that explanatory relations among components of the naturalistic realm—which includes the phenomena of thinking—are very tightly constrained by the sheer size of that realm, its high degree of holistic interconnectedness, and the massive volume of relevant empirical data. Thus, it's not easy to find cases in which there's indeterminacy as to whether one such fact explains another. But the *logical* realm is entirely different.

So it's hardly surprising that an indeterminacy in which logical *facts* are fundamental would not carry over which of S's logical *commitments* are fundamental.[41]

In contrast: the logical realm—and the body of true logical sentences—does *not* seem entirely separated from the facts concerning *which logical sentences S ought to accept*. Arguably, S ought to accept a given logical sentence just in case it's true. And in that case, since it's indeterminate which sentences express the *fundamental* logical facts, it's got to be indeterminate which logical sentences S is *fundamentally* obliged to accept. Thus, it would seem, ironically, that Gibbard's normative account of meaning is *more* prone to attributions of meaning-indeterminacy than are certain fully naturalistic approaches.[42]

---

41 Another conceivable route to Gibbard's 'indeterminacy' critique of naturalistic-dispositionalism about meaning would begin with the general observation that explanations of phenomena in terms of dispositions must deploy further assumptions specifying conditions in which the disposition will *not* be manifested. Only then can we fully explain why a thing, X, exhibited property, Q, in terms of X's disposition to exhibit Q in circumstances, C. For we can point out, not only that C obtained but that none of the potential defeating factors, D1, D2, . . . Dn, were present. And this introduces the prospect of alternative (indeed, *empirically equivalent*) explanations of the phenomena—explanations that invoke different dispositions, but where these divergences are 'cancelled out' by differences in the lists of potential defeating conditions that they postulate. And, in such a case, one might be inclined to hold that there's no determinate fact as to which explanation is right. (E.g. suppose that one explanation combines a disposition to exhibit Q in C with potential defeating factors, D1, . . . , Dn; whereas the other combines the disposition to manifest Q in C & not D1 with defeating factors, D2, . . . , Dn).

Applying these considerations to the case of word-use, we'd again get the result that there's no objective answer to the question of which particular disposition is the explanatory basis for its overall use. But again, although one can appreciate how it may be very hard to see what could favor a given answer over certain alternatives, it's unreasonable to jump to the conclusion that they are empirically equivalent (let alone that they are not determinately distinct). For the fact is that scientists—especially outside of physics—very frequently deploy such explanations. And the usual methodological constraints of empirical adequacy, simplicity, and consistency with established theory, typically suffice to yield objectively plausible answers. I can see no reason to expect that, when applied to the phenomena of language, this technique couldn't be just as valuable as it has proven to be elsewhere.

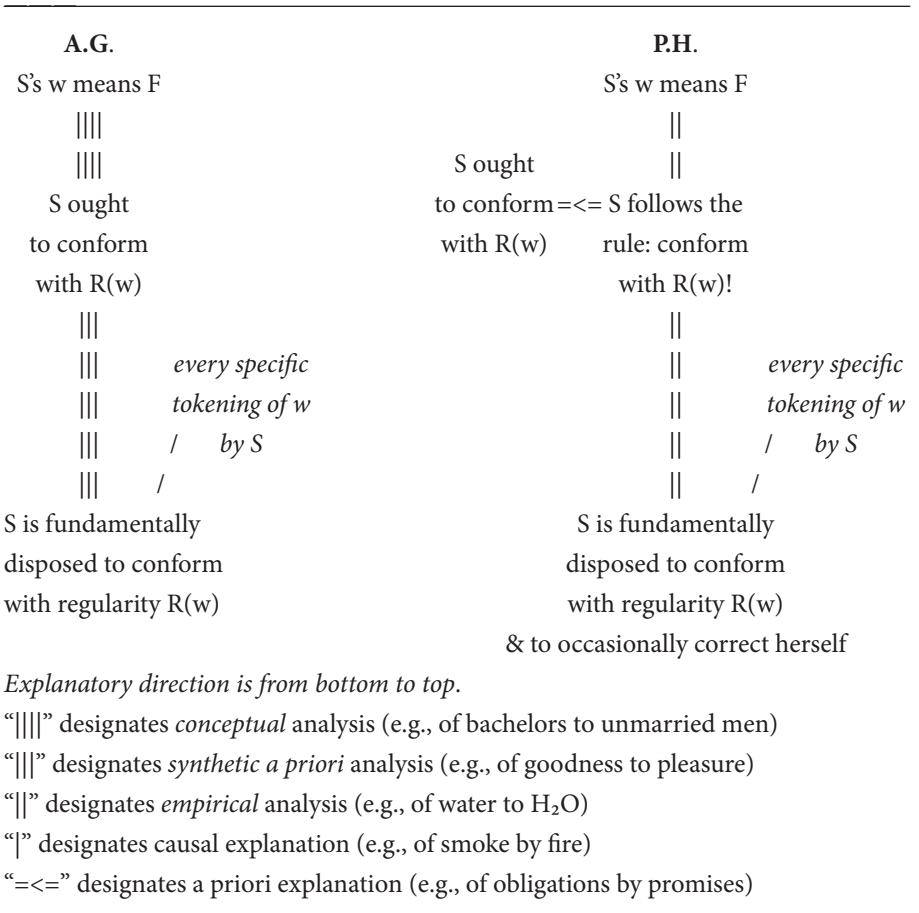
42 A quick word on whether Gibbard's contention would be plausible if understood as the claim *that if meaning is explained entirely naturalistically, then many more words will turn out to be vague than we intuitively feel to be the case*. It might be argued that this is indeed so, since

1. For most predicates there are objects, inaccessible to us, to which we have no disposition to apply "f" or "not f" (and know that we never will have); and
2. What makes a predicate vague (e.g., "rich", "heap", "red") is that there are things to which we aren't disposed to apply the predicate, aren't disposed to apply its negation, and know that we never will have either of those dispositions.

It seems to me, however, that the account of vagueness assumed in (2) is not quite right. On encountering borderline cases of a vague term, it's not just that we *lack* certain dispositions (to apply it, or to apply its negation), and know we'll always lack them. It's rather that we are positively disposed to *refrain* from applying the term, and to refrain from applying its negations. Moreover, these reactions are epistemologically basic; they are constitutive of the meaning of the vague term; and for that reason, we appreciate that no conceivable improvement of our evidential circumstances could make any difference. This is not at all our reaction to the question whether "prime" applies to some inaccessibly huge number or the question

19. Comparisons and Conclusions

Here's a sketch of similarities and differences between Gibbard's position and what I take to be a good alternative that's purely naturalistic:



The extent and depth of agreement here are very substantial: namely,

- That a word's meaning is engendered (at least in part) by a fundamental *tendency* to use the word in conformity with a certain regularity.
- That the speaker *ought* to conform to that regularity.
- That S's meaning-constituting use-tendency for "f" is the one that explains (in conjunction with other facts) all her specific tokenings of the word.
- That this will rarely (if ever) turn out to be the tendency to apply it to fs and only to fs.

whether "dog" applies to some animal in the vicinity of Alpha Centurai—where our uncertainties (indeed permanent uncertainties) are fully explained by the 'remoteness' of the objects at issue.

- Rather, it will typically be S's tendency to accept "$(\exists!\phi T\phi)\rightarrow Tf$"—where "Tf" articulates her fundamental theory of f-ness.

Thus, the two theories diverge hardly at all with respect to what the relevant phenomena are, but mainly with respect to how they see the explanatory relations between these phenomena. In particular, there's disagreement as to:

- Whether our *obligation* to conform with R(w) can be *constituted* by our *disposition* to conform with it. (It's argued in section 11 that the explanatory relation between a naturalistic phenomenon and its normative import is never one of *constitution* or *identity*, never akin to the relation between *being an unmarried man* and *being a bachelor*, or between *being made of H_2O* and *being a sample of water*.)
- Whether, given a disposition's capacity to account for all specific uses of w, we are entitled to attribute that same explanatory power to the associated obligation and hence to the meaning-property that this obligation supposedly constitutes. (An implication of the preceding point is that we're *not* entitled to that inference.)
- Whether there really are any basic norms of the form: "If S is disposed to conform with R(w), then S ought to conform with R(w)". (It was suggested in section 12 that this normative consequent calls for a stronger antecedent. It may need to be, "S is implicitly following the rule: conform to R(w)!"—which is analyzable as "S is disposed to conform with R(w) and to sometimes be dissatisfied with (and immediately correct) his initial manifestations of that disposition."
- Whether meaning-attributions are normative—that is, whether they analytically entail nondegenerate OUGHT-propositions. (See sections 3–6 for arguments to the effect that they do not.)
- Whether the dispositional ground and the semantic-normative import of a meaning-property are always discoverable *a priori*. (See section 16 for reasons to think—*pace* doubts about the significance of the distinction—that such investigations may have to be *empirical*.)
- Whether, as Gibbard contends, pure naturalism about meaning inevitably dictates counterintuitive attributions of indeterminacy, which his own theory does not. (See section 18 for considerations suggesting, on the contrary, that only the naturalistic approach has the resources to block the threat of rampant indeterminacy.)[43]

43 I've learned more from Allan Gibbard than from anyone else what the shortcomings have been of my attempts to give a fully naturalistic account of meaning, and where further arguments were called for. This learning process took a leap forward on reading his *Meaning and Normativity* and seeing the formidable case that can be made for his norm-infused alternative. At the moment, I remain unconvinced by that alternative. But all I really mean by this is that it still seems to me worthwhile to try to develop and improve my own contrasting approach. And for the sake of healthy competition, I hope that he feels the same way about

his perspective. In fact, it could well be that we have pretty much the same convictions and uncertainties here: it's just that I've given myself one job and he's given himself another.

As for the present paper, much of it grew out of ideas that I tried out in August 2014 at the "Global Expressivism" conference in Szczecin (Poland). After that, I delivered what I hope was a better presentation of the issues at the Ann Arbor conference in Gibbard's honor that took place in May 2016. On both occasions, he was present; he raised deep and difficult questions; and I must thank him for any gains in my understanding that have resulted from my struggling with them. I'm also grateful to the other participants in these meetings for their comments and objections—especially, Simon Blackburn, Paul Boghossian, Annette Bryson, Christine Korsgaard, Peter Railton, Lionel Shapiro, and Sharon Street. In addition, I'd like to thank Hannah Ginsborg, Hans-Johann Glock, and Severin Shroeder for valuable feedback on an earlier draft of this essay. And finally, I'm pleased to express my appreciation to David Plunkett for his astute suggestions for how to improve the penultimate version.

19

THE NORMATIVE EXPLANATION OF NORMATIVITY

Jamie Dreier

What does it mean for metaethics if meaning is normative?

That is the question this chapter explores. Mostly it asks what happens to *expressivism* if meaning is normative.

In the first section I try to say in a theoretically neutral way what it means for meaning to be normative, and what reasons there might be to think that it is. Then until section 6 I proceed on the assumption of the normativity of meaning, and ask what follows; only in that last section do I return to the question of whether it is true.

1. The Supposition: Meaning Is Normative

Although it is controversial, the thesis that meaning is normative is not exotic but fairly commonplace. It is defended by Donald Davidson and Saul Kripke, for example.[1] But it is not a thesis whose content is obvious; it is not even obvious what it is supposed to mean. *What* is normative? What sorts of things are alleged to be normative? Meanings? And there is a second worry: what is being *normative*? I will say something about each of these.

1.1. What Is the Supposition?

The supposition I will be making is this:

> *Meaning is Normative*: Judgments about what meanings meaningful things have are normative judgments.

---

1 (Davidson 1985), but see (Schroeder 2003) and (Kripke 1982).

The thesis is about judgments. Judgments are the sorts of things that can be normative.

Some philosophers mean something else. They think that the stuff that the judgments are about can be normative or nonnormative. On this question I will remain neutral as long as I can; later in the dialectic (section 5) I will defend a position on it, but the point of framing the thesis to be about judgments is that it can be accepted irrespective of whether one thinks that, say, the property of being courageous is the right kind of thing to be normative, or instead something other than the subject of the judgment that Carla is courageous gives the judgment its normativity. Similarly, if we think Meaning is Normative, we will agree that Keith's judgment that by "rojo" Lucia means "red" is a normative judgment, regardless of whether we think that the meaning relation is itself the right kind of thing to qualify as normative.

I am taking states of mind themselves to be meaningful things. Your belief that apples are red has a content (that apples are red); it represents the world as being a certain way. So, when someone judges of your state of mind that it is indeed a belief that apples are red, that judgment, according to Meaning is Normative, is a normative judgment. It is possible to hold that judgments about the meanings of words or sentences are normative without holding that judgments about the contents of mental states are, or vice versa, but the two views tend to go together and fit well with each other. And in the context of a discussion about expressivism, they will have to go together, since according to expressivism what sentences mean is fully determined by what states of mind they are used to express.

What about the question of what counts as being *normative*? Here we face a serious problem, I think. "Normative" in philosophy is a technical term, so I should be able to define it. And, it is not at all clear that it means one thing, in philosophical usage; in fact, it seems unlikely that it does. So, I should be able to say which meaning I intend. But, I can't do what I should be able to do. The main reason is that there is no theory-neutral way of characterizing normativity. Different metaethical (or metanormative) theories characterize it in different ways. For example, a nonnaturalist realist view may characterize the normative as being about a distinctive part of reality. A rationalist view may characterize it in terms of the constitutive features of practical rationality. And, of special interest, an expressivist view characterizes the normativity of judgment in terms of the functional features of the state of mind. Gibbard, of course, says that a judgment is normative when it is *plan-laden*. Each of these characterizations is helpful in the context of its metanormative theory. We may hope, at least, that they can be sensibly thought of as alternative views about one subject matter. We hope that there is one thing, which we can recognize, that could be a matter of the special domain of its subject matter, and could be a matter of its connection to the constitutive features of practical rationality, and could be a matter of whether it is laden with planning. But as far as I know there is no further, theory-independent way of saying what this one thing is.

There are some tries at saying. We can try saying that normative judgments are the ones that imply reason-involving propositions (Parfit 2011). But this try is not helpful, in the end. For one thing, it is already entangled in its own controversial theory: that reasons are the

fundamental normative unit out of which all other normative materials are built. It also seems to be committed to a very dubious assumption: that when something *implies* a normative proposition, it *is* a normative proposition. I think this assumption is not only dubious but demonstrably false.[2] But the worst problem with the reason-involving suggestion is that it fails to account for the kind of reasons (or senses of "reason") that are not normative. For example, according to Parfit, Bernard Williams's judgments about internal reasons are not normative at all. And, obviously, if we say instead that normative judgments are the ones that imply propositions that involve normative reasons, we will not have made enough progress (see McPherson 2018).

In section 6, I will return to the important question of just what counts as a normative judgment or claim. For now, I will have to count on there being enough overlap in philosophers' grasp of this common topic to make what I say intelligible.

1.2. Why Believe It?

This section is *not* an attempt to vindicate Meaning is Normative. But without some reason to believe the thesis, the exercise of proceeding on its supposition is going to look sterile or pointless.

It will help to look at a particular version, and it is natural to choose Gibbard's, which although a bit complicated has the advantage of being well-formulated and ultimately very clear. Gibbard's version of the thesis has a view about analyticity at its core. Two sentences are analytically equivalent, according to Gibbard, if and only if we ought to give them the same credence under any supposition. Your conditional credence that Ferdinand is a bull given that he is smelling clover should be the same as your conditional credence that Ferdinand is a male bovine given that he is smelling clover; and likewise no matter what (possible) condition is substituted for Ferdinand's smelling clover. And the biconditional is offered as itself a (rough?) analytic equivalence, not as a striking coincidence, for example. So, analytic equivalence automatically involves an *ought* connection between our doxastic states.[3]

One reason to believe Meaning is Normative arises from Saul Kripke's reflections on Wittgenstein's Rule-Following argument. Suppose we start with the idea that what a person means by a certain symbol, say "+," is determined fully by her dispositions to use the symbol in a certain way. So, for instance, she is disposed to *add* to numbers n and m and reply with their sum when asked, "How much is $m + n$?" If she were disposed always to multiply, then surely it would be very strange to impute to her use of "+" the addition function as a meaning. Surely, we would think instead, "Oh, she is just speaking a slightly different language, in

2 First explained in (Prior 1960). See also Dreier (2002), and especially (Russell 2010).

3 It's important that Gibbard is not merely proposing a Law of Epistemic Rationality, as one might propose an ethical principle. Using a very plausible normative principle to support Meaning is Normative is a mistake, and I think it is a common one; it is not a mistake Gibbard is making here. I will return to this issue in section 5.

which that symbol stands for multiplication." We'd be employing a kind of inferential role semantics, where the role is a causal role, identified by dispositions to infer. But, famously, Kripke's Wittgenstein (Kripke 1982) has a compelling objection to this approach. For it seems to make it impossible for a person to make an error in addition, or at least to have a stable disposition to make an error. For if someone were inclined typically to make a constant mistake in summing fifteen-digit numbers, she would *not* then have the dispositions that we wanted to say determined that she means addition by "+." We might have to conclude that she means *quaddition*, which is just like addition but differs for inhumanly large arguments. Kripke points out that what seems to define her meaning *addition* is not that she *will* or *would* answer our questions by summing, but that she *should*, that it would be *correct* to do so. Thus, fixing a shortcoming of a dispositional or functional characterization of meaning leads us into a normative characterization.

A second reason to believe Meaning is Normative has to do with disagreement. Imagine that two philosophers of language, with competing theories of metasemantics, learn all of the relevant nonsemantic facts about Carla, including her dispositions to form judgments using the symbols in question, her dispositions to infer, the causal connections between her use and the world, and so on. But, these philosophers disagree about what Carla means by "+"; perhaps one concludes that she means addition and the other concludes that she means quaddition. What is the content of their disagreement?

It cannot be about the nonsemantic facts, since they agree on those. One plausible thought is that they disagree about what Carla *ought* to say when asked questions of the form "How much is $m+n$?," for suitably large values of m and n. So according to this suggestion, the substance of the disagreement over what "+" means, as Carla uses it, is a disagreement over what she ought to say or judge. We may have a suspicion that this case is similar to the sort of case G. E. Moore considered in *Principia Ethica*, in which someone knows all the nonethical facts and is wondering whether, given those facts, a certain event is *good*. She is wondering how to behave or what to prefer; the remaining uncertainty, once she is certain of all the nonethical facts, is practical uncertainty. For an expressivist, that means the remaining question is a normative question. So, Moore's example does look relevantly similar to our question about disagreement over what a symbol means. And this consideration bolsters the suspicion that the question of meaning is a normative question.

I don't say (and don't believe) that these two reasons are conclusive. The point of this section is just to motivate the supposition well enough to make further investigation interesting.

1.3. Combining the Supposition with Expressivism

For the next three sections, I will be asking what problem or problems arise for expressivist explanations when we suppose Meaning is Normative. Something worrisome seems to happen.

The project of expressivism is to say what normative expressions mean, and what our normative thoughts mean. Now we are supposing that Meaning is Normative. So, saying

what things mean is normative. In Gibbard's prominent expressivist theory, that means that saying what things mean is expressing our plans. Imagine that Hera says, "Herakles ought to lay down his sword now." More precisely, her assertion is analytically equivalent, according to expressivism, to the conditional prescription, "If in Herakles situation, let me lay down my sword!" (The exclamation mark indicates imperatival force; in Gibbard's notation there is another upside-down exclamation mark at the beginning of the imperatival clause to disambiguate, but we won't need that device.) What does this analytic equivalence amount to? Analytic equivalence is a *semantic* relation; it is a normative relation. As Gibbard explicates it, the claim of equivalence in this case means that it would be *incoherent*, a failure of rationality, for Hera to accept that Herakles ought to lay down his sword now while at the same time rejecting the conditional plan to lay down her sword if in Herakles' situation. Calling it incoherent is not merely describing. It is itself a normative judgment. In explaining the meaning of Hera's normative "ought," expressivism makes a normative judgment of its own, which can in turn be explained only in terms of further normative connections between states of mind. In explaining we express our plans; to explain what we are doing in expressing those plans we can only express more plans, and so on.

This result *seems* problematic. It is a bit dizzying. And it raises the worry that the explanation is incomplete, and maybe uncompletable. There is a threat of circularity. But what, exactly, is the problem? In the next two sections, I look at two ways of fleshing out the worry. I end up saying that neither of these ways amounts to a serious objection to expressivism (even under the supposition that Meaning is Normative). Then in section 4 I identify a serious worry, and in section 5 I spell out one way to respond to it.

2. The Problem of *Circularity*

The most obvious worry about employing normative terms in the explanation of normativity is that it makes the explanation circular. Circularity is a defect in explanations. Let us see if and how the charge of circularity can be made precise.

Expressivism seeks to explain normative sentences in terms of the states of mind their assertions express. There is no expressivist theory of *what it is* for it to be the case that *A ought to F*; instead the theory offers to tell us what it is to *judge* that *A ought to F*. The substitute explanation is disappointing to some, but if expressivism is true it's the best we can get, and it does promise to explain many of the mysterious features of normative thought and talk.[4]

Traditional, causal expressivism could characterize those states of mind in functional terms: by how they are related to experiences, to plans and intentional behavior, to one another. But if Meaning is Normative, the functional characterization is no good. The states of mind must be explicated by their *normative* relations: by how they *ought* to be connected

4 (Blackburn 1984), (Gibbard 2003), Dreier (1992); also Dreier (2015) on expressivism's failure to deliver on the explanation it promises.

to experience and plan and each other. To judge that Herakles ought to lay down his sword now, it says, is to be in a state that *ought* to be accompanied by the conditional plan to lay down one's own sword should one find oneself in Herakles' (present, actual) situation. So now, not only does the theory fail to explain what it is for it to be the case that Herakles ought to lay down his sword, but what it does explain it explains by helping itself to facts about what states ought to be accompanied by what other states. This pattern of explanation surely threatens a regress, if the sort of *ought*-facts cited are different from the sort whose judgment we are trying to explain, or else a circle, if at some point we find that the *ought* of the judgment is the same as the *ought* we cite in its explanation.

Here is a simplified, but in all essentials similar, version of Gibbard's account of what it is to believe that one ought:

> One ought not believe that one ought to F and yet decide not to F.

As Gibbard himself notes, the worry seems clear if we replace the word "ought" with some word we do not understand:

> One blag not believe that one blag to F and yet decide not to F.

This account does not explain *blag*; it doesn't explain what the word means, and it doesn't explain the nature of the blag relation. It does help: after all, it may rule out the hypothesis that "blag" means "will," since (according to Meaning is Normative) some people will sometimes believe they will do something and yet decide not to do it; they will make a rational mistake. So, the account narrows things down a bit, but seems to be a radically incomplete explanation: "These do qualify as restrictions on the use of the word 'blag'," he says, "but they fall dismally short of explaining what the word is supposed to mean" (Gibbard 2012).

Now, one kind of explanation is epistemic: it seeks to provide the uninitiated with understanding. People who do not understand the word "blag" cannot be brought to full understanding by being told something that *uses* that word, for they won't understand the sentence and, in the case of Gibbard's own sentence, they won't be able to deduce the meaning from the context.

That is a failing. But it is not clear that it is a fatal one. In the first place, notice that it will *commonly* happen, in seeking to give the meaning of a class of words, that we will use some of the words in the class in our account. For example, if we try to give an explanation of what logical particles mean, it seems inevitable that we will end up employing some logical particles. It would be a very simplistic speech indeed that avoided all logical particles. So, apparently, we cannot give all of the meanings of the logical particles in a way that pins them down, for people who do not already understand at least some logical particles. But after all that is not very surprising. It would be difficult to explain much of anything to people who did not possess any logical particles at all in their language! The fault, we might say, lies more in the audience than in our attempt.

And similarly, as a practical matter, we do not expect to run in to many people who do not understand any basic normative vocabulary. If we do, well, we will not be able to enlighten them using straightforward definitions and explanations. (Presumably we would have to get them to have certain kinds of experiences and then gesture somehow at aspects of the experiences.) And in that case we will not think that the main problem with the exchange is that our explanation is intrinsically deficient; instead, we'll think it is perfectly natural that we should be unable to get our point across to speakers of such an impoverished language.

In the extreme case, we might have the hopeless assignment of explaining, in language, the meanings of all words, to people who speak no language at all. That is obviously impossible, and its impossibility corresponds to the very large circle of explanation involved in constructing a dictionary. Our limited task, of trying to explain normative vocabulary and ending up having to employ some normative vocabulary to do it, is just a writ-small example of the same problem. So, although the circularity here is real, it does not seem to be philosophically troubling.

There is another kind of explanation: it is a relation that obtains between things in the world. In offering up some words, we sometimes try to articulate the patterns of this relation. Causal explanations are like this. We explain the thunder (not the word "thunder") by appeal to lightning and energy and sound waves (not to "lightning" and "energy" and "sound waves"). This sort of explanation is defective if it is circular, or so we always assume. The relation is transitive (or if not, substitute its ancestral), and it is irreflexive and asymmetric. So, when we find that the chain of explanation we have offered loops around, we can conclude that we have made a mistake.

But expressivist explanation of the normative does not work in this way; it does not attempt, as I have noted, to explain what it is for it to be the case that Herakles ought to lay down his sword. Instead, we said, it seeks to explain what it is to *judge* that Herakles ought to lay down his sword. It explains this in terms of what someone who so judges ought to plan, conditionally. This latter, it turns out, is normative. So, expressivism is explaining the normative in terms of the normative. But it is not explaining anything in terms of itself; the explanation does not loop around in a circle. For nowhere in the explanation of what it is to judge that Herakles ought to lay down his sword, do we find that fact itself. There are a couple of things that may look fishy about this story, but the problem cannot be that something is being used to explain, in the sense of saying what it is to be the sort of thing it is, itself, because that isn't what's happening. Those fishy things will be the subject of the next two sections.

I conclude that the most straightforward charge of explanatory circularity doesn't stick.

3. The *Wrong Project* Problem

If the expressivist explanation of normative judgment is itself normative, one might worry that the original project has not been accomplished and cannot be accomplished. The

project, one might think, was to give an account of normative judgment in naturalistic, or at least purely descriptive, terms. For expressivism is a *metaethical* theory, not a theory within normative ethics or a theory that connects two normative domains with one another.

Compare Consequentialism, which attempts to explain the deontological by means of the axiological. A Consequentialist, like G. E. Moore, offers an analysis of our *ought* judgments (Moore 1993, 2005). Moore helps himself to the unanalyzed property of goodness, or value, in his analysis of what one ought to do. For Moore this is no compromise, for he famously thinks that goodness is a simple unanalyzable property. His Consequentialism, though, is not an integral part of his metaethics. Indeed, consequentialism is perfectly compatible with expressivism, with naturalist realism, and no doubt with other metaethical views; they seem to be independent moving parts. In general, we might say, when one normative property or concept is explained in terms of another, we are doing normative theory. So maybe metaethics proper can try explaining the ethical in terms of other normative stuff or concepts, but when we are explaining one ethical property in terms of another we are still in the realm of normative ethics. Similarly, expressivism is trying to explain normative judgment *in general*, not just some normative domain in terms of another. It is a *metanormative* theory, and metanormative theory is separate from internormative explanatory theories. Expressivism, in particular, is supposed to explain at a different level from Consequentialism, or Deontology. So, if Meaning is Normative, expressivism will fail in its aspirations and become simply a branch of normative theory.

An expressivist might defend herself: if Meaning is Normative, then every metaethical theory is in the same boat with respect to this problem. For explaining what normative expressions mean will necessarily involve engaging in normative theory, in normative explanation. So even if we end up doing something other than what we had originally wanted to do, we will have all other metaethicists as companions in our guilt.

While this is true, it is unsatisfying to leave things at that. For one thing, it leaves open the possibility that expressivist metaethics was misconceived: it had a goal that doesn't even make sense. The philosophical explanation of normativity would have to be reconceived, and if traditional metaethical theories all fell together, that would be small comfort to proponents of expressivism. For another thing, it looks like a thoroughgoing naturalistic reductivism about normativity could end up being in a much better position than expressivism to carry out its own self-conceived project. For suppose we had a reduction of reasons for action and reasons for propositional attitudes to some Humean basis; maybe the program of *Slaves of the Passions* (M. Schroeder 2007) can be satisfactorily completed. Then we notice that Meaning is Normative. No problem! That only means, facts about what a person means by her use of, say, "reason," are partly constituted by facts about human beings' desires. Of course, when we attribute desires to people, we will be employing normative vocabulary; but by hypothesis all of this vocabulary simply picks out bits and pieces of the natural world. So, a reductive naturalist might be quite happy with the picture that emerges. Not so an expressivist, at least

according to the characterization of expressivism that assigns it the task of giving an account of normative judgment in nonnormative terms.

But, in the big picture, there is no real *objection* here. True, if Meaning is Normative, then the project of giving meanings is not exactly what we might have thought it was when we thought that meaning attribution was more like naturalistic description. The thesis, if true, is surprising and has some immediate surprising implications. When you ask me what a word means, you are asking about what you ought to do; you are seeking advice about how to behave and speak and think. So, if I am to give you a good answer, I will have to tell you what to do. My answer will be prescriptive. As expressivists understand prescriptivity, I will be expressing my plans for how to talk and what credences to assign if in your position. Ask me a normative question, get a normative answer; ask me how to think and live and I'll do my best to tell you. The line between metaethics and normative theory will turn out not to be what we thought. It will be like the line between two branches of the theory of reasons, or between epistemic and practical rationality. We would hope to explain one normative domain from another, as a Kantian seeks to explain moral imperatives by reference to the fundamental norms of rational action. There would, at least if things go well, still be explanations available. But they would be normative explanations.

4. The Real Problem

If Meaning is Normative, then an attempt to explain the normative in the only way expressivism offers will turn out to depend on the normative. In the previous two sections, we examined two ways of spelling out how this dependence might be a *problem*, as opposed to an interesting feature, for expressivism. I think there *is* a problem. In brief, it is this: explanations are not supposed to bottom out in appeal to the normative domain, if expressivism is true. Ground-level normative features of the world are the property of nonnaturalist realism. I will now give one way of articulating this assignment of property. It depends on a metametaethical understanding of realism and expressivism that not everyone will share. I suspect, though, that those who do not share my metametaethical view will be able to find some other way of articulating the problem.

It is by now a familiar observation that sophisticated, quasi-realist expressivism endorses many of the claims made by nonnaturalist realism. Moral (and other normative) words and concepts are not definable in naturalistic terms; it is true that slavery is wrong (by the minimalist deflationist account of truth, plus the unobjectionable thought that slavery is wrong); there is some property that sexist language and racist language share, namely being morally objectionable (since property talk arrives without further theorizing once we predicate objectionableness of both kinds of language; Thomasson 2014); moral judgment is a kind of belief (since a belief is nothing more than a state of mind expressible by assertion of a declarative sentence (Wright 1988)). Typically, contemporary expressivism affirms such

truisms, although Ayer (1952) did not. So, unlike Ayer's view, contemporary expressivism is difficult to distinguish from nonnaturalist realism.[5] What exactly is the difference?

A popular view, and one that I have defended (Dreier 2004), says that what distinguishes nonnaturalist realism from expressivism is the *explanations* they give; more precisely, it is what ingredients we find in the explanations they offer of what it is to make normative judgments. The idea is that the ultimate ontological commitments of a theory are to be found in their explanans. Those things (and facts and properties and relations, if like some (Scanlon 2014) you don't want to call those "things") that bear ultimate explanatory weight are the ones the theory counts as "real," in the sense that is supposed to distinguish realism.

What is it for Michelle to believe that sexist slurs are wrong? According to expressivism, it is to plan not to use them; or, to plan to feel guilt if one does use them; or, to have some complex of higher-order emotions toward people use such slurs. In the explanans there is no mention of wrongness itself, or of the constituents of wrongness. Expressivists need not deny that there is such a thing as wrongness. But they do not *use* it in their explanations of judgment or of the meanings of assertions. By contrast, there seems to be no way for a nonnaturalist realist theory to avoid mentioning wrongness, or perhaps its constituents, in explaining what it is to believe that it is wrong to use sexist slurs. Nor would a nonnaturalist see any point in avoiding the mention, of course, any more than a physicalist realist about astronomy would want to avoid mentioning Antares in explaining what it is to believe that Antares is a red supergiant.

It makes sense for the difference between a heavy-duty realist view in metaethics and an expressivist or otherwise anti-realist view to be a matter of what explanatory work the moral facts and properties can do. Gibbard's own conception of the difference between his own view and, say, Moore's squares with this Explanationist criterion (Gibbard 2003, esp. 18–20). And the criterion has some impressive credentials, with support from Kit Fine (2001) and (O'Leary-Hawthorne and Price 1996). It is certainly not universally accepted (see Sturgeon (2006), Chrisman (2008), and Golub (2017) for some interesting dissenting views).

But, it will not have escaped the reader's notice, if Meaning is Normative the Explanationist criterion yields an unexpected result: expressivism is a kind of realism. Indeed, if Meaning is Normative, the criterion tells us there is no escape from realism! For every explanation of what it is to say or to believe something normative will involve normative facts and properties. So it appears, anyway. So, Explanationism plus the normativity of meaning will close

5 There are exceptions. Developing expressivism without the accompanying "minimalism" about truth is gaining popularity; see, for example, Tristram McPherson's chapter in this volume, and Yalcin (2007) along with other work by Yalcin on expressivist accounts of epistemic and probability modals. Gibbard, of course, is happy with minimalist conceptions of many semantic and metaphysical concepts, while leaving open the possibility that there are other, perhaps more inflated conceptions available.

off metaethical possibilities. Maybe that's progress! But it would be bad news for alternatives to realism.[6]

It would not *per se* be bad news for naturalistic realism. A naturalistic realist could be happy with the idea that normative properties and facts figure in the most fundamental description of the world. Those features, she would add, are naturalistic features. Maybe they are reducible to paradigmatically naturalistic features (M. Schroeder 2007) so that they don't really figure in the *fundamental* description of how things are but do show up in perspicuous true explanations of what we do when we make normative judgments, as high-level descriptions of our brains might show up in a neuroscientific account. Or, maybe they are irreducible but meet all the plausible conditions for being natural: they cause things and are caused, they are discovered by the scientific method (Sturgeon 2006), and they are fully grounded in the basic physical facts (Rosen 2017). To say Meaning is Normative is not to say it is *sui generis* or nonnatural.

Expressivists, it seems, cannot be content as naturalistic realists can with the normativity of meaning. If it turns out that their explanations of what we do when we engage in normative thought and discourse is ineliminably normative, then by the Explanationist criterion expressivism is a kind of realism. And, labels aside, it certainly looks like expressivism will have to abandon its naturalistic aspirations. Here is how Gibbard puts the worry, in *Meaning and Normativity*:

> The metatheory of meaning I have developed explains the meaning of "ought" in terms of "ought" itself. It does so in a roundabout way, but still, it might be thought, this makes it not a version of expressivism, but an elaboration of non-naturalism. The story takes the concept ought as primitive: you have to start with the concept to understand the metatheory that explains it.

As I see it, the problem is not really about understanding and concepts. It is about what explains our states of judgment and assertion, what it is to make judgments and assertions. The normative realm, according to the normative metatheory of meaning that Gibbard develops in *Meaning and Normativity*, carries the kind of explanatory weight that entails it counts as real, and its theory as realist.

---

6 In a comment on an earlier version of this chapter, Billy Dunaway suggested reformulating the criterion for realism so that it divides between Meaning is Normative expressivism and realism by focusing on the *way* that normative properties figure in fundamental explanations of normative judgment. For expressivists, *ought*s will enter when we say what it is for a judgment to be of one kind or another: what kind a judgment is will be a matter of what *ought* to be inferred from it, and what it *ought* to be inferred from. Dunaway's suggestion seems very promising but I cannot give it proper attention in this context; I hope to be able to say something about it in future work.

5. A Solution

The solution I offer in this section is, I believe, the *only* solution to the problem as I have posed it. In section 6, I will consider another solution: rejecting the assumption the Meaning is Normative.

Let me pause to make the argument rigorous. I will be pedantic. There are some distinctions that are easy to gloss over and that I will need in my solution. Let me also pick a candidate fundamental normative predicate: being *okay* to do. The subject of the predicate can be any sort of "doing," very broadly construed, including verbs that do not express actions, like "believe" or "sleep." We'll imagine that all other normative concepts can be explained in terms of being okay (Gibbard 2012). And we'll spell out the normativity of meaning in a simplistic but representative way, in terms of being okay.

EX To judge that *x* is okay in situation *S* is to allow in planning for *x* in *S*.

Here to be true to Gibbard's theory we can explicate "allow in planning" as follows: to allow for *x* in planning is to rule out ruling out *x* from one's plans. This explication makes allowing different from simply not ruling out; we can imagine failing to rule out a certain choice not because one regards it as okay but because one has never considered it, or has considered it but been unable to make up one's mind about it.

MN To allow for *x* in a plan is to be in a state that is *okay* to have together with certain states and *not okay* to have with certain others.

MN schematizes the normativity of attribution of intensional attitudes. Plans with certain contents or certain logical features are picked out not by what they are causally, dispositionally connected to but what they are normatively connected to. They are characterized by relations of mutual coherence and rationality, ultimately *okayness*. So, if we put EX and MN together, we get an explanation that looks schematically like this:

EXMN To judge that *x* is okay in situation *S* is to be in a state that is *okay* to have together with S_1, S_2, . . . and not *okay* to have together with S^1, S^2, . . .

Now we want an Explanationist premise.

EXPL A theory whose explanation of what it is to judge that *x* is okay involves normative facts and properties is a realist theory.

Then, since expressivism's explanation of what it is to judge that *x* is okay is EXMN, it seems to follow that expressivism is a realist theory.

But I am being very pedantic, so let's spell out this last step. What we need is:

F That a state is okay to have together with some states and not okay to have together with other states is a normative fact.

F seems trivial. But, I think, the solution to the problem is to deny it.

What I want to deny is not specific to F. You might complain, "Being okay is supposed to be the fundamental normative predicate, in this simplified model. So the fact that a certain state is okay to have with some other state is a normative fact, if anything is." Indeed; but, nothing is. There are no normative facts.

Now this denial may seem to be very much against the spirit of minimalism and deflationism. In ordinary talk we seem to speak of normative facts all the time, and the quasi-realist accommodation of such talk is supposed to accept the idea of facts and properties, in the ecumenical embrace of minimalism. How are we to escape this compelling argument?

Normative Facts Argument

Sex discrimination is unjust.
If sex discrimination is unjust, then it is a fact that sex discrimination is unjust.
If sex discrimination is unjust, then sex discrimination has the property of being unjust.
The fact that sex discrimination is unjust is a normative fact, and the property of being unjust is a normative property.
So, there are normative facts and properties.

The two conditional premises are licensed by minimalism: there is nothing more to a claim about *the fact that p* than there is to the claim that *p*, and nothing more to a claim about *x* having the *property of being F* than to the claim that *x* is *F*. The inferences are licensed immediately by the rules for using predicates and property and fact talk. And the first premise is unexceptionable, and in any case could obviously be replaced by whatever normative claim one prefers. So to resist the conclusion we will, apparently, have to block the move to the fourth line of the argument. And that is the correct point of resistance.

There are normative claims and normative predicates; and, some of these claims are true and some of the predications are accurate; and, true claims state facts and accurate predications attribute properties. But, the facts and properties are not normative. The normativity of predicates and sentences is not a matter of the nature of the parts of the world they are about. It is not to be found in the condition of the world in which they are true and accurate. Normativity accrues not to what we have described but to the way we have picked it out.

There are different ways to fill in this gloss on normativity; we are, of course, most interested in the expressivist-friendly ways, but there are others. To make what I am saying seem less perverse and enigmatic, I will draw an analogy. The analogy will suggest another way of

filling in the gloss on normativity, but I do not mean to be advocating this way. The analogy, I want to be clear, is only an analogy.

Indexical Facts Argument

Game Seven occurred yesterday.
If Game Seven occurred yesterday then it is a fact that Game Seven occurred yesterday.
If Game Seven occurred yesterday then Game Seven has the property of having occurred yesterday.
The fact that Game Seven occurred yesterday is an indexical fact and the property of having occurred yesterday is an indexical property.
So, there are indexical facts and properties.

There are no indexical facts or properties, but the conclusion follows from the premises. Which premise is false?[7] The fourth line is false. It is a fact that Game Seven occurred yesterday, and the thought, the sentence, the claim, is indexical; but there is no indexical fact. Facts are not the right kinds of things to be indexical. Ways of stating facts are. Similarly, there are no indexical properties. There are indexical ways of picking out properties. And given that something occurred yesterday, there is a day, yesterday, on which it occurred. And "yesterday" is an indexical. But yesterday was not an indexical day. There are no indexical days. Days are not the right kinds of things to be indexical; day-naming expressions and concepts are.

"Normative," I'm claiming, is like "Indexical." It is not a feature of the way things are, but a feature of the way we talk about things, or the way we think about them. It is a way of getting at the world, rather than a part of the world. So, although we do use normative expressions to explain meanings, if Meaning is Normative, that is not because we are talking about special normative things, in addition to all the descriptive things we talk about. Normativity is a feature of the words and concepts we use when we attribute intensional attitudes and meanings, not a feature of the states of the mind or the relation between words and world.

What is at stake here? Is it just a quibble about terminology? Gibbard allows two different meanings for "fact" (Gibbard 2012, 68). So in one sense, the fact that I am here and the fact that Dreier is in Providence are two different facts: they are two different thoughts I could have, with potentially very different cognitive and practical significance. And in this sense, the fact that sex discrimination is wrong is a very different fact from any naturalistically expressed fact, for it is a different thought, a normative one, and it has a different cognitive and practical significance in our thinking. I do not use "fact" in this sense, but that is simply a disagreement in terminology. Gibbard's other sense of "fact" is *worldly*. Facts, as I am understanding them, are not individuated by the cognitive role of thoughts that have them as contents; they are what the thoughts are about, and one and the same fact can be thought of

7 The first premise expresses a truth in the context of typing.

in quite different ways (indexically, say, or third-personally). In this sense, the fact that I am here *is* the fact that Dreier is in Providence. And in this sense, the fact that sex discrimination is wrong *is* some fact that we could, if only we knew and had enough vocabulary, express in purely descriptive language.

Is there a deflated notion of a normative fact? There is the notion of a normative thought; we might say, a normative fact is simply a fact that one has in mind when one thinks a normative thought. (Presumably, this is one of the senses that Gibbard has in mind.) But, even for deflation-minded philosophers, it should be clear and reasonably uncontroversial that according to some views of normative language, there is nothing normative about the fact itself; that is surely demonstrated by the analogy with indexical thoughts. There are, certainly, facts that one could have in mind by thinking about what happened yesterday. But those facts could equally be considered by nonindexical thoughts. And analogously, an expressivist can say that although there are facts expressed by normative sentences, they could equally well be expressed by nonnormative sentences, if only we could be sure which ones.

Before I conclude this section, here is a clarification. The analogy between indexical thought and normative thought is *an analogy*. I am not here suggesting that normative thought *is* indexical. That is not a view that comports well with expressivism. It fits better with a contextualist view, like the one presented in Velleman (2013) or Dreier (1990). The point of the analogy is to give a clear and uncontroversial example of a feature of thought and language, a feature of a sentence (or predicate) that is not inherited from a feature of the fact in the world (or property) that the sentence (or predicate) picks out.

We worried that by the explanationist criterion, the normativity of meaning would entail that expressivism is a form of realism. But, we can now see, this need not follow after all. Expressivism can agree that the sentences it uses to say what it is in virtue of which people think normative thoughts are normative sentences. It need not say that there are any normative facts that are essential parts of the explanatory structure of our world. Just as what makes it the case that yesterday it was sunny here is an ordinary fact with no indexicality built into it, so what makes it the case that Allan believes that sex discrimination is wrong is a plain old fact with no normativity invested in it.

6. Is Meaning Really Normative?

I claimed the solution in section 5 above is the only one, given the assumption. But the assumption may be false. Is meaning really normative?

Some objects and properties are, loosely speaking, constituted by norms. To have the property of *being the president of the United States* is, in part, to have the authority to veto laws, the normative power to command the armed forces, and so on. To be a bishop, in chess, no particular shape or material constitution is necessary or sufficient; rather, a bishop *may* move along unobstructed diagonals, capture an opposing piece and land on its square, and *must* remain where it is when pinned orthogonally to the king, and *cannot* move to a

different-colored square. There is nothing more to the property of being the U.S. president, or to bishops, than is laid out in the rules in the institutions that define them. The U.S. Constitution did not encounter U.S. presidents and impose new restrictions on them and give them new powers: it created the status of being president by its rules (compare Hart 1961). And similarly chess players did not discover bishops and decide to impose new restrictions on them.

There is a sense in which statements about bishops or presidents, then, are normative. The judgment that something is a bishop, we might say, just *is* the judgment that it may be moved in certain ways and not in others, which square it starts on, and so on; similarly for the judgment that someone is the president. If someone told us he had a bishop at home that could move orthogonally, or disputed the power of our bishop to move more than one square at a time, we would think there was a misunderstanding: perhaps he is talking about a different game. (Things are trickier with presidents, since the particulars of their legal authority may be in dispute.) To understand what these things are is to understand how they fit into a system of rules.

It is, to my mind, very plausible that Meaning is Normative in this sense of normative. It is plausible that when we say Kurt means addition by "+," we are immediately committed to the conclusion that he ought to accept instances of $x + y = y + x$, in the way that judging that the pointy-topped spindle of wood is a bishop commits us to the conclusion that it must move diagonally. And someone who agreed that Kurt means addition but disputed our contention that Kurt is correct in saying "my age + 5 = 70" only if the sum of his age and five is seventy, would make us think there had to be a misunderstanding.

Here is another example that is a bit different. There is a usage of "murder" according to which only wrongful killings are murders, where "wrongful" is moral rather than legal. Or consider "theft" in like usage: this is the usage according to which we translate Proudhon's slogan as "Property is theft!" (1840, *What Is Property?*). Property cannot itself be *unlawful* taking, but Proudhon was saying that under capitalism it is *immoral* control over what *rightfully* ought to be controlled by others, or uncontrolled.[8] So, part of what it is for an act to be theft is for it to be wrongful, at least *prima facie*. Theft is normatively constituted; calling an act a theft is making a normative assertion.

But why do I say this example is different? I think there is a sense in which "theft" is fully normative and "bishop" and "president" are not. One way of spelling out this sense adverts to *motivational internalism* about moral judgment: to think of something as theft, you have to have some motivation to avoid engaging in it, whereas thinking of something as a bishop need involve no resolve or motive to keep it always on the same color squares. For you might not be interested in playing chess at all, or be playing but only out of interest

8 This is still a tricky example, since we might think Proudhon was using figurative language, or inventing a new variant sense of "theft," or really "vol" (in French). I could use "promise" as my example instead, but that one also comes with some controversial philosophical assumptions.

to see whether the person you're playing with is paying attention. I don't think the idea of "full normativity" here depends on the truth of motivational internalism, though; that is only one way of cashing it out. (See Copp (1995) for an externalist interpretation.) Some philosophers say, the rules of chess tell us how to move the pieces, but only conditionally on our intending to play chess, or on our playing chess; they are *hypothetical* imperatives. And some say that the laws of the United States purport to have genuine authority over us but that it is a difficult, substantial question in the philosophy of law whether they in fact have any, whereas moral law, if it is to be anything at all, must come with such authority (Hampton 1998). One way or another most metaethicists recognize the distinction between purported normativity, a kind that is gripping only once one accepts a given framework, and genuine normativity, which has its force in a more categorical way.[9]

Which kind of normativity is the normativity of meaning?

One test for genuine, robust normativity is whether disagreement appears to survive full understanding. Suppose I say that rooks cannot promote to other pieces, and you say they can promote to dragon kings. We seem to disagree over the powers of rooks. But after a few moments of discussion, it emerges that you are a shogi player and you're talking about shogi rooks; I'm talking about classic chess.

Now when you promote your rook to a dragon king, I can say, "You aren't playing chess!" I suppose I might think you are playing a worse game. I might think you *shouldn't* play shogi. But I couldn't think you are making a mistake in promoting your rook. If I insisted, "No, you can't do that!," I would just be confused. Our apparent disagreement over the powers of rooks would either disappear or transform into some other disagreement (over the relative merits of our games).

A disagreement over the powers of the presidency would not be quite like that. It could be an interpretive disagreement: we could disagree over how expansive the executive war power is, for example, as many have. Nor would we think our disagreement was "merely verbal," as if it could be fully resolved by simply stipulating that the word "president" is ambiguous. It might best be thought of as a dispute within metalinguistic negotiation (see Plunkett 2015; Plunkett and Sundell 2013), or a moral dispute over which more specific concept of the presidency we ought to use to fill in the open texture of the constitution (Hart 1961).

Finally, we might have a dispute over whether it is murder to withdraw life-sustaining support from a patient on the basis of testimony of close relatives that the patient, who is now unconscious, had competently expressed the desire to have support removed under just these circumstances. Our disagreement would be over the wrongfulness of the killing. It would not go away in the face of full explication of our respective positions.

The disagreement over normative attributions of meaning seems *to me* most like the first kind of disagreement. When I have explained that Kurt's own dispositions and considered

9 But not all; I suspect Finlay (2014) is an example of a metaethicist who finds the idea of what I'm calling the genuine kind just baffling.

judgment will or would produce his *adding* two numbers when asked about "+," and you have pointed out that even though this is true his linguistic community all agree in *quadding* two numbers, and we see that the difference between the two practices only shows up for pairs of very large numbers, it does not seem that there is much of anything left to dispute. What *should* Kurt say? Well, if he is *adding* he should say one thing, and if *quadding* he should say the other; according to one rule for "going on in the same way" he must add, and according to the other he must quadd. What should he *really* do if he is going to be true to his own *meaning*? Again, it seems to me there are just two frameworks for evaluating. There are, we might say, $\text{meaning}_{\text{individual}}$ and $\text{meaning}_{\text{community}}$. Staying true to $\text{meaning}_{\text{individual}}$ amounts to adding; staying true to Kurt's $\text{meaning}_{\text{community}}$ would mean quadding. One who (like me) is not inclined to see any further disagreement here is thereby skeptical about the full-bodied, robust Normativity of Meaning.

7. Conclusion

If Meaning is Normative, then expressivism is harder to distinguish from realism. But there is this difference: in its explanation of what we are doing in our normative judgments, realism adverts to (purported) normative properties; expressivism instead explains the role of our judgments and in the course of so doing mentions ordinary, "prosaic" properties by means of normative words and concepts. Just as Explanationists expect, this difference amounts to a difference over what sorts of things appear in more fundamental explanatory layers of the world.

References

Ayer, Alfred J. (1952). *Language, Truth and Logic*, 2nd ed. New York: Dover.

Blackburn, Simon (1984). "Supervenience Revisited." In *Exercises in Analysis: Essays by Students of Casimir Lewy*, edited by Ian Hacking. Cambridge: Cambridge University Press.

Chrisman, Matthew (2008). "Expressivism, Inferentialism, and Saving the Debate." *Philosophy and Phenomenological Research* 77, no. 2: 334–58.

Copp, David (1995). *Morality, Normativity, and Society*. Oxford: Oxford University Press.

Davidson, Donald (1985). "Incoherence and Irrationality." *Dialectica* 39, no. 4: 345–54.

Dreier, Jamie (1990). "Internalism and Speaker Relativism." *Ethics* 101, no. 1: 6–26. https://doi.org/10.1086/293257.

——— (1992). "The Supervenience Argument against Moral Realism." *Southern Journal of Philosophy* 30, no. 3: 13–38.

——— (2002). "Meta-Ethics and Normative Commitment." *Philosophical Issues* 12: 241–63.

——— (2004). "Meta-Ethics and the Problem of Creeping Minimalism." *Philosophical Perspectives* 18, no. 1: 23–44. https://doi.org/10.1111/j.1520-8583.2004.00019.x.

——— (2015). "Explaining the Quasi-Real." *Oxford Studies in Metaethics*, edited by Russ Shafer-Landau, 10: 273–97.

Fine, Kit (2001). "The Question of Realism." *Philosophers' Imprint* 1: 1–30.

Gibbard, Allan (2003). *Thinking How to Live*. Cambridge, MA: Harvard University Press.

——— (2012). *Meaning and Normativity*. Oxford: Oxford University Press.

Golub, Camil (2017). "Expressivism and Realist Explanations." *Philosophical Studies* 174, no. 6: 1385–409.

Hampton, Jean E. (1998). *The Authority of Reason*. Cambridge: Cambridge University Press.

Hart, H. L. A. (1961). *Concept of Law*, 1st ed. Oxford: Clarendon Press.

Kripke, Saul A. (1982). *Wittgenstein on Rules and Private Language: An Elementary Exposition*. Reprint edition. Cambridge, MA: Harvard University Press.

McPherson, Tristram (2018). "Authoritatively Normative Concepts." *Oxford Studies in Metaethics*, edited by Russ Shafer-Landau. 13: 253–77.

Moore, G. E. (1993). *Principia Ethica*, 2nd ed., edited by Thomas Baldwin. Cambridge: Cambridge University Press.

——— (2005). *Ethics*. British Moral Philosophers. Oxford: Oxford University Press.

O'Leary-Hawthorne, John, and Huw Price. 1996. "How to Stand up for Non-Cognitivists." *Australasian Journal of Philosophy* 74 (2): 275–92. https://doi.org/10.1080/00048409612347251.

Parfit, Derek (2011). *On What Matters*. Berkeley Tanner Lectures. Oxford: Oxford University Press.

Plunkett, David (2015). "Which Concepts Should We Use?: Metalinguistic Negotiations and the Methodology of Philosophy." *Inquiry: An Interdisciplinary Journal of Philosophy* 58, no. 7–8: 828–74.

Plunkett, David, and Timothy Sundell (2013). "Disagreement and the Semantics of Normative and Evaluative Terms." *Philosophers' Imprint* 13: 1–37.

Prior, A. N. (1960). "The Autonomy of Ethics." *Australasian Journal of Philosophy* 38, no. 3: 199–206.

Rosen, Gideon (2017). "Scanlon's Modal Metaphysics." *Canadian Journal of Philosophy* 47, no. 6: 856–76.

Russell, Gillian (2010). "In Defence of Hume's Law." In *Hume on Is and Ought*, edited by Charles Pigden. Hampshire: Palgrave MacMillan.

Scanlon, Thomas M. (2014). *Being Realistic about Reasons*. New York: Oxford University Press.

Schroeder, Mark (2007). *Slaves of the Passions*. Oxford: Oxford University Press.

Schroeder, Timothy (2003). "Donald Davidson's Theory of Mind Is Non-Normative." *Philosophers' Imprint* 3: 1–14.

Sturgeon, Nicholas L. (2006). "Ethical Naturalism." In *The Oxford Handbook of Ethical Theory*, edited by David Copp. Oxford: Oxford University Press.

Thomasson, Amie L. (2014). *Ontology Made Easy*, 1st ed. New York: Oxford University Press.

Velleman, James David (2013). *Foundations for Moral Relativism*. Cambridge: Open Book Publishers.

Wright, Crispin (1988). "Realism, Antirealism, Irrealism, Quasi-Realism. Gareth Evans Memorial Lecture, Delivered in Oxford on June 2, 1987." *Midwest Studies in Philosophy* 12, no. 1: 25–49. https://doi.org/10.1111/j.1475-4975.1988.tb00157.x.

Yalcin, Seth (2007). "Epistemic Modals." *Mind* 116, no. 464: 983–1026.

VI

CONSEQUENTIALISM

20

GIBBARD ON RECONCILING OUR AIMS[1]

Connie S. Rosati

In Allan Gibbard's 2008 Tanner Lectures, *Reconciling Our Aims*, Gibbard claims that the primary questions we biological creatures face are "questions about what to do, what to aim for, and how" (Gibbard 2008, 15).[2] But we also face questions about what we *ought* to do, what we *ought* to aim for, and how we *ought* to proceed. More generally, we think not simply about how to live but how we *ought* to live. As Gibbard sees the latter, ought-questions, they can be understood as the former questions, and the former questions are all planning questions.[3] "Questions of moral right and wrong," for example, are just questions about what to do, "with a particular kind of emotional flavor" (16). "Narrowly" moral questions about right and wrong, like all normative questions, are planning questions that concern what to do, think, or feel: more specifically, they are planning questions that concern what moral sentiments to have about acts or agents (17).

Gibbard's lectures thus begin, in part, by summarizing ideas developed more fully in *Thinking How to Live* (Gibbard 2003). But he goes on to argue that his metaethical account of our moral inquiry, though consistent with any coherent answer to our normative questions, may make some answers "more plausible than others" (33). In particular, he suggests, it may make utilitarian answers more plausible than nonutilitarian answers.

---

1 Thanks to David Plunkett, for helpful comments on an earlier version of this essay. And thanks once more to Allan Gibbard—a wonderful mentor and thesis advisor during my years in the University of Michigan philosophy graduate program, as well as a model of deep, imaginative, and respectful philosophical inquiry.

2 All parenthetical page numbers hereafter are to Gibbard (2008) unless otherwise indicated.

3 See Broome (2008, 105–7) for some discussion of why these questions aren't the same.

Gibbard's account of our moral inquiry and its plausible upshot is characteristically novel and in keeping with his long efforts to understand human beings and their normative questioning as part of the natural world. My aim in this essay is to explore a number of important ideas from Gibbard's lectures. In particular, I want to explore and raise puzzles for Gibbard's account of moral inquiry and his attempt to effectuate a connection between his metaethics and the kind of utilitarian normative theory that he has long favored. As I shall explain, Gibbard's account doesn't seem to capture either the phenomenology or the normativity of moral inquiry, and even if his account of moral inquiry were correct, Gibbard's story about how that account might favor utilitarian answers to our normative questions may only be persuasive to those who are already inclined to accept utilitarianism.

Moral Inquiry and the Scientific Picture of Us

Gibbard seeks to situate our moral inquiry within a scientific picture of us. We should look closely, then, at what he takes to be the relationship between that scientific picture and our moral questions. Consider the following passage from Lecture I:

> Suppose I settle on helping a man in need even though I won't get any advantage from it. My coming to this conclusion must be part of any naturalistic, biological story of me. The story, though, won't contain any fact that I've got my conclusion right or not. It doesn't contain a fact that I ought to help or that it's okay not to. It doesn't contain a fact that it would be wrong not to help or that it wouldn't be. Questions of what I ought to do and what it would be wrong to do or not to do aren't questions amenable to science. They are, I have been saying, questions about whether to help and of how to feel about not helping. A scientific picture, then, has us asking questions that don't have scientific answers. The picture shows too why these questions aren't luxuries, but must be central questions for us. And from a scientific picture comes an account of what these questions are: they are questions of what to do and how to feel about things people do or might do. If these are the questions, then we don't need to worry that they concern queer goings-on that form no part of the fabric of the universe, as John Mackie puts it. They are intelligible questions, and they are questions of first importance. (18)

But how does the scientific picture show why the questions of what to do, what to aim for, and how must be central questions for us? And how does the account of these questions come from a scientific picture of us?

As Gibbard presents the scientific picture of us, it begins with a common but controversial background story that he cannot undertake to present fully or elaborate on in his lectures. That story, he says, attempts to answer the puzzle of how "metaphorically selfish genes come to make people . . . who are not entirely selfish" (14). Proliferation of genes depends on forming organisms that can "keep track of the world around them," including the social world. But this "keeping track of," or knowledge of the world, guides action in

ways that proliferate genes "only if actors have the right aims, the right propensities to use their knowledge to guide action" (14–15).[4]

Beings like us thus face questions about how things are—about what the world in which we must act is like—but the primary questions we face concern "what to do, what to aim for, and how" (15). In contrast to other species, we have evolved with particularly intricate and refined emotional propensities, some of them "intensely social," and these emotions provide impulses to action. Gibbard speculates that we are adapted to respond to social situations with certain kinds of emotions, in particular, the moral emotions of "resentment or outrage and of guilt, guided by judgments of fairness" (15). Our emotions "affect reproduction through the actions they prompt, and so natural selection will shape the psychic mechanism of emotions" (15); our emotional tendencies evolved as they did because of this. And we evolved not only to have the emotional tendencies that we do but also to have "a kind of language infused governance of emotions" (15). Our linguistic capacity enables us to discuss with one another what to do and how to feel, as well as to engage in thought that operates (more or less effectively) to control our feelings (15). We may feel a certain emotion, such as resentment, while judging that our feeling is not warranted. Our judgments of warrant come into play in our narrowly moral judgments. To think an act morally wrong is to think the act warrants guilt on the part of the agent and resentment from others. Our moral questions thus concern whether and when the peculiarly moral sentiments are warranted; and our moral inquiry concerns how to live and feel, and how to engage emotionally with people and their actions (20).

We can judge that our feelings are unwarranted. But what do our judgments of warrant amount to? Gibbard suggests that we can understand how our reasoning with a concept like WARRANT works if we "think of judgments of warrant as something like plans" (16). Although we cannot choose what to feel, our feelings are often responsive to our judgments of warrant, and so it is "somewhat as if we had planned what to feel, even though choice doesn't figure in the guidance of emotion in the way that plans for action get realized by guiding choice" (16).

The picture Gibbard presents in the first of his Tanner Lectures is understandably brief; he has elaborated on it more in other writings.[5] Still, the sketch he offers is somewhat puzzling. Consider an example Gibbard provides of how our emotions, though not chosen by us, are nevertheless responsive to judgments of warrant. Suppose that you take the last piece of cake, which I had expected to enjoy. I might, Gibbard writes, feel a "flash of resentment." But then I might consider that you were equally within your rights to take the piece of cake, and upon so reflecting, my resentment may subside. My judgment of warrant, Gibbard proposes, is like a plan not to feel resentment in my circumstances.

4 Gibbard illustrates the point by observing that a person's genes won't proliferate if he knows where a lion is but responds by trying to pet it. Gibbard doesn't otherwise explain what he means by "right" aims and propensities, but given the context, he presumably means right from the standpoint of gene proliferation.

5 See Gibbard (1990, 2003).

One difficulty with the example is that the consideration that another was "equally within her rights" is itself a normative judgment—indeed, a moral judgment—yet it doesn't seem to be akin to planning. My judgment that another was equally within her rights may be the basis or ground for concluding that my resentment is not warranted, but even if my judgment of warrant is like a plan, the ground for it is not. Our judgments of warrant, as in the example, seem to depend on supporting considerations that are not themselves like planning.

A second and more obvious difficulty is that, having already resented you, my judgment of warrant, insofar as it is like a plan, can only amount to a plan not to have like feelings in relevantly similar circumstances in the future. Plans are, after all, future-oriented, even if we can enact a plan only by presently guiding ourselves by it, and here, the judgment of warrant, and so the plan, comes after the fact. Even if my judgment of warrant amounts to a plan for future, similar circumstances, it's bound to be a failed plan, for I cannot choose my emotions, even if I can, so to speak, talk myself down. So my judgment of warrant seems less like planning, even "as if" planning, and more like self-regulation; I don't plan what to feel but, rather, regulate the feeling I already have. And when I (and we) do regulate our feelings, it appears that, even in Gibbard's example, we do so in response to distinctively normative judgments that, as already noted, seem not like plans; my resentment isn't warranted *because* you were *just as much within your rights* to take the cake as I. Judgments of warrant enable a certain kind of self-regulation, it seems, because of their peculiar normative content. In judging that my resentment of you isn't warranted, I judge that it isn't fair of me to feel this way toward you. Perhaps Gibbard would think the latter judgment, too, amounts to planning not to feel resentment in like situations. But that seems to miss the phenomenology of the judgment—the way in which it focuses not on my future possible feelings but on the wrongness of my *present* feelings.

Suppose that when I felt that flash of resentment, I had instead reasoned like this: I ought not to feel resentment, because if I do, I may act out, and then people will think I'm a jerk, and that will make things harder for me. Or suppose I had reasoned like this: I'm already upset and disappointed about not getting the cake, but this resentment is only making me feel worse![6] These thought processes might work to dissipate my resentment just as well as my judgments of warrant, yet they have a decidedly different flavor. Now, I take it that plenty of people do think in these alternative ways and that thinking in these ways, rather than in terms of judgments of warrant, may even come more naturally to them. The scientific picture says that we have evolved to have complex emotional propensities, and in particular, propensities to feel the moral emotions. And we have acquired through evolution capacities for language-infused governance of these emotions. But nothing in the scientific picture as yet explains why the latter capacities would operate through judgments of warrant, rather than through judgments of self-interest. It does not yet explain why the moral emotions would be "guided by judgments of fairness," rather than by judgments of how one's feelings and resultant actions might affect one's social status. Presumably, we should expect it to explain both.

6 See Bittner (1992). Bittner offers considerations like this against the rationality of regret.

In short, the scientific picture doesn't yet explain why the questions of what to do, what to aim for and how, in their distinctively moral sense, are so critical for us, or why they are any more critical for us than questions about self-interest or social status. The difficulty is that the connection between our moral ought judgments and judgments of warrant, as opposed to our self-interest or social status judgments, seems conceptual rather than contingent. Philippa Foot argued in her famous criticisms of emotivism and prescriptivism that we cannot make moral judgments about just anything (Foot 1958a, 1958b). Foot observes that moral concepts, such as RIGHTNESS, OBLIGATION, DUTY, and GOODNESS are related to such concepts as HARM, BENEFIT, and IMPORTANCE (Foot 1958a, 511). A view like R. M. Hare's supposes that "if we describe a man as being for or against certain actions, bringing them under universal rules, adopting these rules for himself, and thinking himself bound to urge them on others, we shall be able to identify him as holding moral principles, whatever the content of the principle at which he stops" (Foot 1958a, 512). But such a view has absurd consequences,

> for it follows that a rule which was admitted by those who obeyed it to be completely pointless could yet be recognised as a moral rule. If people happened to insist that no one should run round trees left handed, or look at hedgehogs in the light of the moon, this might count as a basic moral principle about which nothing more need be said.[7]

Gibbard has elsewhere addressed the kinds of claims that Foot makes, arguing against what he calls a "direct, substance-constrained account" of the meaning of 'morally wrong' and in favor of his own indirect, "sentiment-routed" account, which understands our judgments of moral wrongness in terms of judgments of warranted blame (Gibbard 1992). He contends that the latter account, in contrast to the former, can allow for and explain various normative disagreements. Gibbard's rich and complex argument proceeds by attempting to construct the view with which he disagrees in the strongest form he can devise and then testing it against his own view. As I can't undertake to respond to Gibbard's argument in detail here, I simply acknowledge our disagreement.

Notice that the scientific picture is consistent with alternative views. It might help to see this if we fill in what must surely be one more part of the scientific picture of us. For not only has evolution produced creatures with complex emotional capacities and language-infused governance of emotion, but it has also equipped those creatures with complex capacities for reasoning and a disposition to be responsive to standards of logical inference and claims of rationality. Just as our knowledge of the world may guide action in a way that proliferates genes only if we have the right aims and propensities to use our knowledge to guide action, it may also do so only if we have the ability to draw logical inferences, estimate probabilities, and recognize when situations are relevantly alike or different and how they are alike or different.

---

7 Ibid., 512.

For example, our evolved capacities for gaining knowledge about our environment are such that, once we have them, we have acquired general abilities to learn and reason and so to learn much about the world that may have no particular bearing on gene proliferation, at least not in any direct way, such as abstract mathematics and theoretical physics. The same abilities that explain how we can acquire the latter sort of knowledge might well explain how we can acquire moral knowledge. We can imagine a number of ways the story might go. For instance, the same abilities that enable us to grasp conceptual and a priori truths about mathematics enable us to grasp conceptual and a priori truths about ethics. Alternatively, we can gain knowledge in the sciences and mathematics and in ethics only by having rational capacities that enable us to discern relevant similarities and differences and that prompt us to treat like cases alike. Although our judgments of fairness aren't exhausted by our judgments about whether cases are relevantly similar or not, and so to be treated alike or not, they commonly involve such judgments. Our rational capacities give us a propensity to be responsive to considerations like this, much as they give us a propensity to be responsive to norms of logical inference.

Now, Gibbard maintains that "even claims about rationality in science aren't entirely within the subject matter of science" (15). If he is right, then science's subject matter would seem to include neither claims about rationality in science nor claims about oughts or warrant or fairness in ethics. But insofar as we allow that we nevertheless have knowledge of scientific truths, it seems we would have no reason, without more, to disallow that we have knowledge of moral truths. Those moral truths needn't reside in the natural world any more than truths about mathematics. In maintaining that our ought judgments and judgments of warrant or fairness are aimed at discerning and expressing moral truths, we need not be committed to queer goings-on as part of the fabric of the universe, any more than we need be so committed in maintaining that our logical judgments and judgments of rationality are aimed at discerning and expressing scientific or mathematical truths. The differing stories we might tell may fit equally well with a scientific picture of us as creatures who have evolved with complex emotional propensities and the capacity for language-infused normative governance, and none of them require that we posit queer entities of the sort Gibbard and Mackie would advise us to avoid. My point, of course, is not that one or another story is correct but only that the scientific picture seems to be pretty much silent as to whether our asking what we ought to do or feel amounts to, or might best be understood as, our asking (and planning for) what to do or feel.[8] We can thus accept Gibbard's background story and some of his assumptions about moral ontology, while resisting his suggestion that the scientific picture of us shows why our ought-questions, as Gibbard understands them, must be central questions for us, or that these questions come from that scientific picture of us.

I remarked earlier that our judgments of warrant or of fairness have a different flavor than our judgments of self-interest or social status. Imagine that I feel that flash of resentment

8 But see Street (2006) for a critique of realist theories of value based on the scientific, evolutionary picture of us. And for criticism of debunking arguments against realism, see, for example, Vavova (2014).

toward you; then I judge that I had better get a grip on myself or people might think I'm a jerk. Later in the evening, I tell a mutual friend that I had resented you for taking the last piece of cake but then reminded myself that I should get my feelings under control, so people won't think I'm a jerk. I speculate that mutual friend might well conclude that I am, indeed, a jerk. After all, I failed to consider the real reason why I shouldn't resent you. Or suppose that I am feeling that flash of resentment and sheepishly confess my feelings to a mutual friend. My friend tells me to get those feelings under control so that other people won't think I'm a jerk. Maybe his saying this would lead me to reflect on what he said, thereby diminishing my resentment. But is that what *should* diminish it?

We do ask what (we ought) to do and aim for and feel, and we engage in moral discussions with one another about what (we ought) to do and aim for, how to feel, and so on. We reflect within ourselves and exchange reasons among ourselves. But why do our discussion and reflection have their peculiarly normative flavor? Why do judgments of warrant or fairness tend to come so naturally to us? Why do they often shift in ways that we seem to recognize to be improvements? Perhaps our moral questions, in some sense, come from the scientific picture of us, and perhaps that picture shows why they must be central for us, but not necessarily as Gibbard would have us understand those questions.[9]

Thinking What (Morally) to Do

The scientific picture of us biological creatures, Gibbard maintains, has us asking questions about what to do and feel, and it provides "an account of what these questions are" and of why "they must be central questions for us." Yet that picture contains no facts that would determine our answers to them to be correct. "Questions of what I ought to do and of what it would be wrong to do or not to do aren't questions amenable to science" (18). How, then, does thinking what to do proceed?

> Thinking what to do can go in two stages: In the first stage I form my valences or preferences. In the second stage, if there is more than one thing I equally or most prefer from among my alternatives, I pick one—not out of preference, but out of necessity to choose if I am not to be like Buridan's ass. My strictly normative thinking is a matter of the first part. We could call this part concluding what's "okay" to do and what isn't. When it is okay to do something and not okay not to do it, then I *ought* to do it.[10]

---

9 Gibbard (1990) offers an expressivist account of normative judgments, and of our more narrowly moral judgments, and their role in coordinating feeling and action. He has more to say in that earlier book to indicate the answers he might be inclined to give to these questions, and we may or may not find those answers persuasive. In any case, Gibbard does not, in his Tanner Lectures, directly address them.

10 Gibbard (2008, 19). Gibbard goes on to suggest that there is a third, interpretive stage, but I leave that to one side because it isn't important for present purposes.

The valenced stage, Gibbard maintains, is the stage of thinking what I ought to do. To believe that one ought (or that it is "okay") to Φ is "to rule out preferring any alternative," to "rule out a kind of valence." Our normative judgments, he says, "consist in valences and restrictions on valences" (20).

Our moral thinking what to do or feel does involve intuitions, Gibbard thinks. But these are not the intuitions of nonnaturalism—purported direct apprehensions of moral truths, at least subject to certain constraints.[11] Neither are they the nonnormative intuitions that figure in psychology, the sort that Jonathan Haidt writes about. These "de facto" intuitions are judgments a person makes confidently and on no further ground. The normative or "de jure" intuitions Gibbard writes of are "states of mind of accepting something," though not based on further reasoning. "To think something an intuition in this sense is to plan to rely on it." De jure intuitions are "de facto intuitions to rely on" (23).

Gibbard observes that we can ask a planning question about when to trust our own planning, thereby asking "what conditions are ideal for planning" (23). These might include full information and the sorts of conditions common to ideal observer and ideal adviser theories.[12] It is a psychological question what we would plan were we in those conditions (or better, want ourselves to plan for ourselves in our actual conditions). The answer we would arrive at to the psychological question, however, can provide an answer to the normative question of what we ought to do, though not as a matter of logical entailment. Rather, "in calling conditions ideal for judgment, we mean that judgments in those conditions are ones to trust" (2008, 24). To accept the claim that in ideal conditions, "I would judge that Jones ought to Φ," is equivalent to accepting the claim that "Jones ought to Φ is a judgment to trust," and to accept that claim is to plan to trust the judgment. To trust that judgment is to follow through on the plan and, thus, to make the corresponding judgment, "Jones ought to Φ." Gibbard remarks that if we could settle on the ideal conditions—the conditions to trust for our normative judgments—then we could arrive at normative judgments based on interpreted empirical findings. Our doing so would not violate the "no ought from is" dictum, because there is an "intervening normative premise . . . that what I'd find wrong in those particular conditions *is* wrong—that what I'd then *think* ought to be done *ought* to be done" (24). Planning, Gibbard contends, "requires thinking that the *is* of interpreted psychology—that I implicitly accept an ought, and would accept it explicitly if challenged, on no further ground—supports accepting the *ought*" (26). Planning thus requires trusting intuitions, though that trust is defeasible.

There is a seeming tension in Gibbard's account. For if the stage of thinking what to do is the stage of a person's forming her valences or preferences, then it would seem that the

11 See Sidgwick (1981, 338–42, 400), discussing, respectively, "the four conditions the complete fulfillment of which, would establish a significant proposition, apparently self-evident, in the highest degree of certainty available" and the "two-fold" procedure for assessing common moral precepts

12 See, for example, Firth (1970), Railton (1986a, 1986b), and Smith (1994).

question of what I ought to do *is* amenable to science. After all, it amounts to the question of what I prefer, or would prefer (perhaps under ideal conditions), if I thought about it; and there seems to be a fact of the matter about that. I could be mistaken about my own preferences or valences, and so facts about what I do or would prefer would provide correct answers to my question. Perhaps what Gibbard means is that there are, and that science can provide, no valence- or preference-independent facts that would give correct answers to my question. But that is compatible, as ethical naturalists would maintain, with believing that our moral questions are amenable to science. And if thinking how I ought to live amounts to thinking how to live, where that comes to arriving at valences or preferences, then it would seem the question can only be, at bottom, about a valence- or preference-dependent fact. Unless we are ethical naturalists, which Gibbard says he is not, that would amount to understanding what we had thought was a genuinely normative question as a (merely) psychological question.

Another puzzle about the account concerns his appeal to the notion of trust. Gibbard tells us that to think something an intuition is to plan to rely on it, but it seems we can think something an intuition while wondering whether to rely on it, and so wonder whether it is something to trust. Now, to trust a judgment, in the ordinary sense of the word 'trust,' is to rely on its truth or accuracy. So, if in calling conditions ideal we mean that judgments in those conditions are ones to trust, then to treat judgments we would make under ideal conditions as ones to trust would be to rely on their truth or accuracy (or on them as true or accurate). But then how can we avoid an is-ought problem because of the intervening premise that "what I'd find wrong in those particular conditions *is* wrong—that what I'd then *think* ought to be done *ought* to be done"? After all, if those ideal conditions are not ones to rely on, then we would err in relying on them. Alternatively, if they are ones to rely on, then how are we to square this with Gibbard's contention that the scientific picture contains no facts that would answer our normative questions? Presumably, there are facts about what we would judge under ideal conditions—ones that are to be relied on for their truth or accuracy.[13]

More generally, Gibbard's account of moral thinking doesn't seem to capture either the phenomenology of thinking what we ought to do or the normativity of our judgments. Our moral thinking doesn't, at least in my experience, have the feel of forming preferences or valences; it has the feel of trying to figure something out, of trying to arrive at the right or most defensible answer. If it were the forming of preferences and valences, it would be hard to explain why our moral thinking, when we will be called upon to act, can be so difficult, why we fear making a mistake, and why we are subject to feelings of guilt and remorse when we later conclude that we have erred. And even if accepting the claim, "I would judge that Jones ought to Φ" is equivalent to accepting the claim that "Jones ought to Φ is a judgment to trust," the normativity of judgments about what we ought to do is captured only if

13 Gibbard (2008, 29) says that answers to our ought or planning questions "may in the end count as true or false" but that our initial theorizing shouldn't invoke the notions of truth or falsity.

accepting either claim is itself defensible, and that depends on whether judgments formed in ideal conditions really are ones to trust.

Engaging in Moral Thinking

In the second and third of Gibbard's Tanner lectures, he moves on from characterizing moral thinking to, as he says, *engaging* in it. Although Gibbard acknowledges that the metaethical picture in Lecture I is compatible with "any coherent answer" to our normative questions, he wants to explore whether it makes some answers "more plausible than others" (33). His line of argument covers a variety of issues, but the upshot of it is that John Harsany's two welfare theorems seems to show that the only coherent ethical theory is utilitarian. The implication would therefore seem to be that thinking how to live, at least if we are coherent in our thinking, will lead us to utilitarianism.

Gibbard considers how our moral thinking might go if it is, as he claims, planning how to live with each other. He describes two main "styles" of moral inquiry in philosophy. He depicts the first, utilitarian style, as "humanistic and pragmatic, thinking what's in morality for us . . . and asking what version of morality best serves us" (34). It looks for a value ethics serves that "can be appreciated in non-ethical terms" (34). The other, contractarian, style looks to achieve consistency among our (possibly revised) intuitions and "embraces what emerges" (34). The hope, he claims, is that these styles will ultimately converge, at least when carried out in the right way. Now, one might take issue with Gibbard's characterization of the two styles. One might argue, for example, that the contractarian style is just as humanistic and pragmatic as the utilitarian style, that it just as well captures William Frankena's observation, which Gibbard applauds, that "morality is made for man, not man for morality."[14] After all, Rawls's contractarian thought aims to observe the "strains of commitment" and the need for stability, both pragmatic concerns. And one way to understand morality as being made for man is given by the contractarian picture of people making or "constructing" morality for themselves, through their own ideal reflection. But let's set this aside.

The Alleged Incoherence of Nonutilitarian Normative Theory

Gibbard looks, he says, for a kind of convergence between the two normative styles, but the structure of his approach seems rather different than that, if not unfamiliar. The argument, as we will see, begins with contractarian starting points about the point of morality and about the impartiality provided by deliberation under ideal conditions (for example, behind a veil of ignorance) and tries to establish that these starting points in fact lead to a substantively utilitarian outcome. But this way of supporting utilitarianism might appear to face a problem, for it seemingly affords contractarians a reply similar to a common utilitarian reply to

14 In one place, p. 34, Gibbard seems to acknowledge as much. But on p. 35, he writes, "If morality is for humanity, then we might expect utilitarianism to be right."

the objection that utilitarianism is self-defeating. Suppose it turned out, the objection goes, that what would maximize utility is for everyone to follow Kant's categorical imperative; then wouldn't utilitarianism be self-defeating? A common response is that utilitarianism offers a criterion of right and wrong, rather than a decision procedures, and so even if the theory had this result, it would not be *self-defeating* but merely, as Derek Parfit puts it, *self-effacing* (Parfit 1984, 40–43). No matter how (apparently) nonutilitarian the actions, rules, or principles that satisfy the utilitarian criterion, utilitarianism could still be the true moral theory.

Now, imagine a contractarian faced with a like objection. Rawls developed his contractarianism, or constructivism, after all, to provide an alternative to utilitarianism (Rawls 1971/1999, xi). But suppose it turned out that the principles rational contractors would agree to behind the veil of ignorance are utilitarian; then wouldn't contractarianism or constructivism be self-defeating? Here, the contractarian might reply that their theory provides a criterion of the correct normative principles or rules, not a decision procedure to be followed by individual agents. Therefore, contractarianism is merely self-effacing, not self-defeating, and so it could still be the correct moral theory.

Whatever the respective merits of utilitarian and contractarian responses to the charge of being self-defeating, what this initial puzzle suggests is that Gibbard must not be treating contractarianism as providing a criterion of right and wrong in the same sense that utilitarianism does.[15] So we should understand the structure of Gibbard's line of argument in a more nuanced way.

We might characterize it as *methodologically contractarian*, but substantively utilitarian in its upshot: Gibbard uses the tools of contractarian thinking to show that his metaethics makes utilitarian answers to our questions more plausible than nonutilitarian answers. The rough analogy I mean to make here is to Peter Railton's adoption of methodological naturalism that has ethical naturalism as its upshot (Railton 1993). According to Railton,

> Methodological naturalism holds that philosophy does not possess a distinctive, a priori method able to yield substantive truths that in principle are not subject to any sort of empirical test. Instead, a methodological naturalist believes that philosophy should proceed a posteriori, in tandem with—perhaps as a particularly abstract and general part of—the broadly empirical inquiry carried on in the natural and social sciences.[16]

We might describe Gibbard's methodological contractarianism as holding that moral philosophy does not have a distinctive *a priori* method able to yield substantive truths (contrary to the views of classical intuitionists), and that moral theorizing should proceed by thinking,

15 He might also think that the contractarian reply, unlike the utilitarian reply, is unsuccessful.

16 See Railton (1993, 315).

perhaps under idealized conditions, how to live with one another, so as to arrive at a plan for living with one another that no person could reasonably reject.[17]

Gibbard begins with a problem for utilitarianism—that it "conflicts with strong intuitions"—and then deploys a case that illustrates a "systematic argument for utilitarianism" (35). He informs us that he wants to "use the debate about that argument to explore how moral inquiry might proceed if it consists in doing the sort of thing I claim, in thinking how to live together" (35). The example that Gibbard uses to illustrate the problem for utilitarianism is adapted from Diane Jeske and Richard Fumerton and concerns a man on a riverbank who must choose among children to rescue when their canoes capsize. In the canoe closer to him are two children, not his own; in the farther canoe is one of his own two children. He can rescue the two children nearer to him or his own child, but he cannot rescue all. The strong intuition that utilitarianism seems to flout is that the man is morally permitted (or even required) to save his own child, rather than save the two children.

Gibbard asks us to imagine that in advance of this crisis, the two fathers had planned for such a contingency. What contractual agreement would they have reached? Suppose that the likelihood of being either father is the same and that any agreement reached would be kept. Then given that it is worse to lose two children than to lose one and that acting to save only the single child would mean that the father of the two nearer children stands to lose both, Gibbard maintains that they would agree to rescue as many children as possible. He thus illustrates how the contrarian and utilitarian styles can converge.

Insofar as the motive of fair reciprocity is sufficiently strong to ensure compliance, the agreement the fathers would thus reach would coincide with what utilitarianism prescribes. This shows, Gibbard maintains, that the utilitarian position can derive from motives of fair reciprocity, as well as from benevolence. But suppose that the unlucky father who stands on the riverbank and, per the agreement, must now rescue the other father's two children is *insufficiently* motivated by fair reciprocity (as the other father might be, were their positions reversed). In that case, the agreement would be ineffective. Given what Rawls called the "strains of commitment" a contractarianism that heeds those strains would instead permit rescuing one's own child, thus deviating from what utilitarianism would prescribe.

The motives of benevolence, fair reciprocity, and what Thomas Scanlon describes as "a concern to live with others on a basis no one could reasonably reject" might all coincide, at least in cases of full compliance, as the canoe case illustrates (39).[18] Moral inquiry concerns the planning question of how to live with others, and Gibbard remarks that "the ideals of fair reciprocity and of living with others on a basis that they could not reasonably reject seem

17 Of course, the analogy is loose, in that Railton's upshot is metaethical, whereas Gibbard's is normative. My point is that there is a way of understanding what Gibbard is up to that differs from simply adopting the contractarian framework, and that also differs from Sidgwick's (1981) move from his metaethical intuitionism to utilitarianism.

18 Citing Scanlon (1998).

good candidates for what to want in one's dealing with others" (39). For ease of exposition, I'll refer to these ideals from here on simply as *reciprocity* and *reasonableness*. Gibbard proposes, he says, to think with others who might be brought to share these ideals about how to flesh them out.

Utilitarianism is often criticized on the grounds that it fails to respect persons, but Gibbard remarks that examination of the canoe case "illustrates why coherent, non-utilitarian theories are so elusive" (39). I'm uncertain precisely why Gibbard thinks the case illustrates this. He takes it to show that when compliance is uncertain, nonutilitarian theories would permit people to act against the agreement that they themselves would otherwise accept, but I'm not sure why this would render the theories incoherent. Because these theories do not place total good ahead of respect for individuals or for individual rights, there doesn't seem to be anything incoherent about recognizing that respect for individuals may coincide with utilitarianism in conditions of full compliance, but not in conditions in which compliance can't be assured. Perhaps a contractarianism that heeds the strains of commitment will deviate from what utilitarianism prescribes in something like the canoe case, but what utilitarianism itself would prescribe depends, in part, on the strains of commitment. Utilitarians commonly defend their theory against claims that it violates our intuitions or that it is overly demanding by allowing that the demands utilitarianism makes on us must take into account facts about the limits of human motivation. If the strains of commitment really were so great that the fathers could not be guaranteed to act on their agreement, then surely even utilitarianism would have to prescribe a different course.

Gibbard argues, though, that if we use as our starting points reciprocity and reasonableness—both of which bear on respect for persons—we will be led to a utilitarian moral view. Taking reasonableness as "the point of morality," we can say that no one can reasonably object to a system they would have agreed to in fair conditions, and something like a Rawlsian veil of ignorance is one way to establish fair conditions (2008, 41–42).

The canoe case, he maintains, illustrates how our thinking can yield a utilitarian conclusion, and the economist John Harsanyi, using his important welfare theorems, showed how this result generalizes. Gibbard tells us that the first of Harsanyi's welfare theorems concerns something like Rawls's original position, which places choosers behind a veil of ignorance.

> Think of valid moral rules as the rules one would choose assuming an equal chance of being anyone. Assume one's preferences are coherent in that they satisfy the standard conditions. Then one will prefer the rules that yield the greatest total utility.[19]

Harsanyi's second welfare theorem is "a version of the prospective Pareto condition."

19 Gibbard (2008, 42). In the first lecture, Gibbard notes that "decision theorists have shown that if a way of ranking actions satisfies certain conditions, then it is as if the person chose by maximizing an expected value." By "standard conditions," Gibbard says he means any of those sets of conditions that give this result.

> Suppose that prospective individual benefit is coherent, and so is desirability from a moral point of view. Suppose also that morality is for humanity in at least the following sense: if one prospect is better than a second for each individual, it is the better prospect ethically. Then desirability from a moral point of view, he proved, is a weighted sum of individual benefits. The only way ethical evaluation could satisfy these conditions and depart from utilitarianism is by weighing one person's benefit more than another. (42–43)

Gibbard argues that Harsanyi's theorems pose a problem for nonutilitarian moral theories. A nonutilitarian will end up choosing rules that none of us would have chosen without knowledge of how we, and not all others, stood to benefit. And "any evaluation of the prospects that different moral orders bring must either (i) violate some demand of rationality, or (ii) weigh one person's utility above another's, or (iii) rank some prospect best even though another one prospectively benefits everyone more" (43).

Now, as Gibbard points out, he cannot go into the detail required to fully explain the theorems, their import, or the complex debate about them. Still, the argument is sufficiently clear to wonder whether it need be convincing to those not already persuaded of utilitarianism.[20] There are, after all, a variety of moral considerations that seem to weigh with people apart from benefit, such as considerations of justice and desert. Of course, the argument has us imagining what people would prefer or choose for their own sake behind a veil of ignorance. But what we would choose for our own sake, even under ideal conditions, needn't exhaust our concerns or what might reasonably enter into our thinking when we think together how to live on a basis that no one could reasonably reject. Even if we focus on what people would prefer for their own sake, it isn't clear that they would prefer or choose what utilitarianism would recommend. Suppose that a person, P, is behind a veil of ignorance and choosing moral rules or moral principles on the assumption that she has an equal chance of being anyone. P might prefer rules that yield the greatest total utility. But she might have different preferences. She might consider that so long as her basic needs and wants are provided for, she would like the possibility of acquiring more, should she turn out to be someone with the aptitude, persistence, and interest to acquire it, or she would like simply to be free to explore her interests as she will. She might prefer for her own sake that no matter who she turns out to be, certain sorts of trade-offs among people's preference not be permitted, even if allowing them would yield the greatest total utility. She might prefer some form of sufficientarianism over utilitarianism, or some system of rights together with a generous safety net. And so on. Of course, utilitarians have long made efforts to show that their theory can account for all or most of our apparently nonutilitarian preferences and concerns, but it would be fair to say that their efforts have not been conclusive.

Notice that in preferring some nonutilitarian result or rule or principle, a person needn't be adopting what none of us would have chosen for our own sake unless we knew that we

20 For a more technical assessment of Gibbard's argument, see Broome (2008, 109–19).

stood to benefit in a way other's wouldn't, for by hypothesis, she is behind a veil of ignorance and assuming that she might be anyone. So she isn't, or needn't be, weighing one person's utility above another's. Now perhaps in preferring nonutilitarian results, rules, or principles, she would violate some demand of rationality, though it isn't obvious what such a noncontroversial demand would be. And the criticism that she ranks some prospect as best, even though another one prospectively benefits everyone more, seems question-begging against nonutilitarians. To be sure, she ranks as best an arrangement that does not maximally benefit everyone in the utilitarian's sense, but that is just because, as a nonutilitarian, she has a different sense of what's important. Thus far, then, it seems to me that Gibbard hasn't made the case that the picture he presented in Lecture I makes some answers to our normative and moral questions more plausible than others, and in any case, he hasn't established that if correct, it shows utilitarian answers to be more plausible than nonutilitarian answers.

Rawls's argument, in *A Theory of Justice*, for the claim that persons in the original position would choose his two principles of justice to govern the basic structure of a society in which they would spend a lifetime, was criticized on the grounds that the assumptions he built into the original position already ruled out alternative results. I have suggested that Gibbard's line of argument seems to face a similar difficulty. If we can draw a moral from the criticisms Rawls's argument faced, it would appear to be that we will be hard pressed to argue that our thinking how to live with one another, under ideal conditions, will favor particular normative results without already assuming much of what we would, in our ordinary thinking, take to be up for debate.

A Plan for Living

As we have just seen, Gibbard argues that Harsanyi's theorems pose a problem for nonutilitarian moral theories in that their evaluations of different moral orders will violate a norm of rationality, weigh one person's utility above another, or rank as best a moral order that benefits everyone less than an alternative. The argument relies, then, on a notion of benefit. Gibbard argues, contrary to Harsanyi, that we "can't derive the notion of individual benefit from the preferences people have, or even the preferences they would have in ideal conditions" (47). He credits as quite serious an objection raised by Scanlon, according to which

> There is no one coherent notion . . . that will do the jobs that 'welfare' or 'a person's good' has been asked to do in much ethical thinking: roughly, determining (1) what a person will choose insofar as others aren't affected, (ii) what others concerned to benefit him will choose to promote, and (iii) what counts as a person's good for purposes of moral thinking. (49)

I think the problem Scanlon raises, at least as Gibbard presents, it, is much less serious than Gibbard supposes, at least in light of the contemporary literature on welfare or a person's

good.[21] Consider (i). When well-being theorists talk about a person's "nonmoral" good, they mean to distinguish between what is good *for* that person—what it makes sense to want for her *for her own sake*—and what is good *from that person's point of view*, which will include things she might want for the sake of others.[22] They thus attempt to circumscribe the notion of a person's good to address the scope problem raised by Mark Overvold against desire theories. As Overvold argued, if desire theories include all of an individual's intrinsic desires as ones satisfaction of which benefit that individual, then they render self-sacrifice conceptually impossible.[23] Take the well-worn example of the soldier who desires to throw himself on a grenade to save his platoon. Surely satisfying that desire isn't good for *him*. With regard to (ii), well-being theorists take a person's good, as thus described, to be what others concerned to benefit him ought to promote—after all, it's what they ought to want *for her sake*.[24] When it comes to moral thinking, almost every moral theory includes among its rules or duties a requirement to promote the well-being or benefit of others, and this pertains to the notion of a person's good just discussed. Acknowledging that a single notion of a person's good plays all three roles is perfectly consistent with allowing that our moral thinking may not be exhausted by thinking about a person's good. As Darwall has observed, we need to distinguish between attending to a person's good out of *concern* for her and not interfering, out of *respect* for her, with her (possibly self-sacrificing) pursuit of what is good from her point of view (Darwall 2002, 14–16).

Still, Gibbard does think Scanlon's objection serious. He observes that we are here asking about (iii), and we can't simply look to a person's preferences to determine her good for purposes of thinking how to live (49). We also, he says, can't use something like Rawls's "primary goods" in our moral thinking; Gibbard maintains, for reasons that he doesn't provide in the lectures, that Rawls did not supply "an adequate, defensible rationale for this solution" (50).

Gibbard proposes, in Lecture II, to "characterize a plan for living" that incorporates one version of the contractarian ideal of living with one another "on a basis that no one could reasonably reject" (51).[25] But a person could reasonably reject what she takes to be unfair, and this is largely a matter of disadvantages versus benefits. So, we still need a notion of benefit for purposes of moral thinking. As it turns out, he tells us, "forming a conception of benefit is *part* of ethical thinking, *part* of thinking how to live among other people" (47, emphasis added).

---

21 Scanlon (1998). Regarding both (i) and (ii), Gibbard frames the problem in terms of a person's actual choices, or the actual choices of someone who happens to care about that person. But most welfare theorist would reject such nonnormative characterizations of the notion of a person's good. Gibbard, however, regards questions about a person's good as planning questions.

22 See Darwall (2002, 15–16), explaining and emphasizing the importance of this distinction.

23 Overvold (1980).

24 Darwall (2002) offers an analysis of welfare in terms of what one ought to want for someone for her sake. But we needn't accept Darwall's analysis to think that a person's good is what we ought to want for her sake.

25 Gibbard makes clear that he isn't claiming that his is the only coherent way to work out the contractarian ideal.

Gibbard imagines the retort that might be made to one who objects to a contractarian proposal for how to live with one another: "but you'd have agreed to it!" The problem, he suggests, is that

> If we are to make sense of what we would have agreed to, we can't just look to our aims as they are as a result of the basic moral arrangements we have. The retort, if it is to have specific content, must be filled in with coherent fundamental aims we can take ourselves to have from a standpoint that doesn't just take us as we are. We must be able to look at the various sorts of people we might have turned out to be under various different social circumstances and ask how well these fundamental aims for oneself are fulfilled in these various kinds of lives. (50–51)

Begin with the imperative

> Prefer most to live with others on a basis that no one could reasonably reject on his own behalf. (51)

We must ask, Gibbard says, "what it is to reject a basis for living with each other *reasonably* and *on one's own behalf*" (51). He suggests the following:

> A rejection on one's own behalf of a going social arrangement is unreasonable if, absent information about which person one would turn out to be, one would have rationally chosen that arrangement on one's own behalf. (51)

But what is it to choose on one's own behalf? Gibbard says that we can express this in terms of a person's good.

> One chooses rationally on one's own behalf only if one chooses what is prospectively most to one's good. (52)

Gibbard contends that everyone, in the absence of information about who he is, would choose on his own behalf a social arrangement that "is most to his prospective good" (52). This plan for living together would be a plan "to maximize prospects for the sum total good of people" (52). Insofar as one prefers to live on a basis of mutual respect or in accordance with an ideal of reciprocity, "one prefers to do one's part in an order that maximizes the total good of people, provided that everyone else can be fully expected to do so" (54).

But what is the notion of a person's good that is to figure in what it is to choose rationally on one's own behalf, as well as in the resulting plan of maximizing the total good of people? According to Gibbard, "the question of what constitutes a person's good is a planning question" and planning terms like a 'person's good' cannot be given naturalistic definitions.

> The question of whether there is such a thing as a person's good is a planning question. It is a question of whether to live in a way that takes a certain form. I come to a view about what a person's good is, then, if and when I come to have preferences that take this form. We come to a joint view, in discussion, of what a person's good is if we all come to have preferences that take this form, and—crucially—for each of us the same valuations play the role these conditions assign to a person's good. (54–55)

Adverting to Scanlon's skepticism about there being a single notion of a person's good that can "play the comprehensive moral role of being what the correct moral theory tells us to distribute," Gibbard tells us that what counts as a person's good, on his view, is whatever fills this role (55).

I find puzzling the idea that "whether there is such a thing as a person's good" is itself a planning question "whether to live in a way that takes a certain form," and that coming to a view about what a person's good is involves "coming to have preferences that take this form." To be sure, living lives that we find satisfying requires entertaining and answering many planning questions, at least in our ordinary sense of this. Yet all of our planning rests on what doesn't seem to be a matter of planning or an answer to a planning question at all. Whether there is such a thing as a person's good sure seems to answer itself, and to do so directly, rather than by way of a question whether to live in a way that has a certain form and our coming to have preferences that take this form. And this surely is what we would expect, given the scientific picture of us biological creatures. Reflect on the nurturing required for human development and the sustenance required for continued life and activity. Consult your own experiences of pleasure and pain. Consider the importance in your own life of loving, interpersonal connections, and gratifying pursuits.

I worry that there may be a circularity at the heart of Gibbard's proposal. We are asking the planning question how to live with one another on a basis no one could reasonably reject. The answer to that question is a plan that would maximize prospects for the sum total good or benefit of people. But the question whether there is such a thing as a person's good (and so a sum total good) is a planning question about whether to live in a manner that "takes a certain form," and we arrive at a shared view of what a person's good is if we all come to prefer to live in a way that takes that form. But that form would seem to be one that aims at living with one another on a basis that no one could reasonably reject. Gibbard's proposal is complex, and no doubt I haven't fully grasped how all the parts are supposed to fit together. But my suspicion is that even if our ought questions are planning questions, they have to bottom out, ultimately, in the answer to what is not a planning question. And if that is right, then the normativity of our moral thinking is probably not fully accounted for by Gibbard's metaethics, as presented in Lecture I.

Conclusion

Gibbard's ambitious arguments in *Reconciling Our Aims* seek to forge a connection between his metaethics and his utilitarianism. Each of them—his metaethics and his utilitarianism—has independent appeal, and they are certainly compatible with one another. Each also faces its own problems, as does his account of moral inquiry. But whatever their respective merits, there are reasons to doubt that Gibbard's model of moral inquiry and thinking how to live shows utilitarian answers to our moral questions to be more plausible than nonutilitarian answers. And difficulties with his efforts to show that it does may, in the end, raise more questions for his metaethics.

References

Bittner, Rudiger (1992). "Is it Reasonable to Regret Things One Did?" *Journal of Philosophy* 89: 262–73.

Broome, John (2008). "Comments on Allan Gibbard's Tanner Lectures." In *Reconciling Our Aims*, edited by Barry Stroud, 102–19. Oxford: Oxford University Press.

Darwall, Stephen (2002). *Welfare and Rational Care*. Princeton, NJ: Princeton University Press.

Firth, Roderick (1970). "Ethical Absolutism and the Ideal Observer." In *Readings in Ethical Theory*, 2nd ed., edited by Wilfrid Sellars and John Hospers. Englewood Cliffs, NJ: Prentice-Hall.

Foot, Philippa (1958a). "Moral Arguments." *Mind* 67: 502–13.

——— (1958b). "Moral Beliefs." *Proceedings of the Aristotelian Society*, New Series 59: 83–104.

Gibbard, Allan (1990). *Wise Choices, Apt Feelings: A Theory of Normative Judgment*. Cambridge, MA: Harvard University Press.

——— (1992). "Moral Concepts: Substance and Sentiment." *Philosophical Perspectives* 6 (1992): 199–221.

——— (2003). *Thinking How to Live*. Cambridge, MA: Harvard University Press.

——— (2008). *Reconciling Our Aims*, edited by Barry Stroud. Oxford: Oxford University Press.

Overvold, Mark (1980). "Self-Interest and the Concept of Self-Sacrifice." *Canadian Journal of Philosophy* 10: 105–18.

Parfit, Derek (1984). *Reasons and Persons*. Oxford: Oxford University Press.

Railton, Peter (1986a). "Facts and Values." *Philosophical Topics* 14: 5–31.

——— (1986b). "Moral Realism." *Philosophical Review* 95: 163–201.

——— (1993). "Reply to David Wiggins." In *Reality, Representation, and Projection*, edited by Crispin Wright and John Haldane, 315–28. Oxford: Oxford University Press.

Rawls, John (1971/1999). *A Theory of Justice*, revised ed. Harvard: Harvard University Press.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Cambridge, MA: Harvard University Press.

Sidgwick, Henry (1981). *The Methods of Ethics*, 7th ed. Indianapolis, IN: Hackett.

Smith, Michael (1994). *The Moral Problem*. Oxford: Basil Blackwell.

Street, Sharon (2006). "A Darwinian Dilemma for Realist Theories of Value." *Philosophical Studies* 127: 109–66.

Vavova, Ekaterina (2014). "Debunking Evolutionary Debunking." In *Oxford Studies in Metaethics*, edited by Russ Shafer-Landau, 9:76–101.

21

FREEDOM AND DIRECT BINDING CONSEQUENTIALISM

David Braddon-Mitchell[1]

1. Introduction

In *Reconciling Our Aims*, Alan Gibbard argues that his understanding of moral claims as plans—developed as plans for how to live together—together with an interpretation of the upshot of Harsanyi's second welfare theorem, suggests that a kind of act consequentialism is the right normative ethic. For some, including the current author, this is a welcome and persuasive thought. But act consequentialism is not embraced with delight by very many in the ethics community: in part because it is thought to entail many unacceptable judgements on cases. Much of the criticism of Gibbard's view has focused on whether he is right that act consequentialism does indeed follow from his views: questions have arisen about his interpretation of Harsanyi's theorem, whenever his own "Harsanyi-like" result tells us what he thinks it does, and so on. But this paper defuses another class of objection, as a by-product of something more ambitious.

1 This paper has its genesis in drafts written around 2007; various forms of unreason on my part have prevented it seeing the light of publication before now. Over the years it's benefited from discussion with many: these include but are not limited to Frank Jackson, Justine Kingsbury, Fred Kroon, the late Jonathon McKuen-Green, Andrew Latham, Kristie Miller, Howard Nye, Tahua O'Leary, David Plunkett, Peter Railton, Denis Robinson, Caroline West, and many other philosophers in Canberra, Sydney, Ann Arbor, and Auckland. Thanks to them all, and especial thanks to David Plunkett who has tenaciously insisted that it needs to be more widely seen.

The objection, the reader may already have guessed, is that act consequentialism has few adherents in ethics, and as it is usually understood seems to many to have unacceptable first-order moral consequences. If it is true that act consequentialism is mistaken, and if Gibbard is correct that it follows from his understanding of moral claims as plans together with water-tight theorems, then it looks as though his understanding of morality faces a kind of *modus tollens*. Some have argued that perhaps Gibbard's views might fare better if they entailed not act consequentialism but rather a form of indirect consequentialism, like rule utilitarianism.[2] It's not clear that they do, however, and even if they do, rule consequentialisms face at least equally intractable difficulties. This paper proposes to make Gibbard's conclusions more widely palatable and defuse these objections by showing that a direct act consequentialism for beings like us does not have the unacceptable first-order moral consequences it has been thought to have.

The principal, more ambitions project, is to develop a new understanding of direct act consequentialism that will provide the same evaluations of the rightness of acts as most indirect disposition, motive, or character consequentialisms, thus reconciling the coherence of direct consequentialism with the plausible results across cases delivered by indirect consequentialism. This is achieved by seeing that adopting certain kinds of moral dispositions causally constrains our future acts and limits our freedom so that the maximizing acts ruled out by the disposition can no longer be chosen. Thus, when we act we do the best we can, which is all that is required for rightness according to act consequentialism. I call this view direct binding consequentialism.

1.1. Some Brief Motivational Remarks

Before I start cataloging the problems of existing versions of direct consequentialism with an eye to fixing those problems, some motivational remarks might be in order. Why want to be a direct consequentialist in the first place such that you would want to remedy its apparent evils? Of course, this is no place for any extended argument that you should. The paper is primarily directed at those, among others some of the many admirers of Gibbard's work, who are already attracted to direct consequentialism but put off by its apparently repugnant consequences. But for what it's worth I will briefly outline the two strongest reasons that move me to seek a satisfying direct consequentialism.

The first of these is broadly Gibbardian. Acting consequentially is the *best* way for us to act. If we all acted this way, the foreseeable outcomes for all of us (at least behind some kind of veil of ignorance) would be better. This thought has something of the flavor of contractarianism (if we all agreed to act consequentially, we'd be better off in that, while sometimes we'd give up something precious to us, our expectation would be that this would be more than compensated for by the benefits of others so acting). But contractarianism, in many of its forms, is supposed to stem from rational self-interest. So, we could not make

2 Star (2010).

any form of contract that we could expect to keep. It might be rational to agree to create more utility—help others if you can help them more than you can help yourself or your loved ones—if others also agreed and if we could all expect to keep them. But we know of ourselves that if our self-interest is threatened in unobserved circumstances, we will have no self-interested reason to keep our agreements; and we know the same for others. So, we cannot reasonably make such commitments or believe the commitments of others.[3] Thus, the agreements we can make will be very limited and constrained by observation, enforcement, and strength of motivation.

The Gibbardian thought in response is, to oversimplify, that morality is the substitute for self-interested motivation. If we are motivated in part (and how this happens is a long story) morally and in a consequentialist way, we get the benefits that were seen by the contractarian, but forever out of her reach because she did not allow moral motivation. And of all such moral principles we might have, Gibbard argues, act consequentialism is the one that would have the best outcomes for all.

Some thought along these lines is certainly one of the motivating principles for looking for a version of this act consequentialism that would manifest this benefit (if we were all that way, things would go better for us) without getting into trouble with cases that seem repugnant.

The second is to do with how the act consequentialist thinks of values. Values tell us what the good is. How should we respect those values? One possibility is that we should hoard these in our own lives, the other is that we should try to increase those values in the world. Consider the value of friendship.[4] We might respect that value by having friends. Or we might respect it by facilitating others to enjoy rich and fulfilling friendships. These approaches might come in to conflict—perhaps we can help many people to have rich friendships only at the cost of not having any ourselves. What should we do? Now perhaps we wouldn't trust someone who foreswore friendship to help others have friends, perhaps we think that being *able* to make that choice is a sign of a character that we don't love, or of someone we shouldn't think of as a potential friend. But if someone can make that choice, and making it won't have downstream effects that will make themselves or others so unflourishing it will cancel out the benefits, wouldn't it be the selfish thing to do to keep the benefits of friendship to themselves? Wouldn't it be thus the right thing to do to maximize it?

Mutatis mutandis for other values. Perhaps nonkilling is a value. If so, as a lover of nonkilling, it will be very distressing to kill. But if we do not kill, even though killing results in much nonkilling (and no other disvalue) isn't that the selfish choice? None of this is to deny the thought that failing to restrict one's actions to the kinds of limits that the deontologist

3 It's this problem, of course, which motivates attempts to redefine self-interested rationality in a way that makes it possible to intend or commit to doing things on rational self-interested grounds that we will have no rational self-interested motivation to perform (cf. Gauthier 1986).

4 I first saw a thought like this in Pettit (1991) where he distinguishes maximizing a value from instantiating it.

prescribes may be very psychologically costly to an agent. We feel bad when we kill or otherwise traduce values that are traditional for a good reason (and of course usually we should be very suspicious of arguments that tell us we should kill, etc., on consequentialist grounds. Usually, though not always, they are spurious[5]). But these costs are just costs, and living by a maxim never to incur them when the values we allegedly love are diminished by this is, again, selfish.

And all of this is compatible of course with the thought that perhaps distress and guilty feeling and so forth are not the only disvalue when we fail to instantiate a value—when we for example kill to maximize nonkilling (one shouldn't, though, underestimate the distress and guilt. It may be life changingly bad and thus deserving of very high weight in our deliberations). It may be that we have a theory of value according to which acts of a certain kind—acts of killing, for example—are disvalues independent of the deaths and other consequences they have. That might honor something of the deontologists' intuitions. But if so, it's still a matter of where we want to place the disvalue. If we do not kill in a situation where (unlikely though this is in civilized societies) doing so will eliminate many other killings as well as other deaths, and have no other material bad consequences, then we are simply choosing to take a disvalue—killing—and remove it from our lives into the lives of others. Again, it seems, the selfish choice.

So to conclude these remarks (which of course should make no difference to whether you accept the rest of the paper), this second motivation is pretty much the thought that if you really love a value for itself you want as much of it as possible, rather than wanting just to hoard as much of it as you can. Thus, you maximize.[6]

1.2. The Problems

Direct consequentialism is the view that the right act is the one that will, of all those available to an agent, maximize[7] expected value. Indirect consequentialisms are views, which maintain that the right act is the act produced by a disposition, rule, habit, character, or even virtue, whose inculcation maximizes expected value.

It would be easy to think, given the state of the debate, that neither of these is remotely satisfactory. Act consequentialism is coherent enough, but the answers it appears to give in moral puzzle cases[8] give pause to all but the most ideological consequentialist. It seems we are expected to sacrifice our near and dear when doing so will aid famine relief, or we are expected to commit barbaric acts if doing so will result in slightly fewer barbaric acts being committed, and so on.

---

5 Some good examples of why this is so are to be found in Railton (1984) and in the much read but I fear little understood Smart (1973).

6 Though see Nye, Plunkett, and Ku (2015) for some interesting worries about this second class of motivation.

7 Or satisfice. Throughout the paper I talk of maximizing, but there is a version of everything I say that takes satisficing as the required relation between value and evaluation or choice.

8 Williams (1973).

Indirect consequentialism seems to deliver more plausible judgements on the cases. For example, versions that combine rules, characters, and dispositions may tell us that it is acceptable to have special bonds with our near and dear inasmuch as the best (consequentially understood) kind of rule or character may be one that favors the near and dear to some extent. Unfortunately, the indirect versions face a trilemma. First, they may accept that the reason for choosing the rule or disposition stems from an ethic of maximizing utility. But then the very same consideration—the overarching morality of consequence maximizing—will tell us to break the rule whenever doing so will maximize. But then we will have the underlying principle that grounds indirect consequentialism telling us to do something contrary to what indirect consequentialism itself requires, which is incoherent.[9] The incoherence can be removed by specifying that the best rule must be one that makes us perform the maximizing act on each occasion: but then it seems as though there is a danger of collapse into the act version.[10] Finally, the indirect consequentialist might claim that the desirability of acting according to the rule with the best consequences is not itself justified consequentially but is itself a basic and possibly deontological moral commitment. But then we have strayed from consequentialism: and in any case it is hard to see what appeal this would have as a basic nonconsequential principle.[11]

There are versions of consequentialism that try to have their cake and eat it too: to adopt both rule and act consequentialism as correct in their appropriate domains. A useful term for an overarching position that has this effect is *global consequentialism*.[12] Global consequentialism evaluates every kind of entity by the particular consequences of that kind of entity. Certain kinds of act consequentialism evaluate *rules* not by the overall consequences of adopting the rule but rather by whether the rule produces maximizing acts. Certain kinds of rule consequentialism evaluate *acts* by whether they are produced by the maximizing rule. Global consequentialists, on the other hand, evaluate acts and rules, characters, motives, and so on (as well as other entities) separately by their *own* consequences. So acts are evaluated by whether they maximize, and rules by whether *they* maximize (it should be remembered that the consequences of a rule may outstrip the acts it produces). This means, then, that such a view has some of the features of both direct and indirect consequentialism.

But it is still not a satisfactory reconciliation. When faced with the particular acts that act consequentialism seems to endorse, and which are *prima facie* counterexamples to that view, global consequentialism says that the acts are indeed right: even though the right rule prohibits them, or the right character will not perform them, or the right motive will militate

9 Smart (1973, 10).

10 Lyons (1965). In fact, I am unmoved by this objection. The benefit of the rule and character may be manifested in ways other than the acts it produces. If this is so, the best rule or character may well not be the one that produces the maximizing acts, if there is no way to get the good consequences of the rule or character without the nonmaximizing acts.

11 Perhaps this is the rule worship point of Smart.

12 Pettit and Smith (2000).

against performing them. So, this hybrid view still accepts that the act is right—but somehow the agent is forgiven for not performing it because this omission is out of virtue (Driver[13]) or out of good motives (Adams[14]). Some versions of this view see these cases as one where there are inconsistent obligations: one must perform a certain act, but one must instantiate a certain rule, character, and so forth, that is inconsistent with performing the act. So, a serious cost of such views is either accepting that there are all things considered inconsistent obligations (an attractive feature of consequentialism, to many, is that when all is balanced and you do the best you can, you have done the right thing), or else requiring some story about permissible wrongdoing. Plausibly, however, it is worse than that. Our intuitions about some of the cases are that it would not just be permissible wrongdoing to not kill one's spouse so as to save a few strangers: it would be morally required. But then, on this view, it'll still be wrongdoing, but morally required wrongdoing—and surely that is as close to a clear contradiction as we'll get in the murky realm of ethical intuition.

What is needed is an entirely new take on consequentialism: one that reconciles the coherence and action-guidingness of direct act consequentialism with the plausible answers, with respect to cases, of indirect consequentialism. The solution I'll defend is a reimagining of act consequentialism that sees many important moral acts as being concerned with the inculcation of moral dispositions that bind, or otherwise limit, the freedom of the agent who adopts them. Of course, consequentialists have seen this before. The novelty will lie in the answer to a problem: how we might accept and reconcile three things: (1) The rightness of adopting dispositions that are maximizing, even when it is foreseen that they will lead us to perform what look like unmaximizing acts; (2) the acceptance that many of these apparently unmaximizing acts are right; and (3) assessing the rightness of those apparently unmaximizing acts individually, not via the rightness of the acts that caused the dispositions, or the maximizing status of the dispositions or rules.

This paper proposes just such a new consequentialism. It is the view that an important part of moral choice is the adoption of moral dispositions that causally filter our future choices to make us *unable* to take seriously in practical deliberation certain options. But if we choose the best of the *available* options at the time of act selection, we do all that a thoroughgoing act consequentialism demands of us, and thus choose rightly. I call this new kind of consequentialism *binding direct consequentialism*. Binding because our actions often bind our future actions; direct because it is just that: acts will be evaluated directly, not via the rightness of the dispositions that produce them. Both first-order acts and the acts that affect our dispositions are evaluated and chosen consequentially, and this is made possible by emphasizing one of the roles of moral dispositions as binding us in a way that reduces our future choice set. What is distinctive about this approach is that I show that acts that affect moral dispositions have to be seen as causal interventions in our own psychology,

13 Driver (2001).

14 Adams (1976, 476–81).

which limit our future psychological capacities. Because of this the actions under the grip of these dispositions will be chosen from a reduced choice set, which does not include what might seem to be the maximizing choice. Thus, when we choose the best option available to us, we do all that act consequentialism demands from us. What will undergird my account will be a compatibilist take on free will, according to which choices of habit and disposition are in general free, but many acts in the grip of habit and disposition are not. We will be free to choose from the available options but not to choose from those ruled out by moral dispositions.

One final clarification is in order before I outline the structure of the paper. The view is a consequentialist one, not a utilitarian one. I take utilitarianism to be a special case of consequentialism, which has a particular theory of the good, which results in only one value to maximize, and that value is either a desirable psychological state or perhaps simple preferences. Consequentialism says you should maximize the good, whatever the good is. And plausible theories of the good include much of the texture that makes up a thriving human life: close relationships, the pursuit of important projects, and so forth. None of what I say addresses important problems about how to adjudicate different values, and whether they are comparable or not. And I'll sometime talk of maximizing expected utility, but in that case "expected utility" stands for expected value, whatever the right theory of values. Whether ultimately all values are grounded in preference, or any other foundational value, or whether they might be fundamentally distinct, is also not an issue that I'll be addressing.

The paper is divided into six sections. In section 2, I consider the general idea of inculcation of dispositions as causal intervention in psychology, both in the case of rationality and morality, and introduce a version of act consequentialism according to which options have been removed by such intervention. In section 3, I consider the issue of in what sense we are acting freely when we choose from among the available options, and sketch a compatibilist account of free will that undergirds the theory. In section 4, I contrast my view with some examples of global consequentialism, in particular Adams' motive consequentialism and Drivers' consequentialist virtue theory, and argue that my way of reconciling the intuitions behind direct and indirect versions of consequentialism is to be preferred. Before concluding, I examine and solve some problems for my account raised by considering cases where dispositions are inculcated with the purpose of spreading evil.

2. Dispositions as Causal Interventions

I will start the process of making my view plausible by comparing it with an issue in the philosophy of rationality that has a similar structure. There are interesting cases where it is rational to make oneself irrational (or at least rational to limit the domain of one's future rationality). Simple cases involve things like the beserker drug.[15] Suppose that people have

15 Gauthier (1986).

access to a drug that will make them disposed to vigorous fighting. And suppose further that it is common knowledge that this drug exists, and that I know that I do not have the acting skills to pretend to be an irrational revenge taker. In a situation in which I am about to be attacked, I know that if I flee I will be chased and badly beaten, but if I fight to the best of my abilities I will inflict very serious damage on my assailant, but in the process will be killed or far more seriously hurt than if I were to flee. What should I do? Plainly the rational strategy is to flee, for that will produce the best outcome. Fighting would be foolish and result in death or maiming; much worse than a very bad beating. But before I am assailed, I could take the drug in full view of the assailants. Should I do this? Perhaps I should, for if they see that I have taken the drug they will believe that I will fight to the best of my abilities, and thus hurt them more than they think worth while. I take the drug in the hope that I will never in fact act under its influence, though of course I may be unlucky. And if I take the drug often enough, or take a permanent version of it publicly in the hope of general protection against thugs, then I should expect on occasion to get unlucky: I will eventually come across an irrational assailant who will attack anyway and I will, under the influence of the drug, fight back rather than run away. But if this is not expected to happen too often, then it may be rational enough to take the drug.

When I fight back, do I act rationally?[16] The drug has causally restricted my options. Running away is not an option, though various forms of fighting or standing my ground are. If I choose from the options available to me the one that among them maximizes, then, modulo my inability to choose from a wider set, it seems I act rationally. For rationality (like morality) does not require that we do what we cannot. Rationality is a theory of choice among options. It cannot require that we consider all options, for no finite mind could even list the options. It does not require that I fly with wings when I cannot, even if that would have the best outcome. And similarly for mental acts: rationality does not require that we solve differential equations so difficult that no human can solve them, even if that would have the best outcome.[17] So on my picture, rather than the picture in which the drug makes me act irrationally, I rather act rationally given the constraints the drug places on me. Of course in so doing I do not perform the same act as it would be rational for an agent in my position to perform, who was able to choose from a larger choice set, which included the option of running away.

One attempt to explain why we act rationally in these or similar cases is some extended account of rationality according to which we do indeed choose from a full range of options, but we freely choose the option that is not narrowly maximizingly rational but is the rationally best option in some special sense of rational. The special sense might be that it is rational to do what it was rational to agree to do, or it is rational to do what it is rational

16 Elster (1979).

17 Nuclear deterrence and evolutionary game theory both present cases with a similar structure.

to intend to do. This is the strategy of David Gauthier[18] in Morals by Agreement and Joe Mintoff[19] in the case of the Toxin Puzzle. There are well-known objections to this solution, and I think what unites them is that it is hard to see how this extended sense of "rational" deserves to be called "rational." It shares with rule consequentialism the problem that the considerations that recommend breaking the rule in the one case, or breaking the agreement or intention in the other, are the very considerations that justify the rule or agreement. So if these considerations have authority in justifying the rule or agreement, why do they not have authority in recommending a breach?

These are cases where we are considering prudential reason. We will now consider cases where direct causal intervention applies in the domain of ethics. I will only slightly modify the berserker case to make it into an ethical rather than rational example.

Suppose that I am making the same choices about fighting and fleeing as before, but now I am considering the choices from a moral perspective—part of which is in terms of the influence these choices will have on my philanthropic projects. Once again, consider the option of fighting when confronted by the assailants. Perhaps it is *morally* wrong to fight in these circumstances (unless we think that fighting back will affect the future actions of the assailants: let us suppose we know this not to be so), for fighting will create more disutility—summed over the assailants and oneself—than running. And yet the act of taking the berserker pill in full view of the assailants may maximize expected utility; for it may be expected to prevent any attack and thus generate none of the disutility of either running or fighting. In that case, it would be morally right to take the pill. But is it morally right to fight under the influence of the pill? Well, as the pill has reduced my options, so long as I choose the best option from those available then I do what is right, even if I can see that there is a best option no longer available to me. I do what is right, even if not what is best. Just as when we perform the best available action to alleviate world poverty, we do the right thing, even if there is an option—writing a brilliant best seller that makes billions of dollars that we would then donate—which is not available to us because of causal limitation on our mental powers.

How essential is the pill to our story about the assailant? Suppose that I am unable to go out in the world to pursue my philanthropic interests because of the fear of assailants. If I do go out, it is rational for me to run, and I will. It is also morally right for me to run, for I know that if attacked and hurt badly I can do less good than if I stay at home. But I know that there are people out there who, on the whole, walk the streets unscathed. There is something about their bearing that makes the street thugs think they will fight back. I invest in videotapes about how to walk in an ape-like way that will make assailants fear my tendency to retribution, but it turns out I am no actor. Then I read a book that explains that sadly there is a nomological connexion between seeming like someone who fights back and being one. So, I buy more books and tapes (perhaps I enlist a therapist) and by exposing myself to

18 Gauthier (1986).

19 Mintoff (1997, 612–43; 2000, 339–64).

the right kind of violent film, by visualizations and various methods of psychology, I make myself into someone who will inevitably fight back. I thus walk the streets doing good largely unscathed because my new disposition is written in my walk and bearing in a way that I was unable to merely simulate.

Of course, this new disposition will eventually get me into trouble. I will be badly maimed when I fight back, and my philanthropy will be in recess for months. Will I be doing the right thing when I fight back? Well if the case of the pill is one where we think that I did not have freedom with respect to the option of fleeing, and thus chose the best action from the restricted set, then it is hard to see how we could come to a different conclusion in this case. The methods of therapy are not so very different from the methods of psychopharmacology. The pill alters my brain structure in a way that makes my choice set limited. The therapy acts on my brain structure in a way that makes my choice set limited. If I do, in some sense, the right thing in the first case, then so do I in the second case. If I did the right thing in undergoing therapy, and it produced a consequentially justified disposition, then indirect consequentialism will judge my act of fighting to be right. On my account **direct** consequentialism will also judge my act to be right, for it is the best from a set causally limited by the therapeutically induced disposition. Thus, there is no clash between the judgements delivered by direct and indirect consequentialism.

To take another example, let us look at the case of our moral strictures against killing. Many think that we act rightly when, in choosing from the available options that do not involve killing, we select the one with the best expected value. We do the right thing, even if it has a somewhat worse outcome than an option that did involve killing. How can this be, if we are consequentialists? Well, if I act as a result of a disposition that typically prevents me from taking seriously the killing options, I have done the best I can. My disposition may itself be justified by having the best expected outcomes. It is better to put up with some cases where we will not kill when killing may have had a slightly better outcome, than perhaps to be misled by our future judgements and sensibilities into unwarranted killing.[20]

These are cases, then, where we bind our future selves to limit the range of choices from which they choose. Why would we perform such binding in these cases? Because we think that with a restriction in place we will err on the side of moderation less frequently than we would err on the side of excess if our choices were unconstrained. Or, in a manner inversely analogous to the case of the beserker, we seek the advantages of cooperation, and we know that someone with a pacific nature is more likely to get such benefits, and all the scope for promotion of the good that comes with them. Of course, morally the *best* thing to do would

20 Of course, the very best moral disposition would be the one not to kill except in very extreme circumstances. These might be circumstances where the costs of not killing are so great that it is possible to build a psychology that is generally repulsed by killing but may yet be able to kill when the lives of so many depend on it. This means that someone who thinks that the wrong thing is done by someone who refuses to kill when it will save a nation can still complain that an agent who fails to do so has the wrong moral sensibilities.

be to give the impression of being nonviolent, but reserve the power to kill when calculation told us that, say, killing would have maximizing consequences that outweighed its cost. But, as with thuggery, most of us are unable to fake it. Actually, adopting a nonviolent disposition is the only way to persuade most people that we are nonviolent.

The causal binding feature of this view gains plausibility, I think, when we consider what would be required for a solution to what Robert Frank has called the commitment problem for rational choice theory.[21] The problem here is that whether one is signaling that one will act thuggishly in a narrowly irrational way in the future, or that one will cooperate in a narrowly irrational way, such signals are unbelievable if one is known to be rational.[22] So one must be able to signal that one is not unconstrainedly decision theoretically rational, and for various reasons deception on this front is not a universally successful strategy. What might work is actually to causally limit one's future capacity to act in a way that is unconstrainedly decision theoretically rational. If one is to preserve one's general rationality into the future the best way to understand this process is as causal limitation on one's future behavior so as to eliminate the options that, though decision theoretically rational to perform if available, one must now signal will not be performed if we are to gain trust.

A crucial thing to note is that at least sometimes the good consequences of the best disposition are not produced via the acts that the disposition produces. It might be that a pacific disposition is detected via pacific acts, but equally it may be detected via by-products of the disposition, such as pieces of nonvoluntary behavior and demeanor. It is this case in general, which allows for the possibility that a maximizing disposition could result in quite a lot of nonmaximizing behavior, and thus require that the disposition be causally fettering, since otherwise we will have reason to adopt the disposition but not to perform the individual acts.

The upshot of these examples is that we can see how a direct consequentialist evaluation of the act might give the same results as an indirect consequentialist evaluation. The indirect consequentialist account might judge that in the case where I do not kill despite the slightly greater benefit of killing, I do the right thing in virtue of following the rule. My direct consequentialist account also allows that the act is right, since the act is, of those available, the one with the best expected outcome. So, my account uses the plausible machinery of direct consequentialism but tracks the plausible judgements of indirect consequentialism.

One way to state this connexion is to say that there can be a significant extensional overlap between what acts count as right by the lights of various indirect consequentialisms, and direct binding consequentialism. If the reason someone is an indirect consequentialist is that they think that a certain indirect consequentialism is extensionally correct—gives the right judgements about cases—then noticing this will remove the reason they have to prefer their indirect version over binding consequentialism. Does this mean that I claim that

21 Frank (1989).

22 Of course, backward induction arguments show that this is much sooner than might seem obvious.

binding consequentialism is extensionally equivalent to indirect consequentialism? There can be no answer to that in general, since there are many extensionally different indirect consequentialisms, and so the overlap varies. But the key idea is that in many of the thought experiments where, for example, rule, character, or motivation versions of indirect consequentialism seem to deliver an intuitively better verdict, direct binding consequentialism can deliver the same one.[23]

2.1. Why Is Moral Training the Same as Therapy?

The next claim, then, is that the inculcation of moral dispositions and rules is really very much like therapy; by moral training, I alter myself (or allow myself to be altered) in such a way as to limit my future choices. When I become sensitized against killing—when I adopt a maxim of not killing—I causally alter my brain so that my future choices do not include killing. Luther's stand against the Roman church—*Hier stehe Ich, Ich kann nicht Anders*—becomes literally true.

But it might be objected that moral training, or inculcation of dispositions, is not the same as therapy. Therapy is about manipulation in a causal way, and moral training is about improving moral sensibility. We need here to distinguish between two kinds of moral dispositions: weak and strong.

A weak moral disposition is one where we operate on rules of thumb, but our deliberative powers rest in the background ready to intervene should the disposition be about to cause an action that the agent then judges to be less than optimal.[24]

Strong moral dispositions bind our future selves. In inculcating a strong disposition, we adopt a disposition that ensures that our future choices are causally constrained. We do this precisely because this will maximize overall, either because we cannot trust our deliberative powers in certain circumstances, whether because they would be too slow, too unreliable, or

23 Of course, we could engineer an indirect consequentialism that is designed to mirror exactly the structure of binding consequentialism with different ideology and so will have complete overlap. So instead of saying that an act is right if it has the best expected consequences of all those available, we say something like there is a complicated etiological and contextual fact about the act—that it depends on past choices to restrict availability of various options in various ways and that there are certain psychological features of the alternatives with better expected value that prevent them being implemented—and that it's right indirectly because of having the correct origin. But this would be a mere shadow of binding consequentialism, in the same way as an indirect view that said that an act is right if it produced by a mechanism that on that occasion produces an action that maximizes expected utility is technically indirect (because the rightness stems from the mechanism) but is a mere shadow of the standard direct consequentialist view.

24 Weak moral dispositions—ones where the dispositions are overseen by a maximizing psychological overseer—correspond to the kind of disposition that we would internalize if we were what Pettit and Brennan (1986, 438–56) call virtual consequentialists: on that view the weak moral disposition is the motive for action, allowing us to behave with nonmaximizing motives while the overseer ensures that we in fact do maximize. It is not part of my view that this never happens—it clearly does—but rather that strong moral dispositions are what is required in cases where the maximizing benefit of the dispositions can only be purchased at the expense of acting on some occasions in an unmaximizing way.

where the fact that others see that we are not casually bound may have bad consequences.[25] Thus, we can make a morally justifiable choice to adopt these dispositions even knowing that sometime in the future we will act with worse consequences than we would have if we had deliberated.

Both kinds of moral dispositions are part of our lives, but it is the second kind that we are concerned with here. And this is the kind of moral disposition that it is hard to see in terms other than causal intervention. What we do when we teach our children how to be moral is to give examples that we think will rub off, and when we engage in programs of moral self-improvement we engage in exercises that affect our sensibilities in a way we think will affect our future choices. How can we make sense of the adoption of a moral disposition unless we think that it will causally restrict our future range of choices in some situations?

To answer this question, we need to distinguish between two kinds of moral training. One sort of moral training is what we engage in so as to improve our judgments as to what makes actions right, either by improving the decision-making process or by improving the values built into it. Suppose that we were to change our assessment of what makes actions the best: it would make perfect sense to build that new assessment into our decision-making system without restricting in some sense our future range of choices. We would be modifying our future choice behavior, in light of changed views about what is the best basis on which to choose, and then leaving our future selves free to choose what they think will be all things considered best.

But there is another kind of moral training: one where we have not altered our opinions about what the best calculational formula is, nor what the right values are. In this kind of training, we think that it is best to be such that we will *not* choose according to correct principles of decision-making at some point in the future. This is something, which can only be achieved by genuine causal restriction of options. For we can foresee that if we were perfectly rational, moral, and free, we would act in accord with the correct principles in future situations unless we do something to prevent us from choosing freely, rationally, and morally.

Strong moral dispositions are required, then, in this latter case: where we have reason to think that there is a case to be made for overriding what would be calculated on a case-by-case basis—much as when there is a case for overriding what would be rational to decide on each occasion in the berserker case above. And if that is so then it *only makes sense* to train morally in this way on the assumption that it really will limit our choices. Just as there would be no point in taking the pill if what it did was to enhance rationality (for then it would ensure that we ran rather than fought) there would be no point in moral training if all it did was slightly enhance the accuracy of consequential calculation. The point is that

25 Of course, it is no part of my theory that this is the *conscious* justification of adoption of such dispositions: only that they are rightly adopted because of these factors (and these factors no doubt play some indirect role in the explanation of our propensity to form such dispositions).

strong moral dispositions are supposed to circumvent the deliberations that a free rational and moral agent would make.

Some might object that this second kind of moral training is not the correct account of moral dispositions in cases where we have not changed our opinions about the values or rationality. Suppose instead, they might argue, that what we do when we engage in moral training is not to affect our outcomes by constraining choice but rather to limit the influences on our choices. Perhaps moral training is about making better choices by cutting off influences for ill. It's about reducing our selfish desires, so they are not inputs to future decisions, or alerting ourselves to salient features of future situations we may encounter, which will be important to correct decisions and which may otherwise be overlooked.

I do not disagree with this account of some kinds of moral training, but this is simply a third kind of moral training. When this is the right way to describe the training, then moral training is indeed just like taking therapy to improve one's implementation of an unchanged conception of rationality. In the particular case of the berserker pill, an alternative that enhanced narrow maximizing rationality would not have served the same purpose. But there are many situations in which our rational decisions would be improved by better calculation or more care in choosing relevant data. So, it is with moral decisions—moral training that enhances our ability to calculate and evaluate consequences and attend to morally salient factors in situations is no doubt a good thing. When this kind of moral disposition is what we are talking about, then there is no problem for ordinary act consequentialist evaluations of the rightness of the actions. The disposition improves our capacity to calculate, and we do the right thing when we calculate correctly and act accordingly. But these are exactly *not* the kind of dispositions that indirect consequentialists are concerned with, and where our present task is to show that a sophisticated take on act consequentialism can give the same judgments as the indirect view. Indirect consequentialists concentrate on rules of thumb, moral maxims, and dispositions that make one act in some cases *other* than how one would by merely doing the calculations and then acting in response to them. And only in these cases is there any issue about the way in which the direct and the indirect evaluations stay in track. Thus, they cannot be mere fodder for the making of better decisions.

So the idea, then, is that when we take on strong dispositions we limit the range of our future choices causally. In the future we act freely but only with respect to the available options. In some cases, the best act will no longer be available to us so that in choosing the best available act we do all that consequentialism demands. All of this demands that I say something about freedom, which I do in the next section.

Before I do that, however, I should deal with an objection that might forestall the point of investigating issues of freedom. It might be argued that the right act is not the act that is available to the agent and that has the best consequences. It is rather the act that is available to a *reasonable agent in those circumstances*. This would rule out the physically impossible acts, and those beyond the reasonable mental powers of an agent, but it would include as genuine options things that the agent is unable to do insofar as he is unreasonable. Plausibly,

having taken a berserker drug, for example, makes me unreasonable, however reasonable it may have been to take it. So the option of not fighting is available to the reasonable agent. Similarly, in the case of moral dispositions, if someone has limited his capacity to morally reason, and act with practical wisdom on those deliberations, and as a result is in the grip of a disposition not to kill (except perhaps if the consequences are extreme), then the option of killing is not available to him. But if the right act is the one with best consequences available to the reasonable agent, and the reasonable agent does not have these causal strictures on their deliberative powers and practical reason, then he still acts wrongly by not selecting the option that involves killing.

But this depends on an account of "reasonable agent" that is tendentious. Is the reasonable agent one who has failed to adopt a disposition that a reasonable agent would in fact have adopted? That is what would be required. We would be claiming that the reasonable agent is someone who has not acted reasonably in the past. Of course, there could be an account of "reasonable" according to which what makes one reasonable is unconstrained decision theoretic rationality. But it does not seem to be relevant to the question of evaluation of acts—for it makes the right act the act that would be performed by someone *who is not as they ought to be had they been reasonable*. It thus would not give us an account of right that would help in the business of choosing acts given the kinds of agents that we are and ought to be.

2.2. Free Consideration Rejected

Why must we sometimes see the adoption of dispositions as a causal intervention that reduces our choices? An objection might go like this: if you can have a pill that will eliminate some options, why not have a pill that allows that you consider all options, but which ensures that you will freely choose to decline all the (e.g.) killing options?

The brief answer is that if the pill left us free and unfettered, how could it guarantee that we would always decline a killing option? The longer answer is this: let's start by assuming for the sake of illustration a simple theory of values where the only net lives saved among the existing population is, of all the things mentioned in the example, the only thing that matters. Of course, for any different theories of value, parallel points could be made with different examples. So, consider a case where killing saves a few lives. Suppose that we are free of will, unfettered in our psychological capacity, rational and knowledgeable, and moral. To the extent that we are free of will we have the power to select a killing option. To the extent that we are moral, we will choose the maximizing act (even if we have chosen some maximizing rule that rules it out: we now see that we can create more utility by breaking it in this case). To the extent that we are unfettered we still have the psychological capacity to kill, or can take the option seriously, to the extent that we are rational and knowledgeable we understand the expected outcomes, and have the practical reason to bring about what our morals require. So, we can be sure that will choose the killing option.

If the pill then (or the moral training) prevents us from choosing this option, it must make us immoral, fettered, ignorant, or irrational. The kind of moral disposition I am considering

here makes one psychologically fettered: it takes away one's ability to kill, or to take seriously the killing option, while leaving us unaffected with respect to the remaining options. Of course, this is not the only way the pill could work. The pill or training could work by, for example, giving us false beliefs and making us into deontologists. Or it could work by making us systematically miscalculate the expected value of the outcomes. All of these perhaps happen. But to the extent that there are act consequentialists who can inculcate these moral dispositions—and there are—the option that is sometimes taken is fettering. We become people who are unable, without reprogramming, to kill.

3. A Need for a Theory of Free Will

In the previous section, I said that strong moral dispositions circumvent the deliberations that an agent who is free of will, unfettered, rational, knowledgeable, and moral make. This paper is about the kind of dispositions that circumvent these deliberations by fettering agents' psychological capacities so as to limit the range of choices.

What does it take to count as a limitation on a range of choices? It means that some of the choices are ones that we no longer have because we are unable to choose them. Of course, this is not an inability of the kind that we have when we cannot choose to sprout wings and fly. It is rather a kind of psychological inability. It is an inability such that, while we are free to choose from among the remaining options—in some sense we retain our free will—we are not free with respect to the missing options.

So, in what sense are we not free? Someone who has a disposition against killing, or a disposition not to lie, might not have any kind of obvious impediment that prevents her from killing or lying. She may, even as she does not lie, think to herself that she could lie if she wished, but chooses not to because she has internalized the nonlying disposition.

Of course, there is a full causal story that explains why she will not lie, and it may make sense from that perspective to describe her not lying as a product of her not being free. But of course there is equally a full causal story about how she chooses when she deliberates, or how she chooses when she takes on the dispositions. There is, plausibly, a full causal story about everything from the perspective of which it looks like there is no free will.

If we are not to be eliminativists about free will, or chain ourselves to the questionable metaphysical presuppositions of libertarianism about the will, then some kind of compatibilist account is required. A compatibilist account of free will is one that accepts that an action can be freely chosen if it is fully causally determined by factors prior to the agent. Thus, a compatibilist about free will must admit that we can have a range of choices open to us from which we choose a particular one, even though there were predetermined causal factors that eliminated all the other options. So merely knowing that a range of options has been causally limited does not guarantee that we were unfree with respect to them. So, *a fortiori*, knowing that a moral disposition eliminates certain choices from being actualized does not guarantee we are unfree with respect to those options.

Thus, this objection runs, we might still take an act consequentialist view according to which we have done the wrong thing, because the right option was one of those that we did not perform, even if the fact that we did not perform it was causally determined. Of course, this does not mean that the compatibilist need deny the principle that ought implies can (and the contrapositive that if one can't do something, it is not the case that one ought do it). But it means that there must be some compatibilist reading of "can" with respect to an option, which is compatible with that option's being causally ruled out.

What the present paper requires is that there is some compatibilist reading of "can," and of "free will," according to which when one chooses from options left open to one by one's moral dispositions, one chooses freely from among the things one can do, but according to which the options ruled out by the dispositions are ones one can't select, and with respect to which one is not free. But in its bare abstract form, compatibilism does not tell us that the options ruled out by the dispositions are ones we can't select, nor does it tell us that the options ruled in are ones we can select. Compatibilism only claims that there is some basis for choosing among determined actions those which are determined but freely chosen, and those which are not. To motivate the claim that the genuine options are only those that are consistent with our moral dispositions, we will need to put a much more flesh on the theoretical bones of compatibilism, and some of that will begin in the next section.

3.1. Compatibilism and Moral Dispositions

The full account of compatibilist freedom is not something that needs to be settled here. Nor will it be something remotely uncontroversial. But I will assume that out of the various accounts available there are principles that render it sufficient for freedom that one is acting and choosing on the basis of a well-functioning deliberating device that is reliably connected to action. One acts freely just so long as this device is efficacious and functioning normally, regardless of the causal determination of the device. This extremely abstract formulation captures something in common with higher-order desire accounts,[26] biofunctional accounts,[27] ideal desire or ideal higher-order desire accounts, and so on.

So the thought is that the way to characterize freedom of the will is that for an action to be performed with free will, it must be under the control of one's own decision-making apparatus in some way. In many cases, our day-to-day actions are settled by individual deliberations of this kind, and thus are free. Perhaps even more commonly, actions are under the control of simple habits or rules of thumb—that is, dispositions to produce behavior in a more or less reflex way—but crucially where this is mere calculative convenience: if the decision center is presented with information that the current circumstances are ones where the habituated behavior is inappropriate, the habit is overridden; or perhaps these rules are

26 Frankfurt (1971, 5–20); Dworkin (1970, 367–83).

27 Stamp and Gobson (1992, 529–56).

themselves selected by deliberation.[28] These dispositions are a generalized version of weak moral dispositions and cover all cases where deliberation is moved to a back-up role. These weak dispositions produce freely chosen behavior, since they are still sensitive on a case-by-case basis to the deliberative control unit, if in a slightly indirect way.

The kinds of moral or rational dispositions we are concerned with here, however, are strong dispositions, where even when we realize that the behavior we are about to engage in does not achieve the very goals that justified inculcating the disposition in the first place, we proceed anyway. We have made the decision to constrain our future behavior through inculcating a disposition to limit the range of options considered by our decision-making module.[29] We have treated ourselves—or at least our future selves—as mechanisms to be manipulated by causal intervention, just as we can imagine doing to one another. The future behavior, insofar as it is produced by our decision-making system without external interference is free, though it is not *free with respect to* the eliminated options; just as our decisions are in general not free with respect to what we cannot do. Our causal intervention has changed what we are able to do in the future.

For simplicity in this paper, I'm going to talk as though the distinction between weak and strong moral dispositions is ungraded. For many practical purposes, I think it usually is. But clearly there will be borderline cases; weak dispositions may become so strong that while they can perhaps be overcome by deliberative input, but the probability of success is so low that we may count them as strong. That being so, it follows that there is a penumbra region. There are many ways to go in that case. We could make an account of free will graded, and make moral notions that require it sensitive to this. Or we could make it vague whether the dispositions are strong or weak, and the will is free or not in these cases, and make the corresponding moral notions susceptible to vagueness, or some combination of these. I won't discuss these complications further in this paper.

Compatibilist accounts of freedom of this kind are accounts of when actions are performed with free will. Freedom of the will must entail a connexion between action and the control device that is the will; freedom of action is having one's acts under the control of a control device working in the appropriate way. The behavior of the control device is determined, but on this view it is some kind of category mistake to ask about freedom of the control device insofar as it is working correctly—or at least that puts you in the camp that is

28 The doctoral dissertation by Andrew Latham (University of Sydney, 2019) develops the second of these thoughts in some detail into a view he calls "indirect compatibilism" where much behavior is under the control of autonomous causal selection processes, but counts as free insofar as those processes are themselves created or triggered by deliberative causal processes.

29 Note that this perhaps controversially distinguishes between that part of our psychology that makes decisions about actions and that part which reasons about the desirability or utility of actions. For we might be able to reason that a certain ruled-out option better promoted the very values whose promotion justified the disposition to rule out the option in general, and yet not consider the ruled-out option seriously in actually deciding what to do.

going to be unfriendly to compatibilism of any stripe. What counts as incorrect working is part of the difficult task of coming up with a substantive compatibilist theory of free will, but everyone who believes in rationality and is not a metaphysical libertarian has that task. Some plausible candidates for sufficient conditions for improper working include being under the control of a further control device external to the self—normal environmental causal impacts are irrelevant. Thus, actions can be free or unfree. There can be a lack of freedom due to the control device being improperly connected to another control device, as in brainwashing, hypnotism, and so on. There can be lack of freedom due to the actions not being caused by the control device. So free action is action that is:

(a) Under the control of the correct control device, working correctly (and no other).

Of course, the act that I performed when I took the pill or began the inculcation of the disposition or moral habit is free on such an account, because it is under the control of the control device. The actions performed under the influence of the pill are also free, but the range of options has been causally limited by the pill. So, these considerations give us the following taxonomy of freedom of action:

(1) Ordinary actions under the control of the control device are free when the device is not under the control of a further device (regardless of other environmental influences on the device) and the device is functioning normally.
(2) Actions under the control of weak dispositions are free.
(3) Actions under the control of strong dispositions are free when the control device exerts a synchronic effect in choosing between the available options, but the actions are not free with respect to the ruled-out options.

I do not expect these constraints on an account of freedom to be completely intuitively satisfactory. But this is because of the nature of the debate about free will. I do think that there is a kind of conceptual priority to libertarianism. If it were coherent, and if the properties the libertarian believes in were ever realized in actions, then those and only those acts would be free. Otherwise, if the actual world fails to contain libertarian powers, actions are free if they meet some compatibilist standard. The conditional is much like one I have elsewhere defended for the case of dualism and physicalism about qualia.[30] But nothing here hangs on the idea that the compatibilist account is an account of free will given the very same concept of free will (which the conditional story preserves). We could equally say that the account is a successor concept resulting from reform in the light of being unwilling to accept an error

30 Braddon-Mitchell (2003)

theory, as Frank Jackson suggested in 1998[31] prefiguring conceptual engineering and conceptual ethics. On either understanding, a compatibilist account is a second-best account; it is an account of what we should call "free" given that there are tensions among the core ideas of freedom of the will. As such it cannot hope to be completely intuitively satisfactory. In the next section, I address those for whom second best is not good enough.

3.2. What If the Error Theory Is True?

Many are not convinced that any compatibilist account of free will is correct. Some of these are libertarians about free will and hope that determinism (or determinism plus chance) is false, or that there is direct intervention by some faculty of will that is not naturalistically causally determined or some such. I have nothing to say to such folk that will fit into this paper. I very much doubt that their view is right or even coherent, though if it is both then I agree that freedom of the will would track the operations of such a faculty and thus whether actions under the control of strong dispositions would be free would perhaps depend on whether the special faculty was in operations in each of the actions.

There are however another group of noncompatibilists with whom I have much more sympathy: error theorists about free will. The error theorist agrees with the libertarian on conceptual matters: she agrees that is *a priori* about free will that if there is any, there must be an uncaused and undetermined effect on options, which causes one out of a range of actions to be performed, where there was no prior determination of the outcome. The error theorist, however, disagrees with the libertarian on an *a posteriori* or perhaps logical level. She holds that this condition that is necessary for free will is either incoherent, or at least empirically ungrounded, and thus that true free will is either logically impossible or at least not actual. In either event there is, for her, no free will.

What use are my considerations about the nature of freedom to an error theorist? More than you might think, I suspect. Although the error theorist thinks that there is no free will, this does not mean that there is no responsible action. So, one might give an account of responsible action (in the sense of actions for which one is in some way responsible) that denies an analytic connexion between attribution of responsibility and the nonexistent freedom. Instead of taking the sketch I give above to be an adequate account of freedom, the error theorist might take it to be an account of an important component of responsibility. My account of responsibility would then look much the same as an account that does insist on the analytic connexion with (compatibilist) freedom, except that the component that was labeled "freedom" is now labeled as a psychological precondition of responsibility.

Even if the error theorist about freedom insists on an analytic connexion between freedom and responsibility, and is thus also an error theorist about responsibility, this does not rule out an account of the properties in virtue of which we hold people responsible.

31 Jackson (1988, 44–45).

Of course, this may end up being a revival of Sidgwick:[32] we might promulgate theories about when we should hold people responsible on consequentialist grounds. That is, the best theory or theories about responsibility might be those that have the best consequences.

I think that the account I give here might be recommended on those grounds as well.[33] My sketch of responsibility focuses us on the right level of intervention. When a bad action is the result of a strong disposition, this account encourages us to focus on the disposition, and attempt to change it. When your friend is routinely late, and innumerable instances of annoyance have not changed the pattern, it becomes futile to insist that he simply pull his socks up and exercise more willpower on a case-by-case basis. Instead, the focus of disapproval should move to his not having instigated methods of changing the disposition that causes the lateness. When someone is involved in a car crash when drunk, as a result of recklessness that would not happen when sober, the focus of disapproval should be on the disposition to drive when drunk. If that disposition is also under the control of a further disposition to drive when drunk, the disapproval and intervention should focus on their not having done what it takes to undermine that disposition, or to remove themselves from situations in which they will exemplify the disposition. The controlling disposition, whichever it is, is the one that it will have best consequences to modify.

4. Direct Consequentialism Reconfigured

We have, then, a reconciliation of direct and indirect consequentialism. In fact, it is formally a purely direct consequentialist theory, since an agent does the right thing if and only if they perform the available act that has the best consequences. The reconciliation consists in tweaking the account of "available" so that the evaluations that direct act consequentialism gives track the evaluations that indirect consequentialism would give.

How does this contrast with other attempts to reconcile the intuitions underlying direct and indirect consequentialism? There is a number of such theories, all of which I think are species of what Pettit and Smith call global consequentialism. In this section, I will focus on two of these views—Adams' motive consequentialism and Driver's consequentialist virtue theory[34]—but the basic structure of all such views is quite similar. On Adams' view we evaluate motive sets consequentially based on their overall consequences, and we independently evaluate our acts consequentially. Driver offers a consequentialist virtue theory, where virtues are sets of psychological traits and dispositions (perhaps including motives) that are

32 Sidgwick (1930).

33 Clearly responsibility is not univocal: the concept of Ministerial responsibility in the Westminster system, for example, is one that might be defended on broadly consequentialist grounds, even though it is at odds with the internalist conception of responsibility I am discussing here.

34 Adams (1976); Driver (2001).

satisficingly consequentially good, but the rightness of acts is separately evaluated according to a satisficing consequential calculation.

Both these views treat the evaluation of acts and the evaluation of some indirect property that relates to acts—in Adams' case motives and in Driver's virtues—separately. The right motive/virtue is the one that has best or good enough consequences, the right act is the one that has best or good enough consequences. On these views if we have adopted a disposition that causes us to act in a way that is nonmaximizing,[35] then we have done the wrong thing. But nonetheless, if the disposition is still one that will generally have good consequences, we can approve of the agent for having acted out of virtue. This is why they are both species of global consequentialism. I will call such views *nonbinding global consequentialisms*.

This is not the principal point of difference with my view; although I present my view here as a pure act consequentialism—the evaluation of dispositions, motives, and so forth is limited to *acts of disposition inculcation or retention*—there might well be a version of my view that evaluated dispositions themselves, rather than acts of inculcation, consequentially.[36] The principal point of difference, rather, is that these forms of global consequentialism allow there to be a clash between the two components of the theory. Sometimes the right act will not be the one produced by the virtuous person, or the right motives. On my view, there is a connexion between the right disposition and the process that produces (or monitors) the act: the right disposition rules out causally the options that might otherwise have been right so that the best available act at the time of performance—and thus the right act—will always be the one recommended by the disposition.

Why should we prefer binding consequentialism to nonbinding global consequentialisms like Adams' or Driver's? Their views have a number of costs. First, there will need to be an account of permissible wrongdoing. Where binding consequentialism says that when you, say, fail to kill your spouse and thus save a few lives you do the right thing because you do the best available thing, nonbinding global consequentialism says that you do the wrong thing but permissibly: permissibly because it is done out of virtue or good motives. Alternatively, nonbinding versions might allow that there can be all things considered conflicting

35 Or satisficing; Driver is not in fact committed to maximizing. For expository reasons throughout I will talk of maximizing, when in fact being neutral between the maximizing and satisficing versions of consequentialism is fine for the current purposes (though on other grounds I favor a maximizing version).

36 Though I do not favor such a view, since it would tell us that the right disposition to have would be the best one, regardless of whether there was any way to achieve it. However, I do not wish to take a stand on an in-house issue among binding consequentialists: the choice between what might be called binding global consequentialism and binding act consequentialism. Binding global consequentialism would evaluate all dispositions consequentially, regardless of whether any acts of inculcation or retention are performed, and separately evaluate the acts produced by them. Binding act consequentialism evaluates things only insofar as they are subjects of choice and thus there are acts to consider: and some of the acts will be acts of disposition choice. But both kinds of binding consequentialism will have the crucial feature of harmony between the evaluation of general acts and the evaluations of the dispositions or disposition choosing acts.

obligations—such as to be virtuous and to do the right thing. It is not appropriate here to argue that such consequences are intolerable costs; some are certainly prepared to embrace them. But many are not: and to them my account may recommend itself. Another worry is that rather than permissible wrongdoing, perhaps it may look like there is *required* wrongdoing: if the evaluations of the dispositions and the acts are indeed separate, and generate their own independent obligations, there may nonetheless appear to be cases where one obligation trumps another without eliminating the trumped obligation. Thus, it might be that we are required to prefer to retain virtue or correct motivation and not to kill our spouse so as to save two strangers, but at the same time the obligation to kill remains in force. So, we have done wrong in not killing our spouse, but we were morally required to do so. Morally required wrongdoing is surely a significant price to pay for a view.

Perhaps the greatest cost of the nonbinding global consequentialisms is that it removes the point of "right" as a reasonable evaluative concept distinct from "good" or "best." Consequentialism promotes a connexion between the right action and the best action. But it does not identify them for good reason. Rightness is an evaluative notion connected directly or indirectly with choice. The best outcome is almost always one quite outside of our control on absolutely anyone's conception of control. The option according to which one clicks one's fingers and rights all wrong, the option according to which one sells rights to one's image and banks the huge profits in the name of some ideal charity, the option according to which one prevents starvation by proving Goldbach's conjecture and using the resultant prize money, are all options that are better than any that are usually available to us, and yet they are not what we ought do because, being impossible to us because of physical or psychological deficiencies, they are not really options at all.

This much is agreed territory: Adams' and Driver's version of direct consequentialism respects the idea that the right act is not one of the acts that it is impossible for us to perform. However, it turns out that in the case of moral dispositions or virtues, the maximizing acts prohibited by the virtue or motive still turn out to be right. To justify this, we would need a story about rightness, which explains how we can rule out acts prohibited by physical constraints and (perhaps) mental illnesses, but not those ruled out by the virtues or motives. The deliberative function of rightness does not seem to support such a story. We need the idea of rightness to give us an account of best available option. We need this idea to choose or evaluate the best available action. But the whole point of inculcating motive sets or virtues is to rule out certain options from serious consideration and deliberation. If the justification for not regarding as right any action that is physically impossible is that it is excluded from deliberation because it is pointless to deliberate over actions that we *know* will not be performed, then the very same justification applies to actions that will not be performed because of the motives or virtues that we have inculcated in ourselves. Once accounts of rightness go outside of the ranges of various kinds of possibility, they lose an important connection between deliberation and evaluation. You can no longer make the claim that the agent could have done the right thing in the relevant sense of could. If the right act is

merely logically possible but not physically possible, then it's an account of the best act that is logically possible. If an act is nomologically possible, but not psychologically possible (i.e., psychology could have been different consistent with the laws of nature but isn't), it may be the *best* nomologically possible action, but calling it 'right' is inconsistent with the crucial deliberative role of considering rightness. We might have hybrid kinds of rightness, where we allow certain kinds of impossibilities but not others. All of these kinds of thoughts may play useful roles in our moral thinking—we may need to know what would be right, if certain things became physically possible, or psychologically or technologically possible, and so on. But these accounts of best action would then require us to have a new notion—the best action that we can perform—to do the evaluative work for us. In this sense, it would be terminologically idle to use "right" for the action we already can call best, and to have to create a new term to play the rightness role.

Of course, this only establishes that there are *some* constraints on options. It establishes that "ought implies can" is true in general, but not what the correct account of "can" is. Driver's or Adams' act consequentialist components are presumably not designed to tell us that we ought to do these superhuman things. But focusing on them makes salient the features of acting contrary to virtue or inculcated motives. We treated ourselves as bindable when we inculcated the virtues: *there would have been no point in doing so otherwise*. What justifies a change in perspective so that we no longer see ourselves as bound when we come to act? It is not as though we repeat the calculations that we performed when selecting the disposition, and then act on our current judgment. The central point is that to see the process of strong disposition forming as rational requires that we see ourselves as setting in place a causal change in our psychology that binds us, and removes as options things outside what we are morally disposed to do. When we act on dispositions, the field of our deliberation is the field allowed by the disposition, and the connexion between deliberative evaluation and rightness thus guarantees, on my view, that the right action is not outside that realm.

Similar points apply, I think, to the sophisticated consequentialism whose development starts in Peter Railton's "Alienation, Consequentialism, and the Demands of Morality."[37] On views like this, we are rightly exhorted not to think of consequentialism as a view of act choice, but as a theory of what the right act is. And what seems to us to maximize may often not be the truly maximizing act—thus the phrase "objective consequentialism" is often used. Since we can't rely on our predictive powers on a case-by-case basis, we should focus more on acts of inculcation of rules that look like rules for the good life. The good life includes meaningful projects and committed relationships, and this good would be undermined in sometimes unpredictable ways by acts that seem to maximize, as when I sacrifice my spouse for the lives of a few others. I do not disagree with any of this; but I think it doesn't solve the problems that direct binding consequentialism solves. On this view, we might rightly perform the act of inculcating some maximizing disposition: but we again do the wrong

37 Railton (1984, 134–71).

thing, when acting on that disposition, we fail to do something objectively maximizing (and it's easy to imagine that sacrificing one's spouse for the lives of half a dozen strangers who will go on to flourish could, in some circumstances, fall into that category). Thus, the view inherits the tension of some forms of indirect consequentialism, even though it's a direct one.

5. The Best Choices of Evil People

What should we say about the choices of people who have deliberately inculcated dispositions to do bad things—to limit their future choices in such a way as to eliminate as options many better acts that otherwise would have been real options and which then it would have been right to choose?

We can imagine two ways these dispositions could be adopted. There could be a figure like Milton's Satan who, working under the slogan, "Evil be thou my good!," chooses to adopt dispositions to make him behave in ways that will on balance have bad consequences. Or we can imagine someone out of self-interest adopting moral dispositions that will have on balance worse expected consequences, but better selfish maximizing consequences. Such a person might, for example, deliberately engage in a hardening of the heart that will prevent him from being able to give to the poor, by focusing on tales of fecklessness that make poverty seem like despicable weakness.

The problem is that in either case, once in the grip of the dispositions, if the agents nevertheless choose the best of the limited options left to them, then on my account they have done the right thing, for they have chosen from among the genuine options that one with the best expected outcome. So having eliminated the option of giving to the charity collector, an agent reflects and decides that it would be better to refuse politely than hurl abuse at the collector. They decline politely, and it turns out on my account they have done the right thing, even though there is a very perspicuous action—giving the money—that they have not performed.

The first thing to say is that my story inherits from the general consequentialist tradition a distinction between the rightness of actions and the rightness of blame. And the second thing is that such a person *has* performed relevant actions that are wrong: the actions of inculcation of the wrong disposition.

The combination of these makes acceptable this somewhat peculiar consequence. For there is a focus for attributions of rightness and wrongness, which is at the level of the adoption of the dispositions. And from the perspective of where we should focus our annoyance, anger and disapprobation, it should (consequentially) be at that level. For complaining that the instances are morally poor choices on a case-by-case basis may do little good. The point is that the outcomes on a case-by-case basis are bad ones: and the moral problem lies in a culpable *lack* of deliberation, or limited scope of deliberation, and the moral critic needs to insist on the wrongness of the decision to impose these limits and demand that it be readdressed. It is these limits that should be the focus of our moral concern, not the

deliberation and choice within them. And of course nothing prevents us from saying the acts are bad—have bad consequences. It's only that we can't say that, in most cases, the agent having so hobbled themselves does the wrong thing *at the time of action* in performing them.

It should also be noted that there is something to be said here about the difference between the agent who chooses subject to a strong disposition and chooses the best available option and the agent who chooses subject to the disposition and chooses a poor one. It is possible for an agent who has morally reformed, and has begun the process of eroding their previously inculcated disposition, to at least conscientiously do the best they can (the right thing), and they are in marked distinction to the agent for whom the procedure of choosing the best available option plays no role, even in choosing between the limited options left by the evil disposition. We can at least see that the person who politely refuses the collector is morally superior to the one who abuses the collector. Perhaps it is this that makes moral sense of drama where we see someone thoroughly gripped by evil dispositions trying to do the (morally) best thing under its constraints. Tony Soprano, say, is quite literally bound by mobster dispositions, but at least some of the time tries to choose the least bad option—and (in some admittedly limiting sense) we feel that he does the right thing, if not the best thing, when he does.

Finally, the idea that such an agent acts rightly may seem counterintuitive because in general there is an intuitive problem with bad outcomes being right ones. This is why some people are moved by the idea of strong moral dilemmas: if all of the actions are in some way disastrous, it seems little compensation to say that an action was the right one because it was least disastrous. And in general there is a temptation to ignore the fact that an option was the best available because we are (often rightly) wary of excuses. If an agent tells you that they chose a bad option because it was the best available to them due to some lamentable past poor choice, one might be suspicious. And this suspicion may manifest itself not in questioning the truth of the claim that it was the only option available (particularly since the truth of this is hard to establish) but of doubting that the action was therefore right. But all intuition really is doing is signaling that *something* is fishy; that it is the local action that is strictly speaking not right is just the guess one may have as to the location of the smell. The exact location is up for grabs by the best theory. Importantly, on my account, in these cases rightness of the action itself is not the most important point of intervention. The action may be right, but that is not sufficient for praise or approval if it does not spring from a good disposition. And, of course, we may have a generalized sense that we should somehow intervene in the situation: but it may be subjectively hard to determine whether that is a need to intervene in the agent's dispositions, or to complain about the act. This sense may be the source of our ire and disapproval in these cases, even when it turns out that we can be brought to believe that the right action was locally out of the agent's power.

A final sweetening of the pill, if it is needed, is that much badness is precisely not of the kind where dispositions for the bad have been inculcated. It is rather exactly of the kind where dispositions of a moral kind have *not* been inculcated. Instead, the agent takes on the

perspective of choice on a case-by-case basis. They choose between the selfish and the virtuous act, and choose the selfish. Thus, their act does indeed count as wrong. It's an empirical guess, but I think that most wrong action arises from allowing oneself to unboundedly choose, rather than lashing oneself to the mast of the bad.

5.1. Temporally Extended Acts

The individuation of acts is a very controversial topic, but one that is very relevant to the issues here. For depending on how we individuate acts, it may turn out that when someone inculcates a disposition, they are performing some other act as well—some temporally extended act. In this section, I will deal with some difficulties for my view that such an idea creates but also use it to propose another way to make sense of evil folk who choose the best available option.

Suppose someone presses a button, knowing that it will result in plague germs being dropped on some village or other in twelve months (but not which village). They could plausibly be held to have performed an act of biological warfare against the village. This act, we can assume, would not be a right act. They chose from various options the one of pressing the button and thus having the village destroyed.

I take it that the above is uncontroversial. But what should we say if instead of being a button that causally impacts the external world, it's a (metaphorical) button that causally constrains the agent's future psychology? Imagine that the button is a kind of pill that the agent takes, knowing that it will eliminate as options for him all acts that fail to result in his delivering the plague germs in person in twelve months. Perhaps he knows that he might be talked out of the act in the future if he doesn't constrain himself.

We might not think that the act the agent performs *at the time of delivering the germs* is one that is free with respect to not delivering the germs. However, with respect to the actions performed at the time of taking the pill, it is hard to see how there are relevant differences between the actions performed when the pill is taken and actions performed when the button is pressed. In each case the agent knowingly chooses certain outcomes, and causally impacts on the future to ensure that outcome. In one case, the part of the world used as means is some device connected to the button, in the other it's the agent's own body and psychology affected by the pill, but I do not see an argument for the relevance of that difference. If this is so, we must assume that the agent who takes the plague pill has performed a *temporally extended act* of biological terror at the point of taking the pill.

This situation raises problems for the kind of direct consequentialism I advocate here. For on my analysis, when someone adopts a disposition that is maximizing, and then does some individual act under its sway that is not maximizing, they still can be counted as having acted rightly insofar as they have chosen the best act from a limited range of alternatives. However, what if it turns out that such a person can be accused of having performed a temporally extended act that is wrong, since he earlier acted in a way that he knew would have the consequence of sometimes producing bad outcomes? At the time of choice, the

alternatives were not yet removed, and at that point he performed a wrong act that culminates in the nonmaximizing outcome (of course at the same time he performed many right acts that culminate in all the maximizing outcomes).

The response I want to make to this case is that although the bad effect was foreseeable (in type if not in token) it was the result of a disposition taken on precisely to minimize effects of that kind. That being so, it seems implausible to claim that in this case there was a temporally extended act of that sort. Indeed, if we allowed it to be the case that there was a temporally extended act of that kind, then at the time of taking on the disposition, countless acts—right and wrong—would have been performed. The principle I invoke here is not the strong doctrine of double effect that all foreseen but unintended consequences should not count toward act individuation, just the weaker one that foreseen and unintended consequences that are expected given the disposition chosen, *insofar as the disposition minimizes consequences of that type*, should not count. This is in contrast to the plague case, where bad outcomes (indeed in the particular example the token biological terror) are the intended consequences. This is no doubt a controversial principle but then act individuation is a very shaky branch of philosophy. I think intuitions of moral responsibility, and the rightness and wrongness of behavior in situations, are much more robust that theories of act individuation. Thus, if someone were to charge me with manufacturing a theory of act individuation to match my ethical theory, that might not count as a complaint: I might be tempted to say it is more an observation that best practice is at work. But we can do better than that. To the extent that this principle is odd, it is because we are carving up a single "many headed" temporally extended act into a collection of individual temporally extended acts. When someone adopts the maximizing disposition, the one natural individuation of temporally extended acts is that there is a single act of doing the things that are the natural causal consequences of being so constrained. Suppose Mary forms a disposition to never act aggressively in certain circumstances, and she does this knowing that she lives in a society which is sufficiently peaceful that this is maximizing. This is now out of her direct control (though perhaps she could retrain slowly if she saw that social conditions had altered). On some occasions she, in line with her dispositions, doesn't respond aggressively and it has poor consequences that were foreseeable at the time of the response—perhaps a senior citizen is mugged as a consequence. While the overall balance of good and bad outcomes was foreseeable at the time of inculcation of the disposition, the individual outcomes were not. At the time of commencement of the original act, there was no deliberation over these particular outcomes. Rather, what was deliberated over was a particular pattern of outcomes. So, the primary temporally extended act is the many headed act that terminates in very many places in space and time. It's no different in this regard from pressing a button that, after much delay, delivers medicine to many different places, in most of which it does good, but in some it does bad. That too is (among other things) a single "many headed" temporally extended act. And both of them are right insofar as they have positive expected value. So, the apparent arbitrariness to my individuation conditions for acts sliced out of this extended whole arises

from needing that individuation to respect the organic and indivisible (from the perspective of deliberation) role that the sliced action plays in the whole unsliced action.

It remains to note that this account of act individuation hands us another nice consequence for the purposes of the previous section. If an agent takes on an evil disposition for the purpose of ensuring bad outcomes, then it is open on this account to impute to her a temporally extended wrong action for each consequence that is on the whole bad. For the extended acts are ruled out only in cases where the acts later chosen fail to promote the values that justified the disposition. Thus, we have another way to undermine the worry of the previous section that no wrong act is performed when the evil agent acts (in the best available way—perhaps trivially because the only way) under an evil disposition. There may be no wrong act wholly located at that time; but there may be a temporally extended wrong action. And because the bad outcome was exactly the kind of outcome that the evil agent was trying to bring about by taking on the disposition, the bad consequences can count as the result of a temporally extended action.

5.2. Agents and Time Slices

So far I have talked as though agents should be thought of as local strings of time slices, just long enough to deliberate. So, for example, one way to think of the comparison in the previous section between the earlier and later stages of the good and the evil agent is this:

> Call the earlier, disposition-changing time slices the "bosses."
> Call the later, acting sets of time slices, the "footsoldiers."

1. Evil Boss chooses (wrongly) to constrain the choices of his footsoldier Elvis.
2. Good Boss chooses (rightly) to constrain the choices of her footsoldier Gladys.
3. Elvis chooses (rightly) to do the best he can in choosing the least evil alternative (which is not the maximizing one unavailable to him).
4. Gladys chooses (rightly) to do the best she can in choosing the best available alternative (which is not the maximizing one unavailable to her).
5. If we allow telegraphic acts, then Evil Boss performs a telegraphic act that is wrong, for there is an act that is available to him which is better.
6. If we allow telegraphic acts, the Good Boss does not perform a wrong telegraphic act, for there is no better alternative available to her.

This may be enough to satisfy many, but some may think that all this talk of time slices misses the point: the real issue is whether the person acts rightly in performing an act. In particular, does the person act rightly in performing the unmaximizing act under the control of a disposition that is maximizing?

Given a metaphysics of temporal stages, then the right thing to say is that it is often a contextual matter that temporal stage's behavior we concentrate on when we make a judgement

of rightness. This explains the mixed feelings that we might have when making judgments about the rightness of acts where the acts are not maximizing, or where they are maximizing but we disapprove somehow. The point is that in all of these cases there are two temporal sequences to consider; and one acts rightly and the other does not. Which one is relevant depends on what purposes we might have: correcting overall dispositions or improving the strength of will at moments of deliberative choice. But where we have no particular purposes in mind, the question "did she act rightly" becomes ambiguous. We are happy to answer yes when all stages that are relevant act rightly (as when the boss and the footsoldier act rightly) and we are happy to answer no when all act wrongly (as when the evil boss acts wrongly, and the footsoldier does not choose the least bad alternative) but our moral evaluations may be more confused or nuanced when the evaluations of different relevant time slices are out of synchrony. I take it that is a strength of my view that is explains why these cases are disturbing.

6. One Last Objection: Overriders

An objection which is frequently pressed against this view in discussion is that it seems that the problems are not solved, only forestalled. It's all very well to say that on this view we act rightly in preferring the near and dear, or respecting our core projects. But that's all very contingent. What if we had available to us ways of overcoming our bound dispositions. What if there were a pill we could take, which would allow us to sacrifice our wife for the good of the half dozen potentially thriving strangers. Wouldn't we then, as act consequentialists, be obliged to take it?

It depends. If our dispositions are very simple—we have inculcated a disposition not to sacrifice our projects or our close relationships, but have not inculcated any disposition to protect that disposition against local disruption—then yes, on this view we should take it. But if we live in a world where there are such override pills (or more plausible mechanisms) available, then we have inculcated the wrong strong dispositions: it needed to include a disposition not to take any such pills (unless, perhaps, we see that we are in a world where the very long-term thriving of the world depends on it).

A related worry is that if we are faced with a pill (again, for convenience, standing in for other more plausible things), which could remake us into maximizers that reject any meaningful projects except for maximizing, and eschew all close relations for fear they may get in the way of maximizing, doesn't the view say we should do that? After all, even if the pursuit of meaningful projects and relationships are key among the goods we are trying to maximize, perhaps we can promote more meaningful projects and relationships by eschewing them ourselves.

Again, it depends. If our expectation is that this psychological change would indeed create great utility (which plausibly it wouldn't) then perhaps the answer is yes. It's the difficult decision to become a kind of moral saint, whose life adds little to the sum of utility in itself,

but whose actions add much. We might selfishly hope that such pills or other processes are rare in the world, or at least rare in our neighborhood. But what if they are not rare? Should be all become so transformed? Here the answer is no, on plausible assumptions about what goods we should be trying to maximize. Perhaps one saint, or a few, will make the world a better place. But if we all become saints of this stripe then the world will be no better place. So, the moral decisions will depend on some kind of complicated equilibrium: there will be a right proportion of saints. And the decision as to what to do will then have the complicated game theoretic features that decisions always do when the desired outcomes are an equilibrium state, and one's own decision only a contribution to that.

7. Conclusion

So the judgments in cases that plausible kinds of indirect consequentialism may deliver can be reconciled with an act consequentialist theoretical framework, just so long as we take seriously the causal impact that the moral dispositions that we adopt have on our future capacity to deliberate.

This picture has more far-reaching benefits that simply getting the cases right and making the theory neat. It also, I think, is a far more convincing account of moral life. Moral life is complex: sometimes it really is the case that agonizing calculation and decision has to be made. These are cases where we are relatively unconstrained by moral dispositions, and deliberate on a case-by-case basis. Unusual moral situations will be among the causes of such case-by-case deliberation. And if this is a very different kind of deliberation, then I think it an advantage of my account that it makes it so theoretically.

Weak moral dispositions are also a common feature of our ethical life. We act on a moral autopilot, but feel always able to step in and intervene if we notice that the case is special. When we follow the autopilot, we get none of the phenomenology of being restricted in our choices by who we are.

But there is a widespread class of acts where we may get to choose, but from a constrained set of options. Some options are excluded because, it seems, of who we are, or because of our deep moral dispositions. Even if we see that the act we are about to perform doesn't have as good consequences as another physically in our power, we feel driven to it. This is satisfactory if the deep moral disposition remains one that we judge best. On my account, we still act rightly. That we are not free to choose outside this range is testified to by the fact that if we do decide that our disposition is not a maximizing one, we can't just immediately choose differently. It takes work on the self to reform, to change one's moral dispositions.

I think all of this is of great import to the Gibbardian project. It was too easy to read the picture of *Reconciling Our Aims* as a kind of proof that the state of the world that instantiates most value is a kind of dystopia where acts that seem wildly immoral to the disinterested theoretical observer are widely countenanced for their contribution to a total value on the measures many might accept. This might have been thought to be even worse than

the usual worries about consequentialism. The usual worries just say that a consequentialist moral theory will say that some right things are wrong, and some wrong things are right. But if we accept much of what Gibbard says, and still stand by our first-order moral judgments, we get the even more worrying conclusion that the state of the world that has most value on less controversial measures is also one full of unacceptable acts. If one clung to one's first-order moral judgments, this might lead not just to a rejection of consequentialism as a theory of judging acts but also to a rejection of the theory of value, which says that this total state is the best state. But by reconciling our first-order moral judgments in most cases with act consequentialism, direct-binding consequentialism allows our global judgments of the best overall state to cohere with our judgments about rightness or otherwise of most of the actions that bring it about: making the moral universe safer for us to reconcile our aims.

So when, of all the things we are able to do, we do the thing that has the best expected outcome we act rightly. That is the key insight of consequentialism. What I hope has been added here is the observation that it is our own past choices that influence what the best we can do is, by influencing our future abilities. We act freely within the bounds imposed by those choices. What once we may have been able to do we sometimes cannot, and for good reason. And what we now cannot do, we can sometimes come to be able to do, if that would be best.

References

Adams, R. (1976). "Motive Utilitarianism." *Journal of Philosophy* 73: 476–81.

Braddon-Mitchell (2003) "Qualia and Analytical Conditionals." *Journal of Philosophy* 100 (3): 111–35.

Driver, J. (2001). *Uneasy Virtue*. New York: Cambridge University Press.

Dworkin, G. (1970). "Acting Freely." *Nous* 4: 367–83.

Elster, J. (1979). *Ulysses and the Sirens: Studies in Rationality and Irrationality*. Cambridge: Cambridge University Press.

Frank, R. H. (1989). *Passions Within Reason: The Strategic Role of the Emotions*. New York: W.H. Norton.

Frankfurt, H. (1971). "Freedom of Will and the Concept of a Person." *Journal of Philosophy* 68: 5–20.

Gauthier, D. (1986). *Morals by Agreement*. New York: Oxford University Press.

Gibbard, A. (2008). *Reconciling Our Aims: In Search of Bases for Ethics*. Oxford: Oxford University Press.

Jackson, F. (1991). "Decision Theoretic Consequentialism and the Nearest and Dearest Objection." *Ethics* 101, no. 3: 461–82.

——— (1998). *From Metaphysics to Ethics: A Defence of Conceptual Analysis*. Oxford: Oxford University Press.

Jeske, Diane, and Fumerton, Richard (1997). "Relatives and Relativism." *Philosophical Studies* 87: 143–57.

Latham, Andrew J. (2019). *Indirect Compatibilism*. Doctoral dissertation at the University of Sydney.

Lyons, D. (1965). *The Forms and Limits of Utilitarianism*. Oxford: Clarendon Press.

Mintoff, J. (1997). "Rational Cooperation, Intention and Reconsideration." *Ethics* 107: 612–43.

——— (2000). "Is Rational and Voluntary Constraint Possible?" *Dialogue* 39: 339–364.

Nye, Howard, David Plunkett, and John Ku (2015). "Non-Consequentialism Demystified." *Philosophers' Imprint* 15:1–28.

Pettit, P. (1991). "Consequentialism." In *A Companion to Ethics*, edited by Peter Singer, Ch. 19. Wiley-Blackwell.

Pettit, P., and G. Brennan (1986). "Restrictive Consequentialism." *Australasian Journal of Philosophy* 64: 438–56.

Pettit, P., and M. Smith (2000). "Global Consequentialism." In *Morality, Rules and Consequences*, edited by B. Hooker, E. Mason, and D. Miller, 121–33. Edinburgh: Edinburgh University Press.

Railton, Peter (1984). "Alienation, Consequentialism, and the Demands of Morality." *Philosophy & Public Affairs* 13 (2): 134–71.

Sidgwick, H. (1930). *The Methods of Ethics* (reprinted 1981). Indianapolis, IN: Hackett.

Smart, J. J. C. (1973). "An Outline of a System of Utilitarian Ethics." In *Utilitarianism: For and Against*, edited by J. Smart and B. Williams. Cambridge: Cambridge University Press.

Stamp, D., and M. Gobson (1992). "Of One's Own Free Will." *Philosophy and Phenomenological Research* 52: 529–56.

Star, Daniel (2010). "Review of *Reconciling Our Aims*." *Philosophical Review* 119, no. 2 pp. 259–263.

Williams, B. (1973). "A Critique of Utilitarianism." In *Utilitarianism: For and Against*, edited by J. Smart and B. Williams. Cambridge: Cambridge University Press.

VII

RESPONSE FROM ALLAN GIBBARD

22

REPLY TO COMMENTATORS

Allan Gibbard

I thank all the commentators for this magnificent collection. I am humbled by the great care and attention that such fine philosophers have devoted to my efforts. To many of the commentaries I wish I had better things to say in response, but my hope of course that even if I am mixed up on many issues, the upshot of these exchanges will be to help us see some things more clearly, not to complicate matters unhelpfully.

Much of my philosophical writing has consisted, we might say, in dabbling. When I prepared to write my PhD thesis at Harvard, I resolved to write on cultural relativism in ethics. That suited me especially, I thought, in light of my two years of experience in Ghana in the Peace Corps—though my encounters with traditional African culture were spotty. (Achimota School, the secondary school where I was assigned to teach mathematics and physics, was established in the Gold Coast (which became Ghana) a hundred years ago by the first governor general to recognize black Africans as capable of advanced education. Some of my colleagues and most of my students were Ghanaian, but of course their education equipped them with cultural backgrounds tied to Europe as well as Africa.) By the end of a summer of struggling toward a thesis topic, I still hadn't figured out what new I could say. I feared that if I stayed in graduate school longer than the four years my fellowship covered, then like many of my colleagues, I would have to teach the general humanities introduction Humanities 5, and have scant time for my own writing. I then realized that I could expand and develop a paper I had written in Richard Brandt's moral philosophy seminar at Swarthmore and revised and published while I was in Ghana. And so that's what I did, finishing a draft of the thesis the morning I was to drive off for the University of Chicago and my first university teaching job. As I and a few of my fellows had looked for jobs at the December 1968 convention, we had learned to our shock of a change in the job market: suddenly you

couldn't get a job if you hadn't *started* your thesis. Fortunately, I could send interviewers drafts of substantial parts of mine, and the upshot was I got job offers.

As an undergraduate at Swarthmore I had majored in mathematics and minored in physics and philosophy—though I could have qualified officially as a philosophy major if I had chosen to count my symbolic logic seminar as philosophy rather than math. Swarthmore was wonderful for me, philosophically and otherwise: in those days a small college could have departments that, apart from size, equaled those of major universities. My moral philosophy seminar was taught by Brandt; my "Contemporary Problems in Philosophy" seminar was taught by Jerome Shaffer, an especially brilliant teacher who had drawn me into philosophy my first semester; my symbolic logic seminar was taught by a new junior faculty member Jaegwon Kim. (He told me later that he had been new to the subject, but I could never have guessed. Our main textbook was Alonzo Church's fearsome volume *Introduction to Mathematical Logic*, and W. V. O. Quine's *Methods of Logic* was our accessible relief.) I feared trying to make philosophy my life's profession, wrongly thinking that math or physics or law would be more secure, but I finally decided to take the plunge. A result of all this was that when I entered the Harvard PhD program after my stint in the Peace Corps, I had a fine background in contemporary analytic philosophy (though somewhat weak on Wittgenstein and Austin), but no background in the previous history of philosophy apart from what I had been able to puzzle out from my own reading.

At Harvard, I had wonderful philosophy teachers, as one would expect. Roderick Firth, W. V. O. Quine, and John Rawls were most important to me. I also had an unexpected stroke of luck: In Ghana, just after I had accepted my admission to the Harvard graduate program, a family friend dropped by on his way home to West Virginia from East Africa, and he told me, "No! Don't go into philosophy; go into economics." If I went into philosophy, he argued, I would be confined to university teaching, whereas in economics I would have a wide choice of careers. I was convinced—but I worried that I had perhaps wasted too much of my life before moving on to graduate school, and so I proceeded on the course of life I had planned. My interest in economics had been piqued, however, enough so that I added the graduate introduction to mathematical economics to my program. One afternoon in the little Philosophy library at Harvard, I had stumbled across the little blue book by Kenneth Arrow, *Social Choice and Individual Values*. I was fascinated. Then I learned that not only was Arrow coming from Stanford to Harvard, but that he and Rawls would be teaching a joint seminar on social choice theory with a young economist from India I'd never heard of, Amartya Sen. I spoke to Rawls about it and he firmly said no, that it was only for economics graduate students—but I insisted. Along with us six graduate students, the seminar was attended by stars of economics and decision theory, including Howard Raiffa of Luce and Raiffa, *Games and Decisions*, as well as the editor of *Econometrica* and a young economist who people whispered had been national bridge champion, Richard Zeckhauser. I sat in the seminar gaping at the spirited conversation and thinking, "Never before have I experienced so much intelligence gathered in a single room!" I myself dared speak up only a couple of

times, and when I sought Arrow out to argue with him, he told me to come back during his office hours; when I did, he wasn't there.

Sen gave a presentation showing that if one relaxed Arrow's requirements on a social choice function to require not transitivity but "quasi-transitivity"—transitivity of strict preference—impossibility changed to possibility. I thought about this and fiddled around, and concluded that the only social choice mechanisms that met these criteria were woefully unsatisfactory. They must take the form of what I called a liberum veto oligarchy: there's a group such that if they have a unanimous preference, that preference prevails, but if any of them have opposing preferences, then the society is indifferent. This would mean that no matter how overwhelming a majority, if one person dissented, the result would be social indifference. The time came for three sessions for the papers of us six graduate students, and apart from Raiffa and the three professors whose seminar it was, the eminent economists disappeared. (I have always been grateful to Raiffa—and when I got to know him, I learned that he had been a student at Michigan of my statistician father-in-law.) Sen gave me a xerox of an article of his on the subject, though it was only much later that I learned it had been published. (It wasn't so easy to track down papers in the days before computers, especially if you didn't already know they were published.)

A substantial portion of my initial publications stemmed from this seminar. Social choice theory, as devised by Arrow, has its home in all three of economics, political science, and philosophy. It concerns how social choices can depend on individual preferences and dispositions to action. My first year at Chicago, I investigated something I thought I remembered Arrow saying: that his condition the Independence of Irrelevant Alternatives is equivalent to immunity to strategic manipulation. I worked on how to prove this so that I could include the proof of the theorem in the tiny graduate seminar in social choice theory I was teaching, but to my surprise, proof stubbornly eluded me. Finally, one weekend visiting my fiancée in Urbana, I found the proof. Only trivial, vastly unsatisfactory social choice schemes preclude strategic manipulation. That summer, I worked this result up into a paper for publication. I also figured out how to make the result far more general, applying to all forms of social interaction.

The upshot of my education was that when I began my philosophy teaching career at Chicago, my knowledge of philosophy was spotty, above all in its history, but I had begun to have a reputation among a few top economists. What I went on to write was largely opportunistic; I worked where I thought I could tackle something significant. Every so often, a theme would emerge and give rise to a series of articles or even to books, but many of my publications were scattershot, not fitting into any grand, unified scheme. From my thesis, I developed only one paper, "Act-Utilitarian Agreements", and even though my thesis won an award, I subsequently kept finding that there were more important things to be pursuing.

One example stemmed from another thing I had done at Harvard, namely to question Quine's critique of Carnap over "individual concepts". I did a short paper for Quine's course "Word and Object", and thought I had come up with something pretty neat. Quine found

unclarity on the first page, and gave it a *B*+. I worked all summer to develop the paper, and gave it to Quine. To my surprise, when I saw him a couple of days later and asked if he had had a chance to look at it, he invited me into his office and said, "Now I see what you're doing. It's all in Føllesdal's thesis." Dagfinn Føllesdal was a Harvard PhD (he worked with Quine) and so his thesis was in the library, and when I went there to look at it, I could see that Quine was pretty much right. I thought still, however, that I was on to something in a Carnapian vein, and eventually, while I was visiting at Pitt, I showed my effort "Contingent Identity" to Rich Thomason. He solicited it for the *Journal of Philosophical Logic*, but demanded a much more explicit development of what I was saying. I complied, although I thought that what I was saying should clear enough without the detailed working out. Subsequently, Kit Fine and I sat beside the swimming pool at a conference in Miami Beach pursuing the issues. He wasn't convinced of my way of doing things, but did realize that he had to cope with examples like mine (the statue of the infant Goliath and the piece of clay Lumpl). His paper on this appeared a quarter century later.[1] I took it that normally, a statue and the piece of clay it is made of are distinct, since they don't coincide in space-time, but this, I argued, is contingent—giving us a sense in which identity can be contingent. Fine considers them distinct even when they fully coincide in space and time, and that, I think, is nowadays the predominant view.

Social choice theory and contingent identity, along with various other topics, don't figure in this volume. The organizer-editors had to make choices, and I very much approve of the choices they made. Do unified themes emerge in the pieces that commentators in this volume scrutinize? As I think about it, I think I do find some loose strands. The broadest stem from the nature of philosophy, especially as practiced in the analytic tradition. For any question, we may seek an answer, but we can also worry what the question means if anything, how we might discern its answer, and how to conceive what we are thereby doing. These were the kinds of challenges in Mr. Shaffer's class my first term in college that drew me to philosophy so urgently.

Many of the topics discussed in this volume are normative, in that they concern what's rational—a term Brandt used copiously. I understand rationality as, in more everyday terms, a matter of what it makes sense to do, to believe, and to feel about various things. This faces us with questions both of meaning and of substance: what normative assertions mean, and which ones are true. For the meanings of normative assertions, I have worked on a form of expressivism: questions as to what's rational in action, belief, and feelings amount to questions of what to do, believe, and feel. I thus explain normative thought and language in terms of states of mind that assertions express. For morality, I now accept analyses of Howard Nye, who holds that in a narrow sense, questions of moral right and wrong are questions of how to feel obligated to act. As for substantive answers to broad normative questions, for action I look to instrumental expected utility rather than epistemic auspiciousness. (In cases like

1 "The Non-Identity of a Material Thing and Its Matter", *Mind* 112 (2003): 195–234.

Newcombe's problem and the prisoner's dilemma with one's twin, they diverge.) For degrees of belief to guide actions, I am a Bayesian—not regarding the aims and degrees of belief we in fact have but the ones to have. What we should be working on isn't knowledge and full belief, but degrees of belief and action under uncertainty. For morality and justice, I tend toward utilitarianism, though perhaps of an indirect sort. This, I contend, is what contractarianism delivers and is the only kind of view that fits coherently with our considered judgments. We need a social ethos that benefits people.

Many of the questions the commentators raise concern language, especially meaning and truth. I have argued that subjunctive and indicative conditionals work quite differently from each other, and the conditionals that rationally guide action are subjunctive. Truth is best understood in a way that is mostly deflationary but not entirely so: in an important sense, neither indicative conditionals with false antecedents nor epistemic mights qualify as true or false. I have experimented with interpreting the concept of meaning as itself normative, following in some respects Kripke's Wittgenstein and holding that questions of meaning amount to questions of which sentences to accept if one is the speaker. Putting it this way draws on Paul Horwich, as does much else in my thinking, but he and I go off in different directions to explore the concept of meaning. In this volume, I concede a lot to Horwich's nonnormative explication of meaning; I now think that both his ways and mine are eligible. I write these things and others with vastly insufficient knowledge of the literature in linguistics. My excuse for thinking I'm not entirely wasting readers' attention is that from time to time, I have found ways of analyzing linguistic phenomena that don't seem to be in the extant literature. Examples are expressivism, the ways that indicative and subjunctive conditionals require sharply different explanations, and the communication of normative states of mind.

I have been interested not only in normative questions, their answers, and their meanings, but on our natures as organisms who question and come to answers. That includes our natures as having moral and other normative convictions and being guided by them in our actions, beliefs, and feelings. We gain insight by inquiring how normative judgments fit in with our being shaped by evolution, as interacting organisms that influence each other. That isn't a chief focus of the commentaries, but it does come up with Nye and Rosati.

All this said, I turn to responses to specific commentaries, organized into the six sections of the volume.

1. Norms in Decision and Belief

1. William L. Harper (University of Western Ontario) "Decision Dynamics and Rational Choice"
2. Melissa Fusco (Columbia University), "Counterfactuals and the Gibbard-Harper Collapse Lemma"
3. Zoë Johnson King (University of Southern California), "Who's Afraid of Normative Externalism?"

4. Daniel J Singer (University of Pennsylvania) and Sara Aronowitz (University of Arizona), "What Epistemic Reasons Are For: Against the Belief–Sandwich Distinction"

Both Harper with his collaborators and Fusco write on "causal" decision theory. Johnson King and Singer/Aronowitz write on other topics. So I begin with "causal" decision theory.

William Harper (with Brian Skyrms and Sona Ghosh)

A couple of preliminary points: First, I can't repeat too often, the "causal" view that Harper and I support can't be credited to us. We buried our credit to Stalnaker too far in a footnote, and we have regretted ever since that we didn't make the credit prominent enough. Stalnaker's letter proposing his formulation with conditional propositions is available for all to see in the volume *Ifs*, coedited by Harper—but when Harper and I published our paper the letter wasn't yet published.

I'm putting the term 'causal' in scare quotes, because I think that 'causal expected utility' is not a good term for what to maximize in one's choice of actions. Jaegwon Kim's paper "Noncausal Connections" prompted me to realize this. Many relations besides causality can feed into this kind of expected utility we advocate. So I would prefer to call it not *causal* expected utility but perhaps *instrumental* expected utility. Consider: In many circumstances, nodding one's head indicates agreement. So if one were to nod one's head, one would be indicating agreement—a subjunctive conditional. But not directly a causal one, since nodding one's head doesn't *cause* one to be indicating agreement. Such noncausal subjunctive conditionals could very well figure in the instrumental expected utility of an act. (Which ones do would depend on what the intrinsic value realized by acts consists in—on such matters as whether indicating agreement with praising an atrocity intrinsically bad.) In short, then, the expected utility we claim relevant to what to do isn't properly "causal" but instrumental. The kind of evidential expected utility that Jeffrey developed we might call *auspiciousness*.

Turn then to William Harper, joined by Brian Skyrms and Sona Ghosh. Where I have opinions, I am pretty much entirely in agreement with them—though I'll add some comments and reminiscences. I am in awe of the remarkable, detailed work that the three of them accomplish in advancing our understanding of these difficult issues.

I learned decision theory from Bill Harper. I had already known about von Neumann-Morgenstern utility, and when I visited Stanford for the fall term, 1972, Pat Suppes said at our first lunch, "You do decision theory." I had to ask him, "What's decision theory?" He told me about Jeffrey and Savage, whose work I didn't yet know, and said he preferred Savage. The fall after I visited Stanford, I spent a visiting term at Pitt, and got my first substantial impression of Bill, most of a year before his first substantial impression of me. Rich Thomason told me shortly before Christmas that Bill Harper was coming to town, and I should join in the spaghetti dinner that Rich and Sally were having for him at their house. As Bill says, I had already met him briefly, but we hadn't really gotten to know each other, and so in my mind,

this dinner was the beginning of our intense friendship. Suddenly, at the end of the dinner, Rich brought out the little girls' toy blackboard and directed Bill to give a talk.

I then heard for the first time about David Lewis's triviality proof that no propositional if-then connective could make it the case that in general the probability of the conditional proposition formed with it was the corresponding conditional probability. I heard for the first time of Stalnaker's response proposing moving from conditional probabilities to probabilities of conditional propositions, "If I were to do *A*, then *C* would happen." I sat there gaping. The next year was when I moved to Pitt "permanently" and Bill visited for the year, teaching, among other things, a seminar on decision theory. I sat in on the seminar avidly, and learned what Suppes had been talking about. We also, as Bill relates, saw each other intensively that term—the start of our friendship as Bill remembers it. Our joint paper started with discussions we had at a conference at Western Ontario the next summer that Bill helped organize.

The paper was very much the work of both of us, proceeding from our intense discussions of the things that Bill had known about and I hadn't. We conceived of it first as directed to Jeffrey, but since the background in the Lewis-Stalnaker interchange that Bill had witnessed was unpublished, the real news in the paper wasn't the little that was original with us, but our relaying of the Lewis-Stalnaker points.

The "Death in Damascus" example was from me, though the point it conveys had come up in Bill's and my intense discussions of Stalnaker's proposal. I had heard the story as a child from my mother, but I had no idea where it came from, and only read O'Hara's *Appointment in Samarra* much later. I find decision instability and the nature of mixed strategies in zero-sum games mysterious, and it is wonderful to have Bill straightening so much out. Amidst being honored by all the fine papers in this volume, I am honored by how much significant original work by Harper, Ghosh, and Skyrms appears in this paper.

Stalnaker's proposal should have revolutionized statistics, but to this day, when I chance to talk with a statistician, I get the impression that the lesson has not been absorbed. Since Bill's and my 1978 paper, there have been at least two major sets of treatments that get matters right: the Carnegie-Mellon group of Glymour, Scheines, Spirtes, and Kelly, starting partly with the 1991 paper by three of them,[2] and many works by Judea Pearl starting with *Causality* (2000). When I thought to work further on these issues, I got a surprise. What little statistics I knew I had learned at my mother's knee: how to get a correlation coefficient on a desk calculator, and above all, that correlation isn't causation, and that you can learn about causation from controlled experiments. (She was working on a master's degree in industrial relations to make herself employable now that we kids had gotten bigger. But she failed in that aim because of nepotism rules; positions she would have qualified for were deemed too close to my Dad's.) Her lessons on statistics prepared me for Stalnaker's proposal, and I thought that

2 Peter Spirtes, Clark Glymour, and Richard Scheines, "From Probability to Causality", *Philosophical Studies* 64, no. 1 (Oct. 1991): 1–36.

causation must loom large in statisticians' minds. I thought that they would have much to say on why controlled experiments can force convergence among people who start out far apart on degrees of credence about causality—as de Finetti showed that credences in non-causal claims converge as evidence accumulates. (I follow David Lewis's widely emulated practice of calling degrees of credence or subjective probabilities plain "credences".) Now, my father-in-law was a statistician, and after his death I had a bookcase in my office filled with his books, and so I could look through these to see what statisticians had to say about evidence for causality. *Almost nothing* was the answer, to my surprise. Pearl resolved my mystification, with his treatment of how Fisher had opposed the whole notion of causality and managed to get it banned from respectable statistical publication. I'm still amazed that the extensive lore about properly controlled experiments hadn't until then been backed up by any careful analysis of why controlled experiments yield causal knowledge—though maybe there's something I didn't find. Once I learned, though, that what was needed had been done, if not properly assimilated, I didn't much work further on this.

Bill has done terrific work on ratifiability of decisions. I haven't kept up with the literature on this, and so I won't comment on this topic. I find the discussions between Harper, Arnzenius, and Joyce fascinating, but I am baffled about what to say about decision instability, and puzzled about what to say about cases where the ratifiable strategy only ties for optimality—that is, where if one expects oneself to adopt that strategy, then alternative strategies have expected utilities just as great. This is central to the theory of zero-sum games, but I won't try to add to the conversation any of my own feeble judgments on these matters.

Instead, let me turn to Harper's footnote 5, on which I do have something to say. He asks for a response from me:

> John Cantwell (2010) has argued that future tensed indicative conditionals with acts under consideration for choice as antecedents and outcomes of choosing as consequents are appropriately evaluated as subjunctives.

I have long been inclined to believe something like this, but my recollection is that when I have tried to prove it with intuitive reactions to examples, I have gotten inconclusive results. I take this up in my paper of 1981 that Bill mentions, "Two Recent Theories of Conditionals" (in secs. 4–5, 222–29). In that paper I chicken out and explicitly set aside future conditionals like "If Oswald doesn't shoot Kennedy, someone else will," because I thought I couldn't establish an ironclad conclusion about these. I now think, however, that we can get a pretty strong indication that Cantwell is right. I lay out a lot of the relevant analysis in that 1981 paper of mine, but let me here explore whether we can get this stronger conclusion.

I appreciate that what we call "subjunctive" and "indicative" conditionals isn't what linguists mean by the terms at least some of the time. But we need some terms for the important distinction we use them to make, along with other writers on decision theory, and so I'll use

the terms or misuse them for decision-theoretic purposes. Our prime examples are from Ernest Adams:[3] the subjunctive conditional,

If Oswald hadn't shot Kennedy, someone else would have,

in contrast with the indicative conditional,

If Oswald didn't shoot Kennedy, someone else did.

Subjunctive conditionals have two key features:

(1) a form of 'will' (since 'would' is the past tense of 'will'), and
(2) a tense shift in the antecedent toward the past.

That is to say, although the time the shooting would or would not have occurred is the simple past—the time when Oswald in fact did shoot Kennedy—the verb in the antecedent is put as 'hadn't shot' rather than 'didn't shoot'. So grammatically, I say, this conditional is a past form whose present form is "If Oswald doesn't shoot Kennedy, someone else will." I assert, indeed, that from a single schematic stem conditional like

If Oswald not shoot Kennedy, someone else do shoot him

we get 32 conditionals with various combinations of tenses, such as

If Oswald doesn't shoot Kennedy, someone else will.
If Oswald didn't shoot Kennedy, someone else would.

In the 1981 paper, I call conditionals with these features *grammatically subjunctive*, and contrast them with others that are *grammatically indicative*, prime examples of this last being

If Oswald didn't shoot Kennedy, someone else did.
If Oswald hasn't shot Kennedy, someone else has.

For past tense conditionals, I claim, a systematic thesis emerges: that grammatically subjunctive conditionals are nearness conditionals and grammatically indicative conditionals are epistemic conditionals. By "nearness conditionals" I mean ones that, at least roughly, work as Stalnaker and Lewis depict, and by "epistemic conditionals", I mean ones that, at least roughly, work as Adams depicts. These two varieties of conditionals, I argue in the

3 Adams, Ernest (1970). "Subjunctive and Indicative Conditionals." *Foundations of Language* 6(1), 89–94.

paper, epistemic and nearness conditionals, have remarkably little in common apart from substantial parts of their logic.

What, then, of the future tense, which Cantwell brings up? Our prime example is

If Oswald doesn't shoot Kennedy, someone else will.

"Will" conditionals like this I'm classifying as "grammatically subjunctive". Cantwell says, "Future tensed indicative conditionals behave more like (are semantically closer related to) past looking subjunctives than past tensed indicatives."[4] This might fit my claim that "will" conditionals are nearness conditionals, like past tense subjunctive conditionals.

How can we test whether a conditional like this really is a nearness conditional? For a past tense conditional, that's what I tried to do with my Sly Pete example. In the story, observer Zack knows that Mr. Stone's hand is remarkably strong, and so he gives high credence Pete's having a losing hand. So, the nearness conditional "If Pete were to call, he would win" gets low credence from Zack. But also, since he knows that Pete knows Stone's hand, Zack strongly expects that Pete won't call unless he has a winning hand. So on Zack's view, if Pete's going to call, that's a sure sign that he has a winning hand, and so the epistemic conditional "If Pete's going to call, he's going to win" merits high credence. Return then to "If Pete calls, he'll win." If it's a nearness conditional, Zack will accord it low credence, whereas if it's an epistemic conditional, Zack will accord it high credence. What credence Zack accords it thus indicates whether it is a nearness conditional or an epistemic conditional. My hypothesis, extended to "will" conditionals in general, is that they are nearness conditionals. They are like the past tense subjunctive conditional "If Pete had called, he would have won." Observer Zack will thus accord it low credence.

So if you, dear reader, thoroughly understand all this, I encourage you to apply your linguistic intuitions. Should observer Zack accord low credence to the thought "If Pete calls, he'll win," because Pete likely has a losing hand? If so, that supports my hypothesis that such a "will" conditional is a nearness conditional. Should Zack instead give the thought "If Pete calls, he'll win" high credence, because Pete's calling would indicate that his hand is a winning one? If so, that tells against my hypothesis.

I can't dictate which way to hear the sentence. But I myself find that if it's an epistemic conditional, it's misleadingly put. It would be better to say,

If Pete calls, that will mean he's going to win.

So my own intuitions favor the low credence for "If Pete calls, he'll win," since Pete almost surely has a losing hand. This favors the Cantwell's suggestion that, as I'd put it, a future tensed conditional like this is semantically subjunctive.

4 Cantwell paper on Egan (2010, 7); the paper is Cantwell, "On an Alleged Counter-Example to Causal Decision Theory", *Synthese* 173, no. 2 (March 2010): 127–52.

Harper mentions Brian Skyrms's 2013 treatment of subjunctive conditionals in terms of "an appropriate *K* partition of the relevant alternative chance set ups".[5] Here wouldn't be the place to address the Skyrms paper in any substantial way, but I'll give my off the top of my head reaction, the only point on which I have reservations about what Bill says. It's going to be crucial what constitutes an "appropriate" partition. ("Natural" is another term that Skyrms uses.) One major point of the Gibbard and Harper joint paper was that if we're to invoke partitions for the kinds of purposes Skyrms has in mind, we'll have to place a cause-laden requirement on what qualifies a partition to serve this role. Our prime example was about David and Bathsheba (159–60): David covets her, but fears that taking her would provoke a revolt. With the wrong partition, we get a faulty argument from dominance of a kind that Jeffrey denounces: Where proposition *R* is "There will be a revolt," suppose we partition the possibilities as *R*, ¬*R*. David prefers having Bathsheba in either case. Given either, the conditional probability of matters going as he prefers is higher given that he takes her than it is given that he abstains. The requirement we laid down to rule out such a partition was in terms of subjunctive conditionals; that's doubtless not the only way to put what's needed, but we must have some restriction or other. Harper says of the Skyrms proposal, "This allows conditional chances to do the work needed for causal decision theory without requiring any more detailed assumptions about the semantics and metaphysical commitments of subjunctive conditionals" (9). But we do need to specify in some way what kind of partition will do the job properly. The Savage framework, as we noted in our 1978 paper, involves a partition into "states" that must be in some sense act-independent, but some experts I have talked with insist the independence should be instrumental and some that it should be evidential. We showed in the paper that requiring states to be act-independent instrumentally yields instrumental expected utility, whereas requiring their act-independence to be evidential yields auspiciousness. Often, of course, both requirements are met, but not in zero-sum games where one's choice indicates what the opponent was expecting. Invoking an "appropriate" partition, then, won't let us do an end run around the need for subjunctive conditionals or something else that can do the same job; we must specify what this appropriateness consists in.

Again, let me express amazement and appreciation for the separate contributions of these three investigators to Harper's paper.

Melissa Fusco

Melissa Fusco, in "Causal Decision Theory's Revenge", says many things that I agree with. With some of the things she says, however, I'm in substantial disagreement, as I'll try to explain.

Although "causal" decision theory had enjoyed wide acceptance, she says, Egan's paper has figured in a sea change, and "the tides are turning." Both Harper and Skyrms in this

5 Brian Skyrms (2013), "The Core Theory of Subjunctive Conditionals", *Synthese* 190: 923–28.

volume address Egan's contentions (32), and I very much agree with the things they say. Brad Armendt too argues the case against Egan, in a recent paper—convincingly to my mind.[6] Whether predominant opinion is changing away from "causal" decision theory, or whether it was before these current works countering Egan, I don't know, but I take Harper, Skyrms, and Armendt to show that if opinion had previously been with us, then whether or not it has been changing, it shouldn't have been. I won't comment much further on Egan, since to my eye, these writers have done the job. I will, though, point out a crucial contention of Fusco's that I'm convinced is mistaken.

Egan, I insist, makes an error. In choosing among alternatives, one must apply the same credences to relevant factors in considering each alternative one contemplates. Let me call this the *proper comparability dictum*. In choosing between alternatives *A* and *B*, one can't sensibly evaluate *A* relying on one body of evidence and *B* relying on a different body of evidence. The central thing wrong with evidential decision theory is that it violates this proper comparability dictum. Not everybody finds what I'm saying convincing, I acknowledge; some find evidential decision theory reasonable, and thus reject the proper comparability dictum. But if you heed the dictum and so don't adopt evidential decision theory straightaway, why would you "Eganize" the counterfactuals that figure in the instrumental theory? To do so would be to make the same mistake—violating the proper comparability dictum—as is made by those who go for the evidential theory straightaway.

With the Newcombe choice, heeding the dictum yields coherent verdicts. Whether we go by one's prior credences, or go by the credences one has on taking both boxes, or go by the credences one would have if one shunned the second box, taking both boxes wins. It wins if one uses the prior credences, it wins if one uses the credences one has on taking both boxes or expecting to, and it wins if one uses the credences one would have if one took only one box or expected to. "Eganization" doesn't enter in. In the Death in Damascus case, in contrast, what one believes one is doing bears crucially on what to do, in a way that yields the paradox.

Fusco rejects Egan's claim—a claim I'm inclined to share—that in some cases, "agents should use their anticipated future causal views" and take into account "what they will learn by performing the very act in question" (33). She may well be right to do so. She may well be right that "the Immediate Post-Act Perspective is Practically Unimportant" (48). And if she is, then the Death in Damascus story won't have its paradoxical import. But even if Egan is right on this point and Fusco is wrong, Egan, I say, is wrong to violate the proper comparability dictum.

Fusco appeals to Sarah Moss on updating as communication. I agree that it's very much right to regard memory as communication from a past self to a later self. Moss and I were discussing these things when she was working toward her article and I was finishing my writing of *Meaning and Normativity*. Her article and my book came out the same year, 2012, and

6 Armendt, "Causal Decision Theory and Decision Instability", *Journal of Philosophy* 116, no. 5 (2019): 263–77.

my recollection is that we arrived at treating memory as communication independently. In footnotes referring to the Moss piece, I say that Moss and I don't agree on how communication works, in that Moss's account of thinking and communication is at odds with a central claim of mine: that preserving the proposition a speaker encodes, so that the hearer comes to encode the same proposition as the speaker did, is not a means to communication, but rather a side effect of what achieves communication.

Zoë Johnson King

Johnson King asks how to cope with one's uncertainty as to what the correct moral theory is, and also whether this matters for what one morally must do. On the latter question, she scrutinizes the thesis "normative externalism" which seems to say that one's uncertainty doesn't matter for what one morally must do. She argues however, that what it prescribes isn't different from maximizing expected objective moral value—or at least all this is how I read her. I think she is right, but I'll add a few indecisive comments.

We can read orthodox decision theory is two compatible ways. The central finding of orthodox decision theory as propounded by Ramsey and Savage is that any full contingency plan for how to respond to new information, if it fully meets requirements of coherence, can be put as saying to maximize a quantity we can call expected utility: it will be as if the agent ascribes values to ways things might go, and acts in light of "credences". More intuitively, though, orthodox decision theory tells how form one's contingency plan: Think what matters and how likely relevant factors are, and use these to calculate what to do, feeding them into an expected utility calculation.

This doesn't tell us, though, how to cope with fundamental indecision of the kind that Johnson King is exploring. Can we give directions for how to approach fundamental uncertainty—not uncertainty as to what will happen, what news will arrive, but how at base to deal with it all? One form such instructions might take is just to dictate the answers; that's the spirit in which "Do the right" thing is offered. Saying this, however, won't help us in dealing with fundamental uncertainties—that's Johnson King's central puzzle. Directions for thinking our uncertainties through might go something like this: As I raise questions to myself, certain things seem apparent; call these "intuitions". Intuitions are bound to conflict with each other; how, then, are we to approach matters when they do? At this point, I think, precise directions we could offer give out. I don't see how formal decision theory can much help us; to apply it, we would need to have answers to the very questions we hope it to help us with.

All I can say, vaguely, is that we need to treat our own thinking the ways a good philosophy teacher treats students' thinking. The teacher explores students' ways of pondering issues, offers possible ideas, points out conflicts and confusions in ways we might think about these matters, and encourages responses which are then subject to the same kind of scrutiny. If I right, this can't be regimented as applying decision theory to our thinking.

Johnson King raises questions of the highest general importance. Even if I'm reading her correctly, I haven't come to definite things to say in response, and I'm not sure whether it is possible to do so. As I have indicated, though, I think I accept the definite contention she arrives at: that what normative externalism prescribes isn't different from telling us to maximize expected objective moral value. I thank her very much for this exploration.

Daniel Singer and Sara Aronowitz

"Against the Belief–Sandwich Distinction" is a striking title, but my initial thought was that it addresses different questions from the ones I am calling "epistemic"—important questions, to be sure, but different. I'll pursue this reaction, and then explore a little whether I should stick with it.

As I'm quoted as saying in my paper "The Value of Truth", it is we who aim at truth; beliefs themselves don't literally aim. But I meant that as just the first step from the slogan "Belief aims at truth" toward an intelligible issue worth exploring. The thesis I ended up with was that if my credences are ideally rational, then it's *as if* I were setting them voluntarily to maximize the expected value of a score—a score that's higher the greater credence I accord to truths, and lower the greater credence I accord to falsehoods. But if my belief-forming mechanisms work properly, they aren't in fact under the control of my voluntary actions. Wanting beliefs won't make me have them.

It's intelligible, I agree, to ask how to pursue the goal of true belief (or more properly, high-scoring credences in the sense I discuss in the paper), and so Singer and Aronowitz are discussing intelligible questions—and significant ones. Reasons to act in pursuit of this aim are not, however, what I myself mean by "epistemic reasons". By that term, I have always meant reasons to believe, or more precisely, reasons to have degrees of credence, high, low, or middling. I haven't included reasons to *want* to have high or low credences, or reasons to *act* in order to have credences worth wanting. On this way of understanding the term, "epistemic reasons" don't include reasons to engage in a thought experiment or to consider an issue. I of course agree with Aronowitz and Singer that reasons to consider an issue aren't in general reasons to believe, and so I haven't counted reasons to consider as epistemic reasons in my narrow sense. These reasons are important but different. I have thus accepted what Singer and Aronowitz attack, "the long-standing dogma of meta-epistemic theorizing that epistemic normativity is normativity of belief" (93). Epistemic reasons bear on epistemic oughts, by which I have meant oughts that govern directly what credences to have in various thoughts, not through governing strategies to bring oneself to have degrees of credence in a thought.

Now of course, not much hinges on what to mean by a particular term such as 'epistemic reasons'. And so I shouldn't object strenuously to the meaning Singer and Aronowitz give the term. But I very much need some term or other for what I myself have been calling epistemic reasons—reasons directly for degrees of credence. For one thing, what to do

normally depends on what credences to have in relevant matters, and that goes for what to do in pursuit of having warranted credences. With such questions, then, the sorts of reasons I am calling epistemic come first.

Singer and Aronowitz allow for three states of belief in a thought: belief, disbelief, and suspension of belief. As always, I favor speaking in terms of the full range of possible degrees of credence, from zero to one.

Singer and Aronowitz speak of "what our evidence supports" and "believing strictly more of what their evidence supports" (75–84). This presumably counts as epistemic in my narrow sense, and we can ask how to render it in a framework of credences from zero to one, rather than just believing or not. We can put talk of believing more of what one's evidence supports in terms of scoring thoughts with respect to particular bodies of evidence. I'll make an assumption here that isn't widely accepted but that I think I could argue further for: I'll assume that for any possible body of evidence, there is a credence function that one ideally ought to have in light of that evidence. (By a *credence function*, I mean an assignment of a credence to every thought one might entertain.) That includes the body of evidence that consists of no evidence at all, and the credence function one ought to have in light of no evidence at all will be one's *justified prior credence function*. The credence function that ideally one ought to have in light of a body of evidence will then be one's justified prior credence conditioned on that evidence. We now want to index any credence function one might have by how close it is to this justified credence function—in the words of Singer and Aronowitz, by how close it is to what our evidence supports. Scale credence functions so that the credence function ideally justified in light of one's evidence gets a score of 1. The scores I considered in "The Value of Truth" were scores for credence functions that depended on what's in fact true and what isn't, and these amounted, I argued, to the expected guidance value of a credence function. So for any credence function, we can take the expected value of this score, calculated with respect to one's ideally justified credence function, compared to the score of 1 that one's ideally justified credence function gets. The higher this score, let's say, the closer one's belief state is to "what our evidence supports". This is my explication of degree of closeness to what one's evidence supports.

The degree of closeness we attribute in this way, as an expected guidance value, will depend not only on which credence function is ideally justified in light of one's evidence, but also on the probabilities of encountering various choices, and on which complex of aims (utility function) one ideally ought to have. (Most writers would instead want to put these matters in terms of the aims the person in fact has, and perhaps this can be done. But one's actual aims, like one's actual beliefs, won't in most cases be coherent, and so putting all this in terms of the aims a person in fact has won't be straightforward. I won't inquire further how such a view different from mine could be put.)

In saying these things, I work within a picture of costless, ideal epistemic rationality, not the sort of rationality we can attain. But, of course, it is crucial to inquire into imperfect, attainable rationality. And so it would be fair to ask whether the sharp distinction I'm

making—between epistemic rationality in my narrow sense and rationality in acting to improve my state of belief—survives the move from unrealistic idealization to something realistic. States of mind of kinds we might actually be in don't take the form of perfectly coherent credence functions; they involve various incoherent proclivities to make decisions in states of uncertainty. But just as we can score an ideally coherent state of belief by its guidance value, so we can score a set of dispositions that falls short of full coherence—again, by an expected level of how well that mode of choosing will do. (In that regard, some incoherent complexes of choice dispositions will score higher than some credence functions that are perfectly coherent but foolish.) One's state of belief is thus still scored by its expected guidance value.

This, then, is what I propose when I try to join forces with Singer and Aronowitz to devise a sense of epistemic value that applies to acts as well as beliefs, and is a matter of conduciveness to achieving states of belief that are closer to what they ideally should be in light of one's evidence. And in this sense, I agree, eating a sandwich can, in some circumstances, have high "epistemic value". (Perhaps there's a narrower notion that goes in the direction that Singer and Aronowitz indicate but will never accord epistemic value to eating a sandwich. But if so, I don't know what it is, and so tentatively I accept the contention I find in Singer and Aronowitz that there is no such intermediate notion.)

Singer and Aronowitz, as I read them, argue that we can't reasonably stick to a sense of "epistemic" in which belief states have levels of epistemic value, but actions—even those of directing one's thinking—don't. The most challenging argument they mount concerns the fault in someone who, for the sake of comfort, ignores an elegant and plausible explanation of a result that he has found puzzling. Clearly this fault is epistemic in Singer and Aronowitz's sense, in that it misses an opportunity to raise the score of one's credence function in light of one's evidence. What, though, should I say of the narrower sense in which epistemic reasons must be reasons bearing directly on the credences to have in light of one's evidence. Can one defend the usefulness of this notion in such an application? The issue here doesn't concern ideal costless rationality, since ideally one would have a degree of credence for the proposed explanation, and have it without effort. The question this example raises is how to direct one's mental efforts. Is a sense in which this is not strictly a matter of epistemic rationality useful? Singer and Aronowitz argue that the fault isn't practical; it isn't a matter of directing one's actions irrationally, where actions include ways one directs one's thinking. Comfort and confidence can, after all, be highly worth having.

I'm not entirely sure what to say about this, but here's a try: The fault in this case is one of excessive incoherence in action—a higher degree of incoherence than what one could at best attain. One conceives of oneself as pursuing an explanation of the puzzling phenomenon, but one is failing to take a clearly promising step toward that goal. So the ways one directs one's thinking is at odds with one's image of what one is doing. As I remarked before, changing from ideal, costless epistemic rationality to imperfect, attainable rationality might undermine the narrow sense of epistemic rationality as rational degrees of credence in light

of one's evidence. But I think that even for this challenging case, I can keep to the widespread narrow understanding of purely epistemic rationality and manage to classify the fault of the thinker in this example. It is a case of means-end incoherence in action where one could achieve a lesser degree of such incoherence.

When I first read this paper, I was irritated that in discussing "epistemic reasons" it focused on reasons that I don't classify as strictly epistemic. I thought that just as puttering around as I write in pursuit of rhetorical goals isn't rhetorical, so puttering around in pursuit of epistemic goals isn't epistemic. But as I hope my response makes clear, I eventually came to see great virtues in the challenges they raise. I'm not confident that I have handled them or know how to. How shall we direct our mental efforts in pursuit of epistemically better credences? In answering this, can we move from unrealistic idealization to something that takes better account of our epistemic limitations? I thank these authors for bringing these questions to the fore.

2. Warranted Feelings

5. Simon Blackburn (Cambridge University), "Assessing Feelings"
6. Stephen Darwall (Yale University), "A Gibbardian Account of (Narrow) Moral Concepts"
7. Howard Nye (University of Alberta), "Morality and the Bearing of Apt Feelings on Wise Choices"

Simon Blackburn

Blackburn sets the goal of understanding normative locutions, along with "the qualities of states of mind that they serve to express, on a natural basis" (99). That's a chief goal of mine too, and something I have loved about Blackburn ever since I first encountered him. As we might expect, we have vast areas of agreement—and with Blackburn, I must always search hard if I need to find anything to dissent from. I remember vividly the first time I heard him talk and sat there gaping in amazement at how much he and I thought along the same lines. I'll turn here, as I have before, to the chief remaining point on which we seem to differ.

Blackburn asks whether we need non-Humean materials to achieve the goal of explaining normative locutions on a natural basis. I myself try to explain normative thinking with special states of mind of accepting norms, states whose contents I attempt to characterize through a rough psychological theory. These may count as non-Humean materials. In any case, unlike me, Blackburn doesn't treat such states as playing a systematic, unifying role in normative judgments. Now of course my attempts to rest so much on such states fail if there isn't good empirical support for such a theory—and since specially normative states aren't part of standard psychology, reasonable bets may have to be against there being anything special about norm-infused states of mind. If there does

turn out to be strong enough empirical support for my conjecture, though, that should be good enough for Hume and for Humeans, even if it doesn't count strictly as Humean. If my way fails, we'll have to try to proceed as Blackburn does, but as I'll explain, I find that what he manages to do with the materials he allows himself falls short of what seems to characterize our normative thinking. I suspect that this shortfall isn't due to Blackburn's efforts in particular, but that the "Humean" materials he allows himself aren't up to this job.

"Reasons primitivists," Blackburn says, stop too soon. Am I a "reasons primitivist"? Not exactly, for I think that the notion of a reason is composite, analyzable in terms of oughts of warrant: a reason to do a thing, in the normative sense, is a consideration that one ought to weigh in favor of doing it. I may, though, qualify as a "warrant primitivist". I am not exactly attributing to warrant what Moore attributes to good (or what Ewing attributes to his primitive ought for which I am using the term 'warrant'). I don't just stop with the concept of warrant as understood. I think that we do understand it and can say theoretical things in terms of it, and I think that we can analyze normative concepts in terms of it. Being so analyzable, I hold indeed, is what constitutes a concept's being normative. This isn't a matter of states we can identify by introspection, like impressions and their weaker copies ideas according to Hume. But I wouldn't expect that this is the aspect of Hume that guides Blackburn. I want not just to recognize the concept and appeal to this recognition, like Moore with good, but also to identify the concept of warrant through the role that believing things warranted plays is an adequate human psychology—on the hypothesis, as I keep indicating, that there is such a role.

Blackburn, of course, needn't accept this hypothesis of mine. He recognizes the phenomena that I attempt to treat with my hypothesis, but, if I'm understanding him, he thinks that a kind of treatment different from mine can best handle them, one that doesn't rest on a basic notion of warrant. What he proposes is subtle, perspicacious, and imaginative; I stand in awe. But as I'll begin to indicate, some of the things he says don't quite work, and others implicitly help themselves to the concept of warrant.

Things that in effect help themselves to a notion of warrant without further explanation: "When our desires appear to ourselves not to be appropriate to their objects there is a disruption of the pleasant harmony between our desires, on the one hand, and the sense of their defensibility or propriety, on the other, that we like to enjoy" (99). The term 'appropriate' seems synonymous with 'warranted', and defensibility can be of various kinds: a feeling could be defended as virtuous, generous, or engaging; what's relevant is defending it as warranted—which would take us back to needing to explain warrant. We are worried, he says, about the "justice of our reaction", but this, I take it, is preliminary, a rough indication of what we should be trying to identify. He speaks of some feelings' being "off" and of being in a "defective" rather than "sound" state, but not just any kind of defect or unsoundness gives us the phenomena he and I agree we find. It might count as a defect if friends will feel insulted by ways I feel, but that isn't the particular kind of defect we need to identify. Discounting

one's reaction is right, but our problem is to identify the relevant kind of discounting; it is, I say, discounting as unwarranted. Again, then, we have the explanation we need only if we can draw on the notion of warrant in giving the explanation. Some of the other things Blackburn says needn't succeed. Impartial spectators are good for pursuing thoughts of whether a feeling is warranted, but it isn't analytic that if we become impartial spectators, what we feel will be what's warranted.

"I hope," he says, "that these considerations take away the threat of there being inexplicable jumps in the journey from reactions such as desire or aversion toward judgments of good or bad, right or wrong" (107). My own hypothesis is that introducing warrant is a jump, but the jump is explicable. I agree with Blackburn that the goal should be "to delineate more accurately the kind of dissonance involved" (100), and as I say, I haven't found a way to do that without a basic notion of warrant.

"Suppose, for instance, the naturalist suggests that holding an attitude to be fitting is a matter of identifying yourself with it, and that in turn is a matter of feeling able to stand by it, or defend it, and at the very least feeling no shame or guilt about holding it" (100). Some of this strikes me as possibly moving us in the direction we need. Pride and shame won't do it themselves; these matter, but it isn't analytic that we'll feel ashamed of feelings we think unwarranted, and so considerations of pride and shame won't give us enough of what we should be seeking. But if we could identify the right kind of "standing by" a feeling, that might give us what we need. Hume as Blackburn quotes him is right, I think, about how we evaluate a judge's accuracy of judgment, but thinking that we are justified in holding judges to have such powers isn't the same as thinking that their judgments are right.

When I complain that Blackburn's ways of avoiding taking warrant as basic aren't working, Blackburn might well reply that he is doing something quite legitimate: explaining the phenomena we agree on in terms of a cluster of typical properties, none of which is tied into the phenomena as a conceptual truth. That's an approach that we might be driven to, I concede; it may be the best we can manage if there isn't the special kind of normative state that I hypothesize. My own ambitions go beyond this, but as I concede, they may not be achievable.

Blackburn, I think, is quite right to ask why practices of assessing warrant are deeply worth having, and what that has to do with normative truth. "Why do we have these practices?" he asks. We ponder "what *we ourselves* are to think about it and typically this is expected to be what *we together* are to think about it"—as Hume stressed (103). "Once we have the idea of a common point of view, or the common pursuit of true judgment, the 'right reasons' problem largely solves itself. The right reasons for a verdict are just those features of a subject matter that can be advanced in a common pursuit" (106). I don't, however, see that the problem that worries me solves itself this way. It's not literally a matter of what *can* be advanced; specious reasons can be advanced. We need reasons that help us in the pursuit, and if the aim of the pursuit is to get things right, we need a story of how pragmatic

considerations aim us that way. I greatly hope there is such a story, but I concede I don't have it. I think that things Blackburn says get at something of central importance, but not something I know how to put.

Another kind of line that Blackburn gestures toward seems to me to be very important, something I would very much like to make work or to see someone else make work. "If we insist on asking *why* this character alone entitles the critic to being someone worth listening to, then pragmatism comes to the rescue (for it would be tossing in the sponge just to say that these virtues are indicators of aesthetic truth)" (104). This is the deep question on which I wish we could find more to say. I read D'Arms and Jacobson as taking a hard anti-pragmatic line, insisting that usefulness has no bearing on what's fitting. My hope is to elucidate the way pragmatic considerations do bear on this. "The good critic can show us things we had otherwise missed, enable us to place works in their traditions, to come to understand what is more satisfying and permanently satisfying, and thereby increase our enjoyment." To the extent that these are the reasons to care whether the critic is getting things right, my puzzle is their relation to what succeeding in getting things right consists in. Our ambition should be to improve on these things that Blackburn says, but again, I don't know how to do so, and the things that Blackburn says are, as one would expect, valuable and insightful.

The crudest of pragmatisms won't be acceptable, both he and I agree; we distinguish whether a thought is useful from whether it is true, and doing so is a significant aspect of our practices. "Why," though, "do we have these practices?" Blackburn asks. I do think it is crucial how, as he says, "creatures such as ourselves, starting with an intelligible endowment of mental states, and having typical human problems to solve, might be expected to end up talking, thinking, or feeling as we do" (101). I appeal to this in my 1990 book *Wise Choices, Apt Feelings*. If there's no direct path from the usefulness of a practice to its correctness, then we face the question of what connection, if any, there is between them. I don't credit myself with succeeding in answering this—and I don't know that anyone else has succeeded either. I am convinced that some sort of measured pragmatism could be elucidated and turn out to be right, but we don't have it. As I said above in my reply to Singer and Aronowitz, there is at least one instance of a clear connection between warrant and a component of usefulness: that the credences that are warranted are the ones with maximal guidance value. I don't know if there's any corresponding pragmatic virtue for warranted feelings, any pragmatic virtue that the warranted feelings to have toward a thing systematically enjoy. I would be delighted if someone could answer this—especially if it vindicates an illuminating form of pragmatism for feelings.

I am intrigued by Blackburn's assertion that the nature of observation refutes any claim that the space of reasons is distinct from the space of causes. On my own view, this relates crucially to the contrast I harp on between properties, on the one hand, and property concepts. Properly understood, the "space of reasons" is conceptual. Consider, for instance, Galileo's training his telescope on the heavens. The property of being a reason to peer in one's

telescope is a causal property, whereas the concept of being a reason to peer in one's telescope is distinct from any purely causal concept; it combines causal and normative elements. We can disagree as to what property it is—being a reason to peer in one's telescope—without anyone's falling into conceptual contradiction. How does this apply to observation? Galileo observed that Jupiter has moons. The property of observing this is an ordinary causal property. It involves, as Blackburn says, "putting yourself in line for a causal impact from a state of affairs, and provided the causal impact is one that can be expected if, but only if, the state of affairs obtains, it is a very good way of acquiring a reason for believing in the state of affairs" (108). But note, this specification of the property isn't restricted to causal notions. When you put yourself in line for an impact that *can* be expected in certain circumstances, the "can" is a matter of warrant. I am far from a Sellars expert (though I had undertaken to improve myself in this regard preparing a keynote for a Sellars conference; I was however unable, alas, to deliver on this undertaking). But he presumably didn't mean that the "space of reasons" and the "space of causes" don't combine in many concepts—and the concept of an observation involves such a combination.

Stephen Darwall and Howard Nye

Stephen Darwall and Howard Nye treat many of the same issues as each other, and so I'll respond to them together. (Nye's commentary bears heavily too on Schroeder's, which comes later in this volume—as I'll indicate.) The analysis of the concept of moral wrongness that Darwall adopts, discusses, and draws on is, as he stresses, at least roughly the one that I myself gave in earlier writings. Nye, though, has criticized this analysis and proposed a different one. I some time ago accepted his criticisms and moved toward dropping my previous analysis and adopting pretty much his. I fully accept all he says in his commentary here.

Both Darwall and Nye maintain that if one has sufficient reason to do a thing, that defeats its being wrong. Darwall says one can't "coherently have the attitude of blame toward someone (either yourself or someone else) and simultaneously accept that they had sufficient normative reason for acting as they did". Nye says, "There seem to be genuine problems with the coherence of thinking that one's doing something would be morally wrong but that one has sufficient reason to do it anyway" (137). I am convinced by these things they say.

When I wrote *Wise Choices* (1990), I thought I was drawing my account of the notion of moral wrongness from Mill and Brandt. As Darwall puts it, "In determining what our moral duties and hence wrongful conduct consist in we are in effect determining what kinds of acts we have normative reason to blame if these acts are done without excuse" (111). I put this in terms of guilt as well as blame or resentment, but whether or not we bring in guilt, the definition has two problems. One is that it leaves us to explain the notion of an excuse, whereas if we manage to explain moral wrongness without invoking the notion of an excuse, we can say that an excuse is something that keeps one from being to blame when one does a wrong

thing. Probably, though, we can nevertheless explain being an excuse without having already explained wrongness, and so this challenge is likely tractable. A more profound problem was pointed out to me by Nye some time ago, a problem he takes up in his commentary. Blame and guilt, as he observes, are backward-looking feelings toward an act or action, whereas we should characterize wrongness in terms of forward-looking feelings, the feelings one can have in the course of deciding what to do. Subsequently in my writings, I put my analysis of what 'morally wrong' means in terms of prospective feelings of guilt-laden aversion. Nye's commentary in this volume does even better: he puts it in terms of feelings of obligation, and I think this formulation of his is right. Moreover, he finds in both Mill and Brandt indications that they favored a view like his.

Moral right and wrong, proposes Nye, concern the fittingness of feelings of obligation. To explain what it is for an act to be morally wrong, we should invoke not backward-looking guilt and resentment, but "the fittingness of forward-looking feelings of obligation to perform it, which involve motivation to perform the act" (126). Fittingness here is the notion I just adduced in my response to Blackburn. I'll speak here in terms of a "must" of fittingness: that one *must* feel obligated means that it would be unfitting not to. When I say, then, that it is wrong to razor an article out of a library journal to take it with you, I mean that one must, in this sense, feel obligated not to—at least if without such a feeling one might go ahead and razor it out. Nye ties feelings of obligation to guilt-tinged aversion, and speculates, "For evolutionary reasons, our ancestors came to tend to feel this kind of guilt-tinged aversion as an adaptive inhibition to defecting" (140). This ties in with Rosati's challenge which I take up later as to why such coordination would result from judgments of how to feel about an act, judgments of the "fittingness", rather than of the desirability from one's standpoint, of feeling that way about it. I'll grapple with this later when I discuss her commentary.

Darwall speaks of these matters in terms of holding accountable, and Nye's proposal, we can say, is a matter of holding oneself accountable—in advance. Wrongness on this characterization, though, is not itself a second-person notion, and we still face the question of how all this ties in with retrospective feelings: both with second- and third-person feelings of resentment or blame, and with first-person feelings of guilt or self-blame. Warrant for these prospective and retrospective feelings don't always go together, for as Brandt, Darwall, and the rest of us stress, one can do something wrong but have an excuse. Even so, the default seems to be that an act that in advance merits feeling obligated not to do then merits guilt and resentment if one does it nevertheless. Suppose that doing what's right in a situation would be heroic—say if the only way to save a group is to sacrifice one's own daughter. Should I feel obligated not to save her even if later on, after I save her, I ought not to feel guilty for having done it, because getting oneself to act as one morally ought to act in such a case is beyond what can be demanded? I think so, but I'm puzzled as to why. The solution might be to say that you have to be stricter with yourself in advance than in retrospect, but this calls for more investigation.

Nye says, "Judgments about wrongness and moral reasons seem to have the central normative property of guiding feelings of obligation" (132). This ties in with questions of judgment internalism that will come up when I later take up the commentary of Mark Schroeder.

All this leaves Darwall and Nye somewhat at odds in the analyses they accept. Still, I think that if Nye and I are right on moral wrongness, ample room remains for the wonderfully rich things Darwall says about other moral concepts—now modified to substitute, as their basis, Nye's concept of moral wrongness for Darwall's analysis. Once Darwall's proposals for moral concepts are so altered, I think I can embrace all of them. Darwall suggests that all moral ideas are tied, whether obviously or covertly, to blameworthiness. Could we change this to suggesting that all moral ideas tie in somehow with feeling obligated? That fits something else that Darwall says: "What makes something a moral reason for acting is that it is a *pro-tanto moral obligation-making consideration*" (115). If we switch to Nye's ways, we can continue to say such things as that being morally best means being best supported by moral reasons for action, so that an act may be morally best even if not morally obligatory.

Darwall and Nye both, then, say things I wish I had said and add substantially to the kind of account of the meanings of moral terms that I embrace.

3. Expressivism, Normative Language, and Semantics

8. Tristram McPherson (Ohio State University), "Expressivism without Minimalism"
9. Nate Charlow (University of Toronto), "Metasemantic Quandaries"
10. Alex Silk (University of Birmingham), Weak and Strong Necessity Modals: On Linguistic Means of Expressing "A Primitive Concept Ought"
11. Caleb Perl (University of Southern California), "How to Outfox Sly Pete: A Picture of the Pragmatics of Indicatives"
12. Seth Yalcin (Berkeley), "Modeling With Hyperplans"

Tristam McPherson and Caleb Perl

"There is little reason," concludes Tristam McPherson, "for the expressivist to be committed to minimalism about truth" (167). Normative language and thinking, after all, serve for planning and coordination. That's why we need it, and for this, we do need to attribute inferential significance to our normative claims. But if we move beyond this and attribute truth-aptness to them, we move needlessly, and there are drawbacks. Better would be agnosticism on whether normative claims are truth-apt. (Or this is how I understand his line of thinking.)

This fits much of what I believe, but not all. For normative talk and thought, I reject what Huw Price calls *representationalism*, with its account of truth. Representationalism for a

discourse maintains we can find a single feature that qualifies as truth, that there is a uniform relation ***represents*** between the discourse and its subject matter which figures crucially in *explaining* how the discourse works. What's true is what bears this relation to some part of reality. Whether any kind of discourse fits this pattern raises a difficult set of issues; in any case, we expressivists say, normative discourse does not. Still, once we move away from this picture of truth and truth-aptness, it may still be that significant features of our thinking are as they would be if representationalism held true. One thing we can ask, then, is whether the logic of the discourse is the same as it would be if a relation ***represents*** did explain it all. If so, this could either be because this relation genuinely does succeed in explaining what's going on, but alternatively, the explanation of the pattern of logic may be deflationary—as we expressivists think holds for normative discourse.

Calling something "true" is chiefly a matter of agreeing with it. There is a package of features that I myself count as truth, and although truth minimalism, in its most straightforward formulation, lies near the core of this, it doesn't capture what I think of as truth in its entirety; something more is needed. I reject truth or falsity in my expanded sense for some major kinds of claims: for indicative conditionals with false antecedents ("If Pete called, he won") and for epistemic modals ("The keys might be on the shelf"). Many normative claims do have the package of features that representationalism would yield, and so I count them as true even if no form of representationalism succeeds in explaining their truth. Is this truth literally? More than one package of features will lie in the vicinity of truth, and which truly counts as truth may not have a fully determinate answer. Some claims that I don't count as true are nevertheless true in a less demanding sense. All this is what I'll try to explain in response to McPherson and Perl, and also to Yalcin.

Caleb Perl rejects what I say on being acceptable but falling short of what I call full truth. At the start of his section "Expressivism Can't Work", he tells us that a claim he labels (1) "is definitely appropriate—and its appropriateness is powerful evidence that it's true." He is here speaking of a complex example he has devised: characters Esther and Welsa who, Perl asserts, know things like "if Top Gate opened, all the water ran westwards" (249). Since they know it, he concludes, it's true. In a note, Perl observes that I deny that one can precisely know such things: I deny that indicative conditionals with false antecedents have truth values, and so by factivity, the principle that unless a claim is true one doesn't know it, this indicates that such conditionals aren't things one can know. Often one can reasonably be sure of them, in light of one's evidence, but one doesn't thereby strictly know them. Perl and I both invoke factivity for knowing. Since I deny that it's *true* that if Top Gate opened, all the water ran westward, I therefore deny that anyone can *know* this. And I deny that anyone can be fully clearheaded and believe that a character can know such a thing. In this kind of case, people can properly believe and assert things they don't know. Again, my argument for this, long ago, rested on my Sly Pete example, where one observer can legitimately assert "If Pete called he won" and another can legitimately assert "If Pete called he didn't win." Each of them makes a claim that is definitely appropriate. An omniscient observer, however, would

have no view as to whether if Pete called he won or not. Pete didn't call, he knows, and so for him, the question of whether if Pete called he won can't arise.

The grounds Perl adduces for rejecting expressivism for indicative conditionals, then, I reject. Like things hold for epistemic modals. Take the stock example, "The keys might be on the shelf," or as it could be otherwise put, "Perhaps the keys are on the shelf." These can be definitely appropriate even when in fact the keys are not on the shelf, so that an omniscient observer wouldn't accept them. It can be definitely appropriate to say "Perhaps the keys are on the shelf" even when it isn't *true* that perhaps they are. (I don't know if this violates Williamson's knowledge norm of assertability, but if it does, then I'd think he should qualify the claim.)

Return, then, to truth and representationalism. Both with indicative conditionals and with epistemic modals, we have a pattern of inferential relations that couldn't possibly be explained in a nondeflationary way by a relation ***represents***. This shows that on an important understanding of truth and falsity, neither of these kinds of utterances will in general have truth values. McPherson is clearly right, these cases show, that inferential standards needn't always be rooted in imputations of truth—at least as I use the term. I have been using the term for cases where the logic of a discourse is what we would get if a relation ***represents*** did the explaining, whether or not any such explanation gets matters right. The chief virtue of expressivism, to my mind, is that it gives us a way of getting the symptoms of truth, and does so without a uniform relation of representation between a thought and its subject matter.

McPherson fears that expressivism that includes claims to truth won't be distinctive. I respond that what makes expressivism distinctive is its rejection of representationalism, along with its alternative explanation of how normative concepts work. If we allowed truth-aptness just to thoughts which representationalism genuinely explains, we might get very little truth at all. That might please current Nietzscheans, who stress Nietzsche's attacks on the notion of truth. But for straight normative discourse, the package of symptoms is the same as if a relation ***represents*** did do the explaining, and so that's what I have been labeling as truth.

As I have indicated, though, with indicative conditionals and epistemic modals, lighter packages might have virtues too as candidates for truth-aptness: they may allow for disagreement of a kind and for certain kinds of generalizations. A partially informed observer who agrees that if Pete called he won may be able to say legitimately, in a truth-light sense, "Zack said something true." And it is clear that for some such truth-light discourses, we can satisfactorily have some inferential standards without what I am counting as full truth. Whether it is useful to label truth-light packages in such a way may be worth exploring.

What more is needed, then, for what we should count as full truth or falsity. It's something like this: That there's a maximally specific way things are, and truth is a matter of fitting that way. Equivalently, a fully opinionated observer would agree or disagree. This applies to normative claims: a thinker who is fully decided on what to do in every circumstance a person might face will agree or disagree with normative claims concerning what people ought to do.

This has two important features: First, normative truth is a substantive matter of what to do, of how to live; it goes beyond the facts as conceived nonnormatively. Ought one to weigh benefits to others intrinsically in one's decisions? Both normative egoism and normative universalism fit all formal requirements on normative thinking, but what to do if one's own interests conflict with those of others is a question on which we can disagree. The question is genuine. Second, the requirement to fit in with a maximally specific way for things to be determines a logic, a logic that mere near-truth may not have. There's no maximally specific way things might be, naturalistically specified, such that if things are that way, then, if Pete called he won. There's no maximally specific way things might be, naturalistically specified, such that if things are that way, then the keys might be on the shelf. (Sometimes even when the keys aren't in fact on the shelf, it can be legitimately assertible that they might be.) On the other hand, normative claims can meet this requirement. The sheer logic of normative claims allows a maximally specific way to live such that if things are that way normatively, then one must take the interests of others into account in one's decisions. This logic allows too a maximally specific way to live such that if that's the way to live, one may legitimately give no intrinsic weight to the interests of others in deciding what to do. Which way obtains in truth is a matter of what to do when interests conflict and one must act. This is a normative issue, not an issue of how things are as specified naturalistically. It is an issue of what's true, in a sense that fits the entire logic of truth.

This allows for a near-minimalism for truth: deflation plus this second requirement, that a claim fit the maximally specific way things are. These requirements determine the inferential patterns that characterize truth as I call it. All this is applied to substantive questions of what to do in circumstances one might face; normative truth is a matter of what to do.[7] This is a question on which we can disagree even if we are each ideally coherent and each know all the naturalistically couched truths. I thus accept what we might call "near minimalism" for truth as fitting the determinate way things are. If McPherson rejects such a near-minimalism for normative truth, I find in what he says no adequate reason to think that a normative expressivist must join him. What an expressivist rejects is representationalism for normative claims. I thus concede a great deal to McPherson, but I still insist that the full package that I myself am calling "truth" is a package that matters and that can apply to normative claims. (What I'm saying here must be qualified in whatever way copes with semantic paradoxes and thus allows us to formulate truth minimalism or near-minimalism in a way that isn't self-refuting. There is an extensive literature on how to do this, but I take it no one knows how to do this in a way that is entirely satisfactory. These problems aren't specific to minimalism or near-minimalism

7 Or less precisely, truth as to what one ought to do depends on what is to be done in one's circumstances. The title of Lenin's tract *Shto Delat?* gets translated as *What Is to Be Done?* but more accurately put, it means *What to Do?* A friend of mine who grew up in East Germany told me that they translated it as *Was Tun?*

but to any account of truth: seeming truisms about truth must somehow be restricted, and it's not clear we know how.)

Let me add a remark on "facts". Writers who go for truth minimalism may extend this to facts. For a time I did this, but I now think that it was a mistake. Calling a clear normative truth a fact, as Cornell realists do, can come across often as rhetorical bludgeoning: is it a fact whether utilitarianism gets matters right? The term 'fact' has its home in contrasts: prominently, between matters of fact and matters of law, between what belongs on the news pages and what belongs on the opinion page. (Originally, I seem to remember learning, the term 'factus' meant DEED, the deed on which the claim to a throne in the Holy Roman Empire was based.) So I am now happy to use the term 'factual' to exclude what's purely normative, purely a matter of what to do.

McPherson has more to say that I won't much grapple with (167). Truth minimalism, he says, if included among the theses of normative expressivists,

> saddles their view with additional controversial commitments; it may threaten the distinctiveness of expressivism as a metaethical view, or undercut its apparently distinctive dialectical virtues; And if these worries can be fended off, the reconstructed distinctions will likely show that much of the thought and talk that the marriage is supposed to accommodate reflects commitments inconsistent with expressivism.

He alludes to an alternative set of views that he holds which I won't try to address, but I hope that the things I have been saying address the complaints I have quoted. Likewise with Perl: he says other interesting things, but I think I have addressed the grounds he puts forth for rejecting expressivism for indicative conditionals.

Seth Yalcin

I have been speaking of maximally opinionated states, and I explain them as involving both normative matters—what to do on various kinds of occasions—and factual, naturalistic matters. Yalcin considers "maximally opinionated states of (factual) belief" and says that on my view, they "are not, as Lewis or Stalnaker would have it, maximally specific ways things might have been, understanding the relevant modality as fundamentally non-mental" (286). I heartily approve of Stalnaker's talk of maximally specific way things might have been, and would extend this to bringing in maximally specific ways things may conceivably be. There's a correspondence between the possible belief that things are some way and that way itself, and so at their maximally specific, a maximally opinionated state of mind corresponds to an epistemically possible world as a maximally specific way things may be. I go back and forth freely between talk of ruling out a way things may be and ruling out belief that things are that way. To rule out grass being green is closely tied to ruling out believing that grass is green. These two ways of thinking amount to the same thing; each is more tractable for certain purposes. These are alternative ways of formulating

things, not alternatives that exclude each other. What I'm saying here is sloppy and needs far more analysis, but this correspondence lets us go back and forth between the one way of thinking and the other.[8]

Robert Stalnaker has an important recent paper on my treatments of these things, a paper that it would have been wonderful to include in this volume, if it hadn't been going elsewhere. He objects to starting with states of mind, because doing so doesn't let us handle sentences like "Although I don't believe that grass is green, it is." For such a sentence, a contents-first approach is most natural, whereas starting with states of mind requires us to adduce possible states of mind that aren't transparent to the thinker—a state of mind of believing that grass is green but believing that one doesn't so believe. But this doesn't utterly rule out beginning with states of mind, and doing so has its advantages too. As I say, it lets us handle epistemic modals and indicative conditionals with false antecedents. Yalcin has the definitive paper on epistemic modals like "The keys might be on the shelf." This expresses giving some nonnegligible credence to the keys' being on the shelf. And as I indicated above in my response to Perl, decades ago I wrote a paper on subjunctive and indicative conditionals that argued for Ernest Adams's treatment of indicative conditionals; the sentence "If the keys are on the shelf, they'll get lost" expresses high conditional credence in the keys' getting lost given their being on the shelf. Stalnaker proposes contents first ways of trying to handle such constructions, but even if they work, these states-of-mind first ways of treating them are straightforward.

Yalcin speaks of "normative realism", and says of my view, "Out is the idea of characterizing propositional attitudes as relations to contents." He speaks of "normative realism" and worries that I'm not "characterizing propositional attitudes as relations to contents" (286). This might seem to indicate that I reject representationalism across the board. In an article addressing Price's "universal expressivism", however, I argue that representationalism works for such matters as the layout of surrounding mid-sized objects, but not for good and bad or right and wrong. (I confess to doubts about this, and about which side of the line other subject matters lie on. The line between what's well explained in terms of a general relation REPRESENTS and what's representational only in a sense that is deflationary would raise an intriguing sets of issues I don't try to settle. I realize that the case of mid-sized surrounding objects needs more analysis than I give it, and might not support what I say.) Even where representationalism works, though, one can also explain logical relations in terms of hyperopinionated states. In the normative case in particular, no genuine relation between

8 I skip over, in what I have been saying, the contrast between what's metaphysical, ways things might have been, and what's epistemic, ways things may be. My systematic attempt to work some of this out is in the appendix "The Objects of Beliefs" in my book *Meaning and Normativity*. I'm obviously not disinterested, but I continue to think that the framework I delineate there would clarify many treatments in the literature of epistemic and metaphysical possibility and how they interact.

thinking and the states of affairs one thinks about will serve as a basis for explaining what's going on.

I may well be misreading in casting Yalcin as a representationalist for normativity. Although the things I have quoted may suggest representationalism, one thing he says in explaining his Plan B+ doesn't seem to fit this: Plan B+ explains normative judgment, he says, "as not in the business of representing normative facts" (288). I'd love to get clearer on how all this fits together.

As Yalcin notes, I use terms like 'plan' in a special way that might all too easily be misleading. This has to do with ties in the outcomes of deliberation. Writes Yalcin, "There's my state of normative judgment, which pronounces on what's permissible, and there's my decisional state, which seems more directly related to what exactly I end up doing" (282). I do distinguish these, and "planning", in my book, concerns the former. Writers sometimes distinguish "choosing" from "picking", where one picks among alternatives when one regards more than one as fully eligible. Planning, as I settle on using the term, is confined to choosing in this sense, not picking—choosing for a future or hypothetical case. Perhaps, then, I should avoid the term 'plan', and stick to talk of what, in thinking what to do, one permits oneself, requires of oneself, or forbids oneself. I should adopt the slogan, "Concluding what it's permissible to do is concluding what to permit oneself to do," and instead of the slogan "Thinking what one ought to do is thinking what to do," I should retain another slogan that Yalcin quotes: "To believe that a person ought to do a thing is to require it of oneself for the hypothetical case of forthwith being in that person's precise situation" (284).

My talk of "planning" in *Thinking How to Live* (2003) was spurred by Michael Bratman's wonderful treatment, though I didn't follow him exactly. Speaking the way I did served important purposes. It helps us to keep clear track of the logic of normative thinking, since with full coherence, one plans to do what one is convinced one must do. In later writings, though, I gave up this talk, since it demanded repeated disclaimers of features in ordinary talk that my special sense lacks.

I am using 'ought to' in a strong sense equivalent to 'must'; I turn to this shortly in my response to Alex Silk, but in the meantime, I'll use the terms interchangeably. Says Yalcin, "'Thinking what to do' can just mean 'thinking what ought to be done'—in which case the former marks no special progress" (281). In light of the slogans I now embrace, then, I should explain what kind of progress I think we make. I progress from permitting myself an act to believing the act permissible, and from requiring an act of myself to believing that it's what I must do. Permitting oneself we can explain as a stage in moving toward action, and requiring an act of oneself is one way a person can come to act. In the general case, once one has resolved all uncertainties, one permits oneself one or more alternatives and forbids oneself all others. In case there is exactly one alternative one thus permits oneself, one does it. If one permits oneself more than one alternative, one picks among the alternatives one permits oneself and performs the act one picks. One can, to be sure, act before working all this out, acting in a way one permits oneself without settling what else to permit oneself. But

one may also work out fully what to do and why, acting in a way that one permits oneself to act. We specify the mental acts of permitting, forbidding, and requiring oneself by their role in this process, and explain the concepts MUST, MAY, and MUST NOT expressivistically: we explain MUST by saying that to believe that one must do a thing is to require it of oneself, we explain PERMISSIBLE by saying that to believe an alternative permissible is to permit it to oneself, and we explain MUST NOT by saying that to believe that one must not do an act is to forbid it to oneself.

Yalcin's worry, translated into this framework, is perhaps this: to "permit" oneself an act, I maintain, just *is* to believe it permissible. I thus explain being permissible via believing permissible, explaining the latter, in an empty way, via that state itself. This worry is much like one that dogged Ayer's emotivism: Ayer explained the concept good via approving or favoring: to believe a thing good is to approve or favor it—but, the worry goes, approving is just believing good, and so the explanation is empty. For Ayer's explanation of the concept GOOD to be genuinely explanatory, we must characterize approving or favoring independently as a psychic state. That burden, however, seems easy enough for Ayer to shoulder. Favoring is explained by a tendency to promote what one favors. Likewise with permitting, requiring, and forbidding oneself an act: each of these states is explained by its role is coming to act purposely.

Requiring an act of oneself is thus explained in two ways: in the expressivistic way I just went through, and, for a person who has mastered a word like 'must', as believing that one must perform an act. The concept can be taught in either of these ways. The first explains normative concepts in terms that are psychological, in terms of a psychic state that gets one part way to decision, though it may leave a further stage of picking among alternatives one permits oneself. The second exploits the conceptual competence of a person who already uses terms like 'must'. My thesis is that these explain the same thing. (Yalcin discusses the phrase "the thing to do" which I got from Railton, and like points apply to this phrase. The phrase is appropriate only when alternatives aren't tied for being the thing to do. Absent any such tie, we can bring a person to grasp the concept of being the thing to do either by teaching a word like 'must', or by saying that to believe that an alternative is the thing to do is to settle on doing it.)

With this on the table, I turn to two questions Yalcin poses. "There is a gap between this ideally-conceived planning and ideally-conceived normative judgment. Does Gibbard grant this point?" (284). I don't grant it if "planning" is conceived as ending in permitting or forbidding oneself various alternatives. I do, of course, if "planning" is deciding apart from settling what to permit oneself and what to forbid oneself. In that case, "planning" includes something that isn't part of normative judgment.

Yalcin asks also, "Is the normative state modeled with N the same as the planning state modeled with P?" (285). That is, is *believing* that one *must* pack the same as *planning*, in my special sense, to pack? Now that I've eliminated the misleading term 'plan' in this response, the question becomes whether believing that one must pack is the same as

requiring oneself to pack—and I claim that it is. Where I here answer Yes, Yalcin perhaps seems to say No. "Normative judgement—or more specifically, one's views about what is permissible to do in various situations actual, hypothetical, and counterfactual—is its own thing" (293). I'm not at all sure whether or not a represenationalist view that I reject lies behind his words.

I accept Yalcin's proposal, "We might just as well view hyperplans as functions on subjective predicaments", that we "reconstrue hyperplans as functions on sets of centered worlds, rather than on centered worlds". I agree that for purposes of explaining Holmes's judgment that he ought forthwith to pack, "You might as well have two separate sets, a set of centered worlds (factual belief) and a set of hyperplans" (297). Whether an observer such as Watson believes that Holmes ought forthwith to pack, however, can't be simplified in such a way; it depends on which pairs of sets of centered worlds and hyperplans Watson allows.

This brings me to important things that Yalcin says that I haven't managed to understand. He considers a case where Holmes learns that it's too late to catch the train, and so concludes that he needn't pack. In this case Yalcin says, "Holmes goes from thinking he ought to do something to thinking he needn't" (298). We need to be clear on the times of these oughts. As I read them, they are oughts in the subjective sense, in light of the information Holmes has at the time for action and in light of the credences for him then to have. We are considering Holmes's rational views at time t and at a later time t+1. At time t, the story is, Holmes hasn't gotten the news that it will be too late to catch the train; at time t+1 he has the news. What is he to think in advance, at time t, that it will be the case that he ought to do at time t+1? That, I would think, depends on what at time t he justifiably expects he will know at time t+1. There seem to be two possibilities, and I don't know which Yalcin has in mind. First, perhaps at time t, Holmes expects that at time t+1 he won't have such news. He thus expects that at time t+1 it will be the case that he ought to pack. He then gets the surprising news that it is too late to pack, and so changes his mind. Or perhaps at time t Holmes is uncertain whether he will get more news by time t+1. In that case, Holmes won't have a definite view as to what it will be the case that he ought at time t+1 to do. He'll think that if he'll have such news, it won't be the case that he ought to pack, and if he won't have such news, then it will be the case that he ought to pack.

Whichever of these cases he has in mind, Yalcin speaks of "a case of strict gain in information" (298), and he says, "What changed is his view about the world, and hence which aspect of his (stable, unchanging) normative view speaks to the situation he takes himself to be in" (298). That's right, I think, for either of these variants: his purely normative views are given by his hyperplans, which don't change; what changes is the prosaically factual information he has at time t+1, as opposed to what he earlier expected he would have.

What I don't get is the moral Yalcin draws: that a set of options can be fixed by a state of information which is not itself a possible doxastic state. I'm not clear either why he says, "not all thinking what to do can theorized in terms of hyperdecided states being ruled in

or out" (300). At time *t*+1, if he has gotten the news, he rules out all hyperdecided states that include believing he has time to catch the train. If he hasn't gotten the news, he doesn't blanketly rule out all such states. These things that Yalcin says may well be true, but if they are, I'm not following why.

Yalcin's commentary is profound and helpful. Some of what he proposes I very much accept, and his criticisms spur me to try to fix some things up. At the same time, as I have indicated, some of what he says I don't yet understand. I know he's a philosopher who tends to get things marvelously and insightfully right, and so I expect that he's right in these claims. But I have tried to explain points which so far I'm not following and might disagree with.

Alex Silk

I've always been puzzled by 'ought' and other "weak necessity modals", and I'm still not at all confident that I understand how they work. I'll stick to oughts of action in what I say and ignore epistemic oughts and oughts of what to feel. Silk has complex and illuminating things to say about the terms 'ought' and 'must', but I'll start with a simplistic version of what he says, and only later get to a crucial qualification he offers. His position oversimplified is this: 'must' is a "strong" necessity modal and 'ought' is a "weak" one. 'Must' works pretty much as I depict for 'ought'; it is a "strong" necessity modal in the sense that it doesn't just present a consideration as having weight, but prescribes what to do. The word 'ought', though, works differently from the way I say it does. It is a "weak" necessity modal in the sense of indicating defeasible considerations in favor of what it is said one "ought" to do. In short, then, for action, weak necessity is a matter of something's being a consideration, whereas strong necessity is a matter of what to do, all considerations taken into account.[9]

If Silk's full treatment were this, it wouldn't fit the linguistic phenomena. It would do pretty well for 'must', but not always for 'ought'. Consider Sartre's famous dilemma: I'm a young man in occupied France whose mother needs his care, but who feels a strong obligation to join the resistance—to "join up", as I'll put it. I can't do both, and I ask you, "What do you think I ought to do?" You might respond firmly, "You ought to join up!" (In this discussion, I'll use the exclamation point for things said in a firm tone of voice with falling intonation at the end.) To my linguistic sensibility, that settles the issue against caring for Mom at the expense of joining up. I can reply, "So in light of all that, you don't think I ought to shirk joining up in order to care for Mom."

As the oversimplified version of Silk I just gave adjudicates these matters, you might instead reply, "You ought to care for Mom. Then again, you ought to join up. It's a dilemma." I can hear that as a proper reply, but the intonation will be different. "You ought to care for Mom" must end with a slightly rising tone, as before a comma, indicating that the train of thought will continue. We will hear this as declining to answer my question of what really

9 Hare suggests that 'must' is a better word than 'ought' for the strong necessity modals he treats, including the moral ought.

I ought to do. Suppose instead you answer, "You ought to care for Mom! Then again, you ought to join up!"—again with the exclamation point indicating a tone of finality. Then to my ear, you are then being deliberately paradoxical. You are giving two conflicting answers to my implied question of what I ought to do. Or suppose I put my question to you this way: "What in the end do you think I ought to do?" You can't unparadoxically reply, "In the end, you ought to care for Mom. And, in the end, you ought to join up."

My uncertain conclusion, then, is that 'ought to' is ambiguous between a "strong" and a "weak" necessity reading, and we have various conversational devices to disambiguate. As Silk indicates, many philosophers assume the strong necessity reading—and they aren't wrong to do so, I say. The strong necessity sense is often what people mean by the word. This fits what Silk quotes McNamara as saying about 'ought': "Its typical uses are to offer guidance, a word to the wise ("counsel of wisdom"), to recommend, advise or prescribe a course of action" (227). These qualify as strict necessity uses in the sense in which Silk and I are using.

Once we heed the qualifications Silk offers, we can perhaps see that he agrees. As I myself might put it, the term 'must' is always emphatic. Where an unemphatic strong necessity modal is needed, we use the word 'ought'. Silk quotes Campbell on the emphatic nature of 'must': "The idea that one *must* (not) do something . . . does not generally indicate that one option is simply better than another option, but that the other is 'out of bounds,' 'off the table,' 'closed off'; alternative possibilities become '*unthinkable*'" (225). When instead the considerations balance roughly but one set outweighs the other, 'must' doesn't fit. We still need a word to convey an unemphatic strong necessity, and the word we use is 'ought'. I take this to be Silk's more refined view, and it agrees with what I have been saying. 'Ought' is required rather than 'must' when we need a strong necessity modal but one that is unemphatic. It is this use, Silk might recognize, that philosophers like me pick up when we use 'ought' for strong necessity. We then use the term 'ought' regardless of magnitude, both for cases where one course of action wins by a hair's breadth and for cases where one greatly *must* do a thing. This fits the "primitive ought" which I adopt from Ewing.

I read Silk as mostly dealing with another meaning the word often has, a pro tanto sense, as Hare and others put it. In the pro tanto sense, "I ought to go now" means there is a significant consideration that favors going now. I think my discussion of various Sartre cases shows that this is not always what 'ought' means. Some of the other things Silk says about 'ought' may, however, not fit this reading, but I'll set them aside as demanding more thought than I have managed to give them. "The apparent 'weakness' of weak necessity modals derives from their bracketing whether the necessity of the prejacent is verified in the actual world" (205). And "We may want to communicate information about a body of norms without necessarily registering commitment to them or enjoining others to share in such commitment" (238). These things aside, I conclude that the word 'ought', is ambiguous between an unemphatic strong necessity modal and a weak necessity modal. The technical usage that I adopt along with Ewing and others takes in emphatic uses as well.

Nate Charlow

Charlow's is a very rich paper, and whereas I welcome much of what he says, there are many aspects I'm not equipped to assess. I'll try to respond to some of his criticisms of me on what normative assertions mean, although I'm not at all confident that I'm appreciating their full force. When we seem to disagree, some of it may just be matters of different ways of using terms—especially the term 'plan'. On the loosely Gricean account I take up, he and I do perhaps disagree genuinely.

Especially in my 2003 book *Thinking How to Live*, I used the term 'plan' in a sense much broader than is usual. In later writing I mostly drew back from this, because the term remained misleading even though I had specified ways in which my sense was misleadingly broad. As I intended the term, one can "plan" not only for what to do but also for what to believe in light of various bodies of evidence one might come to have. One can "plan" too for how to feel about things people do. The point is not that all the things we "plan" for are voluntary acts. Rather, to take an instance, "planning" how to feel affects how one does feel without fully determining it. When Charlow says that practical claims are just one kind of cognitive prescription among others and urges us to recognize kinds of prescriptive content distinct from planning content, what he calls cognitive prescriptions may be what I called "plans". He speaks of "cognitive advice", and this term of his, I take it, would cover advice on such matters as what to believe and how to feel about things people do. Advice on such matters, as he says, won't have "imperatival prescriptive force" bearing on the advisee's action-guiding states. But as I use the term, "planning" advice extends to what he calls "more broadly cognitive prescriptions" (190).

I adopt a broadly Gricean account of typical reasons to believe a speaker, reasons that implicitly guide us when we believe her. Whatever would undermine explicit Gricean reasoning will intuitively derail believing her. We won't believe her if we think her insincere, that she doesn't believe what she says, and we won't believe her if we think she has no way of knowing what she's asserting. We may of course still believe what she asserts on independent grounds, but that isn't to believe *her*. Grice-like guidance won't often be reasoned through explicitly, but when I call it implicit, I mean that hearers adjust their beliefs as if they were so guided explicitly. I maintain that an implicit Gricean rationale applies to normative claims as well as to claims of matters of naturalistic fact.

Charlow seems to deny such an implicit rationale. He objects that, in the kind of case I describe, the speaker could get the hearer to believe what she says by reporting her state of mind rather than expressing it. That's right; expressing one's state of mind can do the job, but it doesn't follow that nothing else can. (Indeed we often self-attribute and express a state of mind more or less interchangeably. Instead of telling the election skeptic "Trump lost," she can say, "I'm convinced Trump lost." These two have different meanings, but either can bring him around to sharing a belief of hers—do so less awkwardly, perhaps, because asserting a state of mind puts conversational pressure on the hearer to share it, pressure that can be uncomfortable.) Meanings, I agree, aren't fully tied down by the role of an assertion in

bringing hearers to share the speaker's state of mind and so add to the common ground; they are matters of what constitutes agreeing or disagreeing.

Communication, he says, is not "a *side-effect* of forming a specific belief about the internal state of the message's source". I don't entirely follow Charlow's alternative. "The subject simply *apprehends an instruction*, or way to plan, and decides—possibly, but not necessarily, after engaging in higher-level reasoning about the source of the instruction—whether or not to adjust their plans accordingly" (199). But why does apprehending the instruction, when he believes her, typically lead him to take the proffered advice? This may be a point on which Charlow and I do disagree.

Pain, as Charlow says, delivers a sort of message (prescriptions, I take it, for what injuries to nurse and what threats to avoid). We don't need to psych out the source to heed the message, and so why should we psych out the source when someone tells us something? (Or I think that's Charlow's challenge.) Pain delivers a message in the sense that its functional role is to get us to nurse and avoid; we are genetically adapted to feel pain when we do and respond to it on this pattern. If we are genetically adapted to believe speakers, it is, I have been saying, implicitly on a loosely Gricean pattern. I'm not clear on Charlow's alternative.

"And so one may begin to wonder what explanatory role the assignment of *non-propositional semantic content* is supposed to fill in Gibbard's semantic and pragmatic theory for normative language" (196). I see it as explaining the tie of what one believes one must do, believe, or feel to what one does end up doing, believing, or feeling. The tie is far from totally reliable, but there's a kind of crisis when it is broken. What Huw Price calls representationalism, supposing that a represented state genuinely figures in explaining the state of mind—as we picture what it is for the cat to be on the mat as explaining the nature of the belief that the cat is on the mat—doesn't explain a state of normative belief, I maintain. A different kind of explanation is needed.

So I thank Nate Charlow for a fine, thought-provoking commentary, but am convinced by only some aspects of what I understand him as saying.

4. Disagreement and Objectivity

13. Mark Schroeder (University of Southern California), "Convergence in Plan"
14. Lauren Olin (University of Missouri—St. Louis), "Comic Disagreement"
15. Peter Railton (University of Michigan), "Expressivism and Objectivity"
16. Billy Dunaway (University of Missouri, St. Louis), "The Metaphysical Conception of Realism"

Mark Schroeder

A central focus of Mark Schroeder's commentary is judgment internalism for moral wrongness, the view that "it is literally impossible to think that stealing is wrong without being in a norm-acceptance state that would motivate you not to steal" (312). Nye's analysis of wrong

bears heavily on this. It doesn't entail precisely judgment internalism for moral wrongness; for all it tells us, it isn't utterly impossible to believe an act wrong and be unmotivated not to refrain. Rather, if you believe the act wrong, that commits you to believing you *must* feel obligated not to perform it. (I'll use the word 'must' to mean it would be unfitting not to.) Feeling obligated to refrain is a motivational state, but you might believe you must feel obligated to refrain and yet not feel that way—and so this perhaps is not a motivational state.

We may well, though, get a kind of judgment internalism for the *must* of being unfitting not to feel obligated. Perhaps believing you must feel motivated involves some tendency to feel that way. Nye says, "Judgments about wrongness and moral reasons seem to have the central normative property of guiding feelings of obligation" (132). The concept of wrongness, then, somehow alludes to this guiding role. If you missed how the way you must feel bears on how to feel, you wouldn't understand this must. You can't sensibly say: "Yes, I true enough must feel obligated not to razor this article out of its library journal, but so what? That's not how I do feel." Likewise, if I contemplate razoring an article from a library copy, I can't genuinely believe it would be wrong to do so but ask "So what?" It seems incoherent not see the wrongness as bearing on whether to razor it; if one didn't see this, it seems to me, one wouldn't have the concept of moral wrongness that the rest of us have.

That's the question of judgment internalism as applied to the "must" of being required by fittingness. We may have, then, a kind of judgment internalism for wrong at second iteration: perhaps one doesn't genuinely believe the razoring wrong unless one is guided toward feeling obligated not to do it.

One thing I have tried saying is this: When you don't feel obligated but you are convinced that you must, you are in a contradictory state of mind. It's like thinking that everyone will die some day but that you won't.

Now to my term 'planning': As Schroeder and other commentators note, I have used this term in a special sense that may be misleading. I use it for deciding what to do if in a situation, current, anticipated, or hypothetical. Or more precisely, since one can regard more than one alternative as eligible, I should characterize "planning" this way: it is coming to a view on what to permit oneself and what to forbid oneself if in a situation. In case the answer is to permit oneself only a single alternative, forbidding oneself all others, one *requires* that alternative of oneself. This includes the situation of being exactly as one is, in all one's circumstances and with all one's characteristics. To believe you must get up early this morning is to require it of oneself for the case of being you—for the case of being exactly like you this morning. You might be settled on staying in bed, but for the case of being in your situation, determined to stay in bed as you are, I can require myself to get up early. Does this entail judgment internalism for 'must'—namely that people who don't get up don't genuinely believe that they must, at least at this very instant? I have sometimes taken this view: I have thought that if you don't get up, you don't really believe that you right now must. On this picture, what you might keep thinking is "I must get up, but not quite yet." On a widely held alternative view, you can be fully convinced that you must get up but firmly decide not to.

Now I agree that there might be senses of the word 'must' that act this way. But I am specially interested in a sense of this term—Ewing's primitive ought—that doesn't act this way, for which judgment internalism does obtain. This is the element, I say, that renders concepts normative. It applies to acts, and also to feelings about things. The moral must is different, but say I along with Nye, it contains this conceptual element: that I *morally must* tell the truth means that, in this primitive sense of 'must', I must feel obligated to tell the truth.

Scanlon rejects this judgment internalism for a 'must' of action. He pictures a person convinced he must call his doctor to get news of possible cancer, but fearful of a dread diagnosis and thus loath to call. The person keeps thinking "I must call" but doesn't. Now I think the likely phenomenon is the one I have indicated: thinking "I must call, but not quite yet." But exactly how to describe my thoughts as I struggle to get myself to call may not be determinate. I do insist that in a significant sense of 'must', there's no such determinate state as genuinely believing one must and firmly intending not to. Under the pressure of the occasion one can shrink back, but that's changing one's mind, whether explicitly or not.

Another view is to concede that it is possible to believe one must right now call and not to, but it is inconsistent. In recent times I have retreated to this view, but I'm not sure whether we need to. I do believe that if you don't see what you must do as bearing on what to do, you are missing the point and not using the word as the rest of us do. You may hear others use it and think it stands for a property that they care strongly about, without any full idea of what that property is. Maybe this gives meaning to the word as you use it, as a speaker of a language we share. But in that case, what you mean by the word in your thoughts is parasitic on its meaning in the rest of our speech community, and you are missing the point. Can what you must do be one consideration among others in settling what to do, a consideration that may be outweighed by other considerations? Or do some considerations that bear on what to do bear on what you must do and others not? I don't know how to fill out such a view as coherent and plausible; I'm not clear what this defeasible 'must' is supposed to mean. So I'll await an explanation, and in the meantime, use the word 'must' as settling what to do.

While on the subject of judgment internalism, I should refer readers to a treatment of these and other issues that is far superior to anything I myself can offer—the wonderful 2018 Ph. D. dissertation of Doug Kremm, *Practical Cognitivism: An Essay on Normative Judgment*.[10]

Now to Schroeder's central subject: he is quite right to stress that on the view I develop in my 2012 book *Meaning and Normativity*, my account of meaning judgments as normative applies to itself. He then asks why, if I'm right on this, we get a high degree of convergence on what words mean and a much lower degree on moral matters. Not that convergence on meanings is complete: Philosophers disagree, for example, on what we mean when we say that an act is open to one, that one *can* perform it—that, say, when one razors an article from

10 Posted at http://nrs.harvard.edu/urn-3:HUL.InstRepos:39947186. I much hope it can be published soon; that will be a major contribution to our understanding. On judgment internalism, see especially Chap. 4, "Agency and Akrasia", pp. 203–127.

a library volume, one could have done otherwise. I agree with Schroeder, however, that our tendency toward convergence on meanings is strong, and he is right to want to explain this.

He speaks of some arguments as "High Octane" and others as low or medium, and for our high degree of convergence on questions of meaning, he looks to an explanation that will be "medium octane", tying meaning to a goal of communication. That's the kind of explanation I myself would seek, but I'm puzzled how it should go. Whether communication is achieved depends on meanings: I tell you "I robbed a bank yesterday," suppose, and you say to yourself in the language you think in, "Gibbard robbed a bank." If these words of yours have their standard meanings, we have achieved communication. If, however, those words mean, in the language in which you think, that I withdrew money legitimately, then I haven't succeeded in communicating. So can we explain meanings in terms of what's needed for communication, or must we explain communication starting with meanings? When I try to answer, I find myself stymied, and so I beg for help.

Schroeder speaks of how cognitivists on meaning will treat these matters. I suspect that there are many positions that would count as cognitivist, and we should say different things about them. With Nye's analysis of wrong on the table, we can discuss two kinds of views: A theorist who accepts Nye's analysis will think that the question of "cognitivism" pertains to the *must* of fittingness: is it cognitive? A theorist who hasn't taken up Nye's analysis may ask directly about the status of wrongness. The question of how all this applies to meaning fits in best with cognitivism with Nye's analysis, but I'll start with cognitivists who don't take Nye on board.

Cognitivists are presumably representationalists: They think that there's a property that we're talking about with the word 'wrong'. They might be analytical naturalists of the kind that Moore attacked, or they might be nonnaturalists. I'll take it that Moore succeeded in refuting analytical naturalism—or more precisely, since Moore was talking of good, we should consider Ross, who thought that wrong was its own thing and argued against analytical naturalism for wrong. I'll assume that Ross's arguments work for wrong, or that other arguments along these lines can be made to work. As for nonnaturalism for wrong, Ross holds that there's a nonnatural property that 'wrong' picks out, and we can directly conceive of it. We are able to think of this property, and decent people think of it with aversion.

Our interest will presumably be in cognitivists who are fully coherent in their thinking, and we expressivists may be able to argue that there's no fully coherent way to fill out the cognitivist views we are considering. I won't say much to establish this for the various forms of cognitivism that I mention. The debates are familiar, and I don't know anything I can quickly say that would add to them. Ross's view, I say, just doesn't explain enough. How is it possible to conceive of such a nonnatural property?

The most natural kind of cognitivist to be for 'wrong' is perhaps not someone with a full theory worked out, but one who picks up the term, hearing it used in certain tones of voice, and makes it his own, acquiring various beliefs couched with it. He assumes that people are talking about something and thinks in terms of it, without having much of a view of what

it is. Such cognitivists might be moralists or amoralists: the moralist sides with morality as he conceives of it, holding it important to shun whatever's wrong, whereas the amoralist is indifferent to morality as he conceives it. People are concerned to avoid what's wrong, he observes, but he doesn't join them in this. Then too, a cognitivist might also be undecided between moralism and amoralism. In any case, the cognitivist I have in mind here takes himself to be able to think in moral terms, whether or not wrongness appalls him. But he doesn't get it; he doesn't comprehend how the rest of us are thinking.

I'm puzzled by a tentative conclusion that Schroeder comes to. (He puts it in terms of 'ought' and it bears most fully on what I have been saying if this stands for what I have been calling the must of fittingness [317].) Says Schroeder,

> If, even while endorsing the expressivist plan for what to mean, we allow that the cognitivist plan is another reasonable plan, then we can allow that although every agent who understands the meaning of 'ought' is motivated in accordance with their 'ought' judgments, among these are agents who fail to understand what 'ought' means only because they endorse a different plan for what to mean with it. This makes much more intelligible how these agents could fail to be motivated by their moral judgments, making the resulting form of judgment internalism more palatable.

My puzzle is this: If each has a plan and keeps to it, do they each mean the same thing by 'ought'? Does the cognitivist Coggy mean the same thing by this as does the expressivist Expy? If not, we might fear, they don't engage each other with their respective words 'ought'. If one asserts "No one ought to steal" and the other denies what he himself means by "No one ought to steal," isn't any seeming disagreement merely verbal? It could perfectly well be that judgment internalism is correct for what Expy means by 'must' and incorrect for what Coggy means by 'must'. If, on the other hand, the two do mean the same thing by 'must' and they know they do, then they have competing accounts of the thing they both mean, and it must be that one of those accounts is mistaken. These issues are hard to analyze, and I may not have them straight; David Plunkett and Tim Sundell question the kind of argument I have just given. Lack of engagement does, though, seem to me to be a problem, and so I don't follow what situation Schroeder is asking us to imagine.

Amidst his wonderful presentation of my ways of thinking on meaning, I find one other thing to balk at. I don't think "meaning is something that we bring to the world, rather than something that we find there" (311). What we bring to the world is *beliefs* as to what things mean, which we form on the basis of tendencies of usage and the like. It is in part because of what we are like that we come to the normative conclusions we do from premises about what speakers are like, but those conclusions aren't about us; they are about the speaker. If anyone brings meaning to the world, it is speakers and those who comprehend. That includes us, to be sure, but in our roles as objects of study, not as the theorists doing the studying.

Lauren Olin

I love Olin's anecdote about Russell, Shaw, and Webb; it's really funny! I'm delighted to have something fresh to think about.

I expect there must exist some psychodevelopmental truth as to what makes a thing funny to a person, but the attempts that Olin surveys come across to me as lame, through no fault of hers. I don't know how to do better; I wish someone did. I do think that the comic response, finding things funny, is a human universal, even if what people find funny varies from culture to culture.

As for how to analyze the meaning of claims as to what's funny and what isn't, I might possibly agree with Olin, but I'm not sure I understand all of what she says. So let me give my own rough view, which ties in closely with what she quotes from Ted Cohen. Oftentimes, I'd say, we engage in working toward a common sensibility—and Cohen explains some of why this is. It's then that bald claims as to what's funny and what isn't are meaningful. Here we could emulate Stalnaker's talk of proposals to add to the common ground in a conversation: asserting "That's funny" proposes adding finding it funny to the common ground. Note that here I depart somewhat from Stalnaker in that as he formulates matters, the common ground consists of propositions, whereas I treat it as consisting in states of mind on which we have openly come to accord. These states of mind can include emotional attitudes toward things, degrees of belief ("Yes, probably!"), and acceptance of norms governing belief. You can propose adding opposition to police brutality to our common ground, or wanting drunken driving and sexual assault to be effectively deterred, or both. Sometimes too, though, we find we aren't getting anywhere with this pursuit, and we give up on it. We can then retreat to another subject, one on which we can still hope to agree: what's funny to me and what's funny to you. So on my view, when I said that the Russell anecdote is funny, I'm not literally saying anything about my own state of mind; I'm not issuing a subjective attitude report, a psychological description indexed to me. What I'm saying has no such reading. That contrasts with talk of what's "funny to me", which of course does index what I say to me—explicitly. Going from the one to the other, from saying "It's funny" to saying "Or at least it's funny to me" is a retreat. If I say baldly "It's funny," you have various possible responses: not only "Yes it is," or "No it isn't," but also "Well, it isn't funny to me." This last doesn't directly contradict what you said, but indicates that we're unlikely to achieve accord on the matter, and so we'd better shift to a related topic and stick to what's funny to whom. At this point, you may well allow further that there's no truth of the matter as to whether it's funny, but you don't have to.

Olin distinguishes apparently faultless comic disagreement from genuine, and speaks of views on what's appropriate. Predicates of personal taste are a prime example; they don't give rise to any real disagreement. I think we can characterize what it is to regard a predicate as one of personal taste. To do so, we can invoke what I misleading called planning, thinking what to do or how to feel if one is such-and-such a person in such-and-such circumstances with such-and-such characteristics. I can't stand asparagus, but my wife and many others love it. I can ask myself how to feel about eating asparagus in case I am she with her tastes,

and for this case I settle on loving it—whereas for the case of being me with my tastes, I settle on abhorring it. Delight in torturing people is different: for the case of being someone who delights in torture, I demand a change. This is no mere matter of personal taste, say I, and it is far from faultless. My conviction that torture isn't just a matter of personal taste consists in settling on abhorring it in case I'm someone who loves it. (I knew an Australian who abhorred wearing socks with sandals and didn't seem at all to regard this as a matter of personal taste; I thought this view of hers silly.)

What is it, then, to improve one's comic sensibility? Again, I answer in terms of what it is to believe—to believe, say, that exposure to the right people would improve my sensibility. This is to settle on responding as I will once I have undergone their influence. Alternatively, I could think that the ways exposure to these people and their jokes will make me constitutes not improvement but degeneration. That's to settle against responding the ways I will once I have undergone these influences. As for wondering whether a joke is funny, a joke I understand, that consists in thinking how to feel about the joke, not having yet settled on an answer.

Olin objects that one can't plan for mirth. But can't one ask oneself how to feel in a circumstance—including whether to feel mirth? True, if the case arises, what I've settled on won't necessarily control how I do feel. But as we are constructed, I think, the opinion can have some influence. In a like vein, a person can keep feeling awful about himself while being convinced that this isn't the way to feel. This just means that normative governance, like all governance, sometimes prevails and sometimes doesn't. For the most part, one's view on how funny something is adheres to how funny one finds it, but not always. And I agree that if you don't find a joke funny, you can think that it must be funny—because, say, a friend whose comic sensibility you admire finds it funny—and still be far from cracking up.

Phrases like 'that damn Kaplan' (which I wouldn't say for any Kaplan I have known) we might call *expletive adjectives*. As Olin indicates, they do behave in special ways. Like most other adjective tied to feelings, such as 'horrible' and 'wonderful', they act syntactically, under conditionalization and negation, much like descriptive adjectives such as 'crafty'. Consider

> If that crafty Figaro messes up, he's in trouble;
> That crafty Figaro is winning.

Both carry the implicature that Figaro is crafty, but no straightforward response like "No he isn't" dissents from this characterization—as it does from "Figaro is crafty." What's special about these terms is not their syntactic possibilities but how their meaning is to be explained. What Huw Price calls "representationalism" won't work for them; it isn't truly explanatory.

When I talk of what's funny and what isn't, I am indeed talking of whether amusement is appropriate, but this requires not only more careful formulation but also that we explain the term 'appropriate'. On my view, to say that amusement is appropriate in a circumstance is to allow it to oneself for the case of being in that circumstance. When I say a joke is funny, I'm

saying that for anyone who understands it, amusement is appropriate. As Olin and Cohen stress, understanding a joke is a matter of being aware of the lore involved, whether or not one goes along with it. To my taste, most ethnic or racist jokes aren't funny—but I seem to remember a time when I found a racist joke funny even though I deplored it. (Fortunately, I can't remember what the joke was.) I do find the "hungry for power" joke funny and I'm not disturbed by it—even though I reprehend automatically stigmatizing Germans as pro-Nazi.

I've been expressing views off the top of my head, sort of, on questions Olin so delightfully raises. I don't know how much of what I've been saying departs from her. But reading her and thinking with her about the comic has been great fun.

She and Peter Railton both say helpful things about what I would call emotivistic expressivism. I'm glad to learn of Russell's clear statement of it, preceding Ayer's. As I keep saying, Stevenson got there only later.

Peter Railton

I am of course immensely gratified by Railton's commentary, by his description of what I have maintained and how it may have influenced his own directions. Railton says things I believe, and says them far better than I can. His "tendentious histories" remind me of some things I had been aware of, and also of things I had forgotten. I have thought of Ayer rather than Hume or Stevenson as the founder of expressivistic emotivism, but I'm glad to learn from Olin of Russell's clear statement of the thesis, preceding Ayer's. I think I've learned of other sources, only to forget what they were. But Hume isn't a clear source; as Railton says, he "was more focused on psychological and explanatory questions than on semantics" (347), and in more than one instance—with causation as well as with morality—he will treat as equivalent three things, in three successive sentences, that we now read as offering competing accounts. One of the three will neatly formulate expressivism, but Hume didn't distinguish this from the other formulations, as Ayer did. Railton is entirely right that Ayer makes no attempt to distinguish justificatory reasons from other influences on our attitudes; he has no account of what it is to claim something as a reason supporting an ethical claim, as opposed to other ways of trying to affect a hearer's attitudes. He seems to distain trying to make such a distinction, and this is a serious defect. In his practice, though, Ayer pursued moral causes in the same sorts of ways as the rest of us, justifiably when he opposed Jewish quotas for "public schools", and unfortunately when he joined so many others in Mao worship.

As for Stevenson, he had things to say, as Railton reminds us, about the looseness of language and how "We all like to be neat" can be more than a plain lie. But he favored the subjectivist locution "I approve of *X*" as loosely getting at what he was saying. (In the book, I mean; in his original article it's "I like *X*.") Ayer, unlike Stevenson in his initial writings, makes clear the difference between reporting a state of mind and "evincing" it, the crucial distinction for expressivism. I had thought that the "Do so as well" was Stevenson's contribution, but Railton reminds me that Ayer says that we try "to affect another person in such a way as to bring his sentiments on a given point into accordance with one's own" (349). So I don't

think that Stevenson's was a "more sophisticated version of emotivism" (350). I was always puzzled by Stevenson's descriptive semantic subjectivism, as Railton accurately labels it, as a good approximation to what he wanted to say, and indeed I seem to remember Frankena and Brandt saying that he insisted on it. Kevin Toh, when he was a graduate student, revealed to me what had happened: Stevenson's 1975 collection of papers *Facts and Values* includes a long recounting of his later thinking on a number of issues, including this one. He there explicitly comes over to the expressivist side. Stevenson must, though, be credited with the idea of "disagreement in attitude", which I think tremendously important—and his discussion of first-person deliberation, which Railton highlights, is masterful.

I worry slightly over things Railton says about naturalistic definitions of ethical terms, as Moore called them. Some of what he says might be taken in a sense I would reject, and that I don't think Railton intends. Moore's point in rejecting "naturalistic definitions", as I see it, was that if we define *better* as meaning, say, encompassing more happiness, we then can't use the term to ponder whether to favor all and only happiness. The things Railton says that worry me are ones that seem to indicate a boundless falliblism: "Any account or criterion we develop could still be inadequate to capture all that is at stake," and "We cannot take our current point of view or commitments as guaranteed to be correct, but must be remain open to the idea that we could wrong even in our most central convictions or assumptions" (358). I regard myself as highly fallible, but not in every respect. I'm certain that I mustn't torture a person for amusement, even if I indeed would find it amusing, and in less artificial terms, I'm certain that the Nazis were morally appalling. Railton is precisely right, however, that I insist on the "logical or conceptual independence" of what to do from any "definition to which we might attempt to resign our capacity to deliberate, decide, and act" (359). That's not necessarily because of doubt, but because this concerns, as he says, "logical or conceptual independence". I don't think that Railton accepts the extreme sort of fallibilism that would worry me, and so perhaps I shouldn't even have brought it up—but I thought I should be sure to distinguish the things that we should agree with in from any such a position. What he says also shouldn't entail that we could never arrive at a correct comprehensive account of how at base to live. Maybe we can't, but that's a further matter. Railton is right, I think, that even if we did, it we would still need to have terms in which to inquire into it. Even if we found the view beyond doubt, we'd need to understand what it would be to doubt it.

Railton quotes me as saying, "Well, at this point I'm more interested in developing my anti-utilitarian intuitions than in pushing them aside" (359). I don't remember saying this, and don't know what I would have meant. But it's true that I don't think we can get along without intuition entirely. Rational inquiry can only proceed from some cautious, defeasible faith in our powers of judgment.

I am fascinated by Railton's elucidation of a sense in which Ayer, despite rejecting "descriptive semantic subjectivism", was in a broad sense a "radical subjectivist", allowing "no question of ethical truth, accuracy, evidence, or knowledge" (353). The expressivist thesis by itself—that normative meanings are to be understood as expressions of being, in some sense,

for or against—doesn't entail standards like the ones Railton lists as "marks of *objectivity* in a domain of thought and practice", his (i), (ii), and (iii) (349). But when I start thinking in the ways that expressivism describes, I join Railton in these claims. One of my first papers on norm expressivism I named "Normative Objectivity", but I never got nearly as far as Railton has in elucidating how to understand objectivity in the sense we need. This is one among the many ways that I find his commentary immensely helpful and clarifying.

Billy Dunaway

Dunaway characterizes realism in a way I find promising, and raises questions that I don't know how to answer. I'll try briefly to explore some questions that arise in my mind, but I'll have to be quite inconclusive.

His aim, he says, is "to characterize realism in a way which does not leave it susceptible to quasi-realist accommodation" (362). He carefully doesn't claim to have accomplished this aim definitively, and he doesn't pronounce on whether Blackburn's or my own version of "quasi-realism" is a form of realism. He does, though, claim to have shown that if he has characterized realism correctly, we quasi-realists haven't shown that expressivism can capture the whole of normative realism (or more narrowly, of ethical realism). This contention strikes me as fairly plausible, and explaining a notion of realism should allow us to evaluate it—but I won't try to come to a definite conclusion on this score.

Dunaway explains realism as a matter of the comparative degrees fundamentality one ascribes. This explanation has the great virtue that it allows us to classify the standard view of "polywater" as irrealistic along with behaviorism. It also has what might be a vice: doesn't it classify the predominant scientific view of what it is to be alive as a form of irrealism? After all, the kind of fundamentality the view attributes to living falls short of the kind that vitalism attributes. I don't see this, however, as decisive grounds for rejecting Dunaway's analysis. If one feels hesitant about this, it's perhaps because irrealisms have a flavor or unmasking. When vitalism was a prominent alternative, the view that has become standard served to unmask it, but by now we may be so accustomed to the standard sort of view that we don't find a mask to tear away.

An alternative might be drawn from Huw Price's talk of "represenationalism", which I mentioned at the start of these responses. Many of Dunaway's examples of realism in theories fit this proposal—and pretty clearly expressivism isn't representationalist. However, I end up agreeing with him that this story of what's at issue may not work for everything; his example of polywater is pretty convincing. On the standard view, perhaps, the term does refer to a kind of substance, but in a way that isn't of much fundamental interest. So the standard view is representationalist for polywater, but not realist. Score this as a vindication for Dunaway.

So fundamentality is what we'd better pursue. One question I have is how normative fundamentality fits into this. One kind of normative egoist, for example, holds that the thing to do is always what maximizes one's hedonic prospects; that, she will claim, is

the fundamental normative principle for action; it is what lies behind reasons not to lie and reasons to be kind. This is a kind of fundamentality that I recognize. What is fundamental normatively won't be a matter for a metatheory of normative concepts, but a substantive matter of normative theory. Is it a matter of metaphysics? The answer, I presume, is no, but I would like to know how the contrast works. Dunaway helps himself to an understanding of being metaphysical, without a need for much explanation. I understand the notion for important contexts. For example, the Copenhagen interpretation of quantum theory, I might say, tells no coherent metaphysical story. What it treats as real differs for measuring devices and for what's measured. The many worlds interpretation does have a metaphysics, however fantastic we may find it: what's real, it says, is the universal wave function, of which you and I are features. On this interpretation, when I decide by the flip of a quantum coin whether to shoot myself in the head, there's a feature of the universal wave function that constitutes how things went "in my world" where the coin landed heads and so I didn't shoot myself, and a feature that constitutes how things have gone "in a world" where the coin lands tails and the continuation of the old me shoots himself. So for this sort of case, I recognize what's metaphysical. I don't, however, know how to extend this to the question of whether, on an expressivist's metatheory and the normative egoist's theory, what to do is a metaphysical question.

My central question about talk of "normative realism", though, remains this: Take something that has a highly fundamental bearing on what to do—say, being worth wanting. This, Dunaway's normative realist might tell us, is a real property of high fundamentality. Why go for it, then? As a general question, this seems empty. For each thing that's worth wanting—pleasure, for instance—we can sensibly try to find an answer, but not for plain being worth wanting. One answer I've heard to this sort of question is that to say that something is worth wanting just is to say that one ought to want it or that it's all right to want it. So we're asking why it's the case that one ought to want what one ought to want, and of course that question is empty. My point, though, is that the concept of being what's worth wanting gets its nature from its role in thinking what to do. Someone who didn't accord it that role wouldn't have the concept.

True, she might pick up the term from someone who had the concept, so that the phrase in her mouth signified what it signifies for us who have the concept. It might be like the concept of Madagascar for someone with no idea of where that is. The person with such a transmitted usage might be a normative realist, insisting that being worth wanting is a real property, and we can work further to determine its nature. The crucial question, though, concerns the primary possessors of the concept who understand what they're thinking. That it's a real property won't be important for how they conceive things; what will matter is this role in settling what to do.

So can I be a normative realist? Is being worth wanting (or something else normative) a property, a property with a high degree of fundamentality? As I have indicated, I'm inclined to think that for being worth wanting, this question misses what's crucial.

V. The Normativity of Meaning

17. Paul Boghossian (New York University), "The Normativity of Meaning Revisited"
18. Paul Horwich (New York University), "Obligations of Meaning"
19. Jamie Dreier (Brown University), "The Normative Explanation of Normativity"

Paul Boghossian, Paul Horwich, and Jamie Dreier focus on the metatheory of meaning in my 2012 book *Meaning and Normativity*. Before I respond to their critiques individually, I should highlight, of things I messed up in the book, the greatest blunder I'm aware of. I gave no explanation of what we might call normative explanations of things that happen. What a person's words mean is, I claimed, a normative issue, but obviously what a person's words mean helps explain many things about the person—especially what happens in her discourse and explicit thinking, the sentences she utters or thinks to herself. How does this work, on my views?

A few decades ago, there were controversies about "moral explanations" of things that happen: that, for instance, the badness of a regime caused a revolution against it. "Cornell realists" argued that the cogency of such explanations showed that moral properties are natural properties, signified in the same ways as other natural properties. Simon Blackburn and I, among others, argued that our expressivistic analyses of moral terms allowed for such moral explanations. "The revolution was caused by the badness of the regime" can amount to a long, perhaps infinite disjunction of the form, "The regime was bad in that it was brutally harsh, and its harshness caused the revolution, **or** the regime was bad in that it was permissive to the point of laxity, and this permissiveness caused the revolution, **or**" and so forth. This same pattern allows too for explanations of the causes of events when they are couched in terms that are normative more broadly, normative but not specifically moral—in epistemological terms, for instance, in terms of what to believe. This, according to the hypothesis I develop and explore in the book, includes explanations of events in terms of meanings. "The insulting meanings of the things St. Stephen said caused him to be stoned to death," a reader of the New Testament might say. This could amount to an infinite disjunction of claims of the form, "Having an insulting meaning consists in having property *P*, and St. Stephen's words' having this property *P* caused him to be stoned to death."

Why, though, go through such contortions? Because the concept MEANING must answer to two demands, for explanation and for guidance. The demands, again, are that meaning figure in explaining certain linguistic phenomena causally, and that it guide us in engaging with each other's thinking and with our own. Now it may well be that the properties that make it advisable to engage a person in certain ways are the ones that Horwich identifies. My example of interpreting what Newtonian physicists meant by the term 'mass', though, was intended to show that this might not always be the case, that one could coherently maintain

otherwise, still recognizing all the facts as specified in causal/explanatory terms. Newtonians applied the term 'mass' to what they took to be the actual ways the world is, and would be confused if they felt called upon to apply the term to the ways Einstein's special theory of relativity says the world is, with his distinction between rest mass m_0 and relativistic mass m. Horwich, in his critique of my account, extends his theory of meaning to apply to such thinkers: as applied to special relativity, he says, the meaning of their word 'mass' is indeterminate. This, I say, amounts to telling us not to settle, for the case of being such a thinker, what sentences in one's language to accept for cases where the universe is relativistic rather than Newtonian.

That may well be good advice, but I have heard knowledgeable people voicing firm opinions that Newtonians mean one or the other, REST MASS m_0 or RELATIVISTIC MASS m, by their term 'mass'. I try to identify what is at issue, among other things, between those people and Horwich in their substantive theories of what Einstein's terms meant.

I should add that there is much to be said for taking the advice that Horwich in effect offers—namely, for the case of being a Newtonian contemplating the possibility that the world is relativistic, not in one's thinking to apply one's old term 'mass' to cases where rest mass and relativistic mass diverge. Here the pragmatic issue of how to engage people is relevant: one should engage fellow Newtonians not by making straightforward claims about "mass", but rather by bringing them to change their language to one with clearly different terms for rest mass and relativistic mass. We should exclaim, "Forget what we meant before!" My metatheory of meaning explains what one is then doing. One is changing the language in which one thinks, and bringing those who have been Newtonians to change theirs. Once we all make the change, the simple core of Horwich's theory, that the meaning of a word is its basic acceptance property, again applies to us in a clear way and explains which sentences in our thinking we accept, which we deny, and which we neither accept nor deny.

Paul Boghossian

I owe aspects of my thinking on these issues to Paul Boghossian in more ways than I can keep track of. I think indeed that I read his work on Kripke's Wittgenstein before I read the Kripke treatment itself. As I'll indicate, there is much in Boghossian's ways of thinking about these issues that I accept and applaud. On the other hand, in some ways, our framings of the issues are so different that it is hard to mesh them together into positions on the same set of subissues. The statement of conclusions that Boghossian ends up with I entirely accept as he words them, but on the way to this, I in some ways frame things quite differently from the ways he does.

Begin with two preliminary questions. First, Boghossian objects, even if I give the right explanation of beliefs and their content, still it won't apply to propositional attitudes other than beliefs. What about hopes and fears? What about desires and intentions? "Mental content features in a host of *other* propositional attitudes besides belief, desiring and hoping, for example, to which *different* norms apply, if any" (395). Whatever normativity is in play,

Boghossian proposes, lies not in concepts of particular items of mental content, such as the concepts SOMETHING and NOTHING, but in propositional attitude concepts such as the concept BELIEVES.[11] This highlights a choice we face. What's normative most clearly is combinations: combinations of attitude and content such as "believes SNOW IS WHITE". We can try attributing this normativity to either component, either to the content or to the attitude. More precisely, we can try saying either that what's normative in all this is concepts of particular items of mental content, such as SOMETHING, or that it's concepts of particular propositional attitudes such as belief. The normativity of "believes SNOW IS WHITE" might stem either from the normativity of the concept of the content "that snow is white" or from the normativity of the concept of the attitude "believes". I go for the former; Boghossian goes for the latter.

On the standard schema, various distinct attitudes can govern various distinct propositions, so that I can *intend* to dance and *believe* that I'm about to dance—and these amount to intending and believing the same proposition I'M ABOUT TO DANCE. Whatever normative is in play comes, Boghossian proposes, from attitudes and not from items of content. What I propose instead is to take as basic the various beliefs one might have, as mental states to be identified in normative terms. Then understand intentions and the like as relations to corresponding beliefs. This has advantages, as I'll shortly indicate.

Consider my intention to dance. We can specify it via its relation to the possible belief that I am about to dance. If we want a quasi-regimented language for this, we can render intending to dance as *intending-realized* the belief that I am about to dance. Beliefs stand in conceptual relations to each other, such as the one between believing I'M ABOUT TO DANCE and believing I'M ABOUT TO DO SOMETHING. We can render the entailment relation between these propositions in normative terms as follows: consider the entailment

> "I'm about to dance" entails "I'm about to do something."

This, I say drawing on Hare, amounts to saying that if I ought to believe the former then I ought to believe the latter. This can give rise to conceptual relations among other propositional attitudes such as intentions, but I propose taking possible beliefs, identified in normative terms, as the primary basis for identifying propositions. Normative relations among other propositional attitudes will ensue. For example, if I intend to dance, I intend to do something. This comes out, in a variant of the pattern that Hare proposed, as

> If I intend-realized belief in I'M ABOUT TO DANCE, then
> I intend-realized belief in I'M ABOUT TO DO SOMETHING.

---

11 I intended to address this criticism of Boghossian's years ago in a session at a conference, but as I listened to this complaint and thought about how to respond, I kept feeling sicker and sicker. I soon realized that I wasn't sick at the thought of how to respond, but from a stomach bug. So, I had to dash away from the session before I could say anything.

Oughts governing beliefs thus give rise to relations among intentions. The same will go, I suspect, for other propositional attitudes.

A chief advantage of proceeding this way is that it gives us an informative way of specifying and distinguishing items of content. We can of course specify items of content by the words we use to voice them, as I AM ABOUT TO DANCE is voiced, among us speakers of English, by the words 'I am about to dance.' That lets us draw on our language competence, but it isn't informative in any systematic way. A more informative way to specify items of content is via relations like entailment. I harp on the concept SOMETHING, identifying it not only as the one we voice as "something", but also in terms of entailment relations and the like. These, as I say, can be put in normative terms in the way proposed by Hare; instances are

> If one ought to believe SNOW IS WHITE, then
> one ought to believe, if the question arises, SOMETHING IS WHITE.

Normative ties to experience will figure as well in the specification of other concepts such as GREEN. If instead we go Boghossian's way, placing whatever is normative entirely in attitudes like belief and intention, entailment must come into the story from outside, either without explanation or with a separate story of its own. Boghossian's story is this: "In a suitably broad sense of 'evidence,' I have undefeated evidence that it can't be the case both that snow is white and that nothing is white" (395). This impossibility isn't attributed to anything in particular, such as the logic of the term 'nothing'. But the impossibility does seem to follow from this logic, and we can put the relation in Hare's way. Thus by starting with beliefs and normative relations among them, we can say systematic thing about the logic of the concepts in play. I urge, then, that we should identify items of content in the normative way I have sketched, and value the systematicity we thereby attain.

I'm reasonably satisfied with my contention that explanations of meaning start with beliefs. I'll turn now to a matter that I find more difficult: how to specify beliefs and their content. As Boghossian reports, I treat belief as accepting sentences in one's own language. Very roughly, this follows Paul Horwich—but as Boghossian points out, if this were my entire account, it would fail in a way that Horwich's own account doesn't. Horwich has in mind a language of thought, whereas I speak of the ordinary language that a person thinks in. This, however, is not my entire account. I recognized that, as Boghossian points out, some of my beliefs will be nonlinguistic, in that they aren't formulated explicitly in a language I know and think in. He gives the example of looking at a scene; I will believe many things about it that I couldn't put in words. How, then, do I want to handle aspects of beliefs that aren't encoded in any language available to me?

I did talk about this in the book, but tersely—and, I discover too late, a reader would be hard put to locate where. (Indeed it isn't even in the index, which should have included the term "nonlinguistic" under the heading "belief".) The relevant passage in the book is the first full paragraph on p. 132. What I say there is vague, but the idea is this: A belief formulated

in one's language gets its meaning from its inferential ties with other sentences—along with observation. This can be generalized to include nonlinguistic beliefs. My project owes a careful explanation of this, and I'll shortly say a little more, but not enough. Very roughly: Start with Hare's explanation of inferential ties such as entailment. SNOW IS WHITE entails SOMETHING IS WHITE in that if one ought to believe the first, then one ought to be disposed to believe the second. Similar patterns will hold for beliefs that aren't explicit, that aren't couched in language. So with Hare's account of entailment, we can say that any belief, linguistically couched or not, gets its content from entailment relations between it and other belief states—including observational states. I owe a more careful development of such an account, but this would be the rough idea.

Saying this requires, however, that we be able to designate a possible belief in some way that doesn't presuppose the inferential relations it stands in. As Boghossian notes, I deviate from Horwich in an important respect: I speak of belief as a state of accepting a sentence in one's own language, whereas Horwich adopts a "language of thought" hypothesis. I now don't think that either my characterization or Horwich's captures what needs capturing. The language of thought hypothesis may be right, but it seems dubious. The mental language is supposed to be like natural and familiar artificial languages in some ways, but without the usual limitations of such languages as ambiguity and uncertain reference. Various artificial languages have many of these features, but I don't know any good enough reason to think that such an idealized language figures in normal thinking. When a sentence such as 'I had a book stolen' is ambiguous, we do seem to be able to keep track of alternative readings, and accept the sentence on one reading and reject if on another. Whether we do this by translating each reading into the language of thought, though, isn't clear. Like things go for ostention: With a sentence like "That's a dog," we are often able to keep track of what's being indicated with the word "that". What role a language of thought might play in this would need elucidating.

Early Chomsky had tree structures doing the work of disambiguating, and perhaps that is the way to handle ambiguity, and so I might have included in accepting a sentence in one's own language its tree structure. To take Chomsky's example, the string of words "I had a book stolen" can be thought with three different tree structures. That leaves other problems, though. Boghossian notes perceptual beliefs that don't seem to be linguistically encoded. And with "That's a dog" or "C'est un chien," we need something giving what one is talking about. (The same goes for the time of a past tense sentence, and for seeming reference to God if there is no supreme being.) What's needed can't literally be a relation of reference, since as with God, there may be no entity to stand in that relation. I'm not aware of a standard way among linguists to identify beliefs apart from specifying their content, or even a satisfactory way. I can identify a thought as "the thought I just had", but that doesn't get us far on the way to systematicity.

I might need correction on this score, but my impression is that these are perennial problems against which going theories crash. So I'll have to be agnostic about how to characterize

a belief state in a way needed for Horwich's purposes or for mine. That's unsatisfactory, to be sure, and we can ask whether there is any way of referring to a belief state without ascribing particular entailment relations to it.

One more small and inconclusive matter: Boghossian thinks that a dispositional account must appeal to ideal conditions, and he finds a problem with such an appeal. The justification must be internalist, he says, in that "it must be something that I am aware of and whose relevance to my action I can somehow see." This can't hold "for a justification grounded in some counterfactual about what I would do if conditions were ideal" (400). My own thinking on this is not in terms of ideal conditions. Perhaps one could get an account of ideal conditions out of it, as conditions of using words as one ought to, but I don't depend, in the characterizations I offer, on anything about ideal conditions.

Another difference between us is that I don't center my analyses on rule-following. As Boghossian recognizes, most language isn't a matter of following rules explicitly, but things Boghossian says seem focused on the explicit case (399).

> If S is following a rule R on a given occasion by doing A, then (1) S has *accepted* R, (2) S's acceptance of R determines whether what S did was *correct*, (3) S's acceptance of R *explains* why S does A, and (4) S's acceptance of R rationalizes or justifies her doing A.

In the usual, implicit case, there has been no initial explicit act of acceptance. Moreover, if there has been, the person may not be guided by it accurately. We don't get at what the rule is by learning what the person thinks it is; what rule a person is following must be discerned by an observing theorist. It's a matter of how to explain the person's proclivities—including reactions of satisfaction or dismay at what one has done. In many cases, to be sure, an initial act of acceptance may help explain, rationalize, and justify what the person goes on to do, but that, I would think, isn't centrally what we should be explaining. More basic is what's going on in following rules implicitly. Despite Wittgenstein and Kripke, in many cases, putting things in terms of "following a rule" may not much help. Agreed, in the usual case, the person is disposed to act *as if* there had been some occasion of explicit acceptance that guided later responses and lets us as observers assess match between the rule and the performance. But we are left to discern what it is about the subject's dispositions to act and react to her actions that makes for this "as if".

Note also that although in the case of language, it may be that whatever rule one is following justifies whatever one does that accords with it, that's special at best for language, with its legitimate arbitrariness. The vicious torturer, say, who follows a rule isn't thereby justified in what he does.

So turn now to the "really difficult" question that Boghossian poses and somewhat answers: "how my grasp of the meaning of a word could *justify* future applications of that word" (401). That's officially out of the range of things I try to address in my metatheory of meaning; it is a question in the substantive theory, not the metatheory. But I do make vague assumptions

as to what an answer might be. I suggest (in a way that alludes to Quine) that what's justified in usage is determined by a balance of simplicity and closeness of fit—where what precisely constitutes the relevant standards of simplicity, of closeness of fit, and of the balance between them must be specified if such a theory to be fully determinate. Again, how to specify these things are questions for a substantive theory of meaning to answer. (An alternative substantive theory to use as an example might be drawn from Horwich.)

Back to metatheory, then: once we have a candidate for such a theory, how is it to be assessed as giving a right or wrong answer? At root, this will be in the same sort of way as we assess other questions of justification—by intuition, as we might say, the same general sort of way as we might come to accept moral standards. We are asking in broadest terms what sentences in one's own language to accept. (Or better, as I said earlier, what belief states to be in, whether or not they consist in explicit acceptance of sentences in one's language.) In addressing questions like this, we will be guided by our judgment as to what justifies what and how those judgments fit together. This is close to what Rawls calls seeking reflective equilibrium.

All this, I confess, is frustratingly vague and general, insofar as it is meant to be more precise than the adage "Balance closeness of fit with simplicity." Boghossian has other things to say, mostly in other things he has written, but I'll stick, unfairly, to what he says in his commentary here, and make a few remarks. I read Boghossian and me as pointing to roughly the same things. He centers what he says on "grasp" of a concept. "I have a *reason* for using that word as opposed to another, a reason that is provided by what I mean by the word. So, our grasp of the meaning of an expression has to be able to justify our use of that expression" (399) What does such grasp consist in? We must appeal "to the notion of an intuition or an intellectual seeming. Our grasp of the meaning of 'green' supplies us with an intuition to the effect that it correctly applies to the presented object" (400). That is close to what I am saying: I agree that we'll look for an account that vindicates and systematizes the bulk of such intuitions.

So why accept those sentences in one's own language that fit a certain balance between simplicity and closeness of fit? On this, intuitive and pragmatic considerations agree. Doing this is both conservative and mildly reformist. Its pragmatic virtues are those of predominantly doing what comes naturally without elaborating contortions. That raises important old questions that I won't here try to address. How do pragmatic qualities bear of things that aren't actions, such as believing—which often consists in accepting sentences in one's own language. How do the norms one accepts for what one is doing relate to what one does, when it's something nonvoluntary like accepting a sentence in one's own language. As for intuitive considerations, I speculate that intuitive judgments of what people mean will match the upshots of balancing simplicity with closeness of fit. As I say, I think of myself as following Quine in this. But of course this speculation would have to be checked out, which is far beyond anything I could do here. "Grasp", I take it, is a mental state that immediately gives rise to intuitive judgments that a word applies. Boghossian, I presume, doesn't think that such judgments will be infallible, but they bear great weight, I agree. In

these senses, I agree that grasp "sets a *standard* of correctness for that behavior; it explains it; and it *justifies* it" (399).

I thank Paul Boghossian for this marvelous commentary, and I hope that despite our different ways of framing the issues, we are converging on a common view—both on what the terms 'meaning' and 'mental content' mean and, insofar as I hint at a vague substantive view, on what meaning and mental content consist in.

Paul Horwich

It is thrilling to have such a powerful theorist considering my efforts on the meaning of meaning so intensely and incisively. His critique leads me to move substantially toward his positions. But in a sense I'll explain, his theory may endorse my own metatheory of meaning.

Horwich suggests in a note at the end of his commentary that perhaps "it's just that I've given myself one job and he's given himself another" (436). I think that's right, in more ways than one. First, as I keep saying, Horwich chiefly elaborates a substantive theory of what meaning properties consist in, and does so wonderfully. I'm asking a different question: what such a theory means, what it is claiming. I say at the start of my 2012 book *Meaning and Normativity* that I would much rather have devoted my efforts to the chief job Horwich takes on, the job of developing a substantive theory of meaning—and I would have done so if I had thought I knew how to contribute to the project. That brings us to the second way: Horwich and I develop different views as to how the concept MEANING works, and we each take on the responsibility of trying to get our respective views to be as well developed as possible, so that they can compete in their strongest forms.

That said, Horwich too has a metatheory of meaning for naturalistic terms—one that is powerful and plausible. In his commentary, he adds to this metatheory in a way I find plausible. Are there then two eligible interpretations of the concept of meaning, his and mine? I'll still argue that there are—although with Horwich's metatheory as he refines it, it may now be less urgent to develop a metatheory like mine that treats the concept of meaning as normative. I still claim that my way of developing this metatheory makes best sense of the line that Kripke's Wittgenstein leads us into, though I won't be arguing this exegetical point beyond what I said in the book. If I'm right in what I now believe, there are indeed two different coherent ways to elucidate our concept or meaning, but I now see the advantages as weighing more in Horwich's favor than I previously thought they did.

My own metatheory of meaning, in rough terms, is that what a person's words mean is a matter of how to use those words if one is that person—more specifically, under what conditions to accept sentences with those words in one's own language. An attribution of meaning, on this view, is thus not purely causal/explanatory, and a purely causal explanation of the person's responses doesn't fully settle what her words mean in this sense. In presenting this metatheory, I did keep using a roughly couched substantive theory of meaning as an illustrative example. I didn't advocate this theory, but I did claim that a theory along such lines isn't refuted by anything that Kripke's Wittgenstein establishes. Alternatively, I might have

used Horwich's substantive theory of meaning as my example for illustrating what accepting a substantive theory might consist in.

As I say, Horwich does have a metatheory of meaning to go with his substantive theory. In the first place, he says, "The meaning-properties of words are engendered by underlying properties of the sort that can best accommodate our pre-theoretic convictions about meaning" (436). I agree. But that might be done not only with causal intuitions but also with normative intuitions as well. That is how I proceed with my account of the concept of meaning as normative. More specifically, he holds that our term 'meaning' works in a purely naturalistic way, like our term 'valence' in chemistry. Its role is entirely naturalistic/explanatory: its role is to contribute to explaining in purely causal terms, among other things, which sentences, in a person's language, that person accepts in various circumstances. Taking my cue from Kripke's Wittgenstein, I myself think that the concept MEANING has a second role. Along with causal explanations of how our language works, we have decisions to make about how to engage each other in our discourse, and how to engage one's own thinking. These decisions are mostly implicit, but sometimes we puzzle them through explicitly. "I don't know what you mean by that word" can be a challenge in conversation, and "What do I mean by that word?" can be a question I pose to myself in order to clarify my thinking. Meaning in the sense of how to use our words figures centrally in this, in deciding how to engage the thinking of others and of oneself. Of course, how *to* use a word depends heavily on how we *do* use it, and on what, in purely causal terms, explains our usage. Still, there can be disagreements on how to use a word even among people who agree on all purely causal matters. The hypothesis I explore in the book is that such disagreement is normative; my project in the book was to understand what could be at issue in questions about meaning.

Before I engage the bulk of what Horwich says, a couple of side notes: First, after presenting a beautifully accurate summary of my views, he sometimes puts matters in ways I would not. In a note, he says that I deploy a coursed-grained sense of "property" whereas his sense is fine-grained. I find it hard to discern whether this is just a matter of our using terms in different ways, or whether, rather, the view Horwich criticizes isn't quite my own. He does acknowledge my sharp distinction between properties and property concepts. (Correspondingly, I distinguish *states of affairs*—which, roughly, are structures of properties—and *thoughts*, which are structures of concepts.) In my lingo, the property of what a person's word means is natural, whereas the meaning of an assertion as to what it means is normative. For example, Frenchman Pierre's sentence 'C'est un chien' signifies the *natural* state of affairs that it's a dog (concerning what conversational attention is focused on). In saying this, however, I am making a *normative* claim; I am voicing a normative thought: I am saying how to use one's word 'chien' if one is Pierre—how to *use* it in the sense, again, of which of one's own sentences to accept in which circumstances. (In the example, the question is for what circumstances to accept one's sentence "C'est un chien" if one is Pierre.) Horwich renders all this accurately, but he also puts it in a way I would not: "Regarding the *phenomenon* of a given word possessing its particular meaning: is such a fact normative 'all the way down'?

Or is it ultimately and entirely constituted by the word's naturalistic properties?" (402) He speaks of the issue between us as whether "meaning-properties are constituted naturalistically" (418). This is misleading, it seems to me: I say that meaning "That's a dog" is a natural property of Pierre's sentence, but the claim that his sentence signifies this property is normative. When Horwich also puts the issue as concerning the "fact" of what a word means, I protest that word 'fact' can mean either of two different things. It can mean STATE OF AFFAIRS THAT OBTAINS or it can mean TRUE THOUGHT. The state of affairs of a word's meaning what it does, I maintain, is natural, whereas the true thought that it so means is normative. The *state of affairs* of Pierre's word 'chien' meaning DOG is natural, whereas the true *thought* that Pierre's word 'chien' means DOG is normative.

Second, my claim is not that understanding meaning claims to be normative resolves indeterminacies of meaning. It tells us what is at issue between people who resolve them in different ways. My account can tell us what's at issue between thinkers who agree completely on the facts about the speaker or thinker as naturalistically specified and explained, but who attribute different meanings to her words. With the Newtonian physicist who grapples with Einstein's special theory of relativity, what explains her responses is that she accepts Newtonian physics and hasn't clearly conceived of relativistic physics as a possibility. But that doesn't by itself settle how her words apply—truly, falsely, or indeterminately—to the possibility of special relativity's being correct. So we have a full explanation of which sentences in her language she accepts, which she rejects, and which she neither accepts nor rejects, but the explanation is consistent with different accounts of what her words mean.[12]

I now agree that Horwich extends his theory in a way that resolves the kind of indeterminacy that I had been claiming. His extended theory is not just that the meaning of a term is its basic acceptance property, the property that explains its use. It is also that for cases in which such explanations give out, the meaning is indeterminate. We could put what I'm now saying as follows: Understand the *basic acceptance property* of a word in a person's language as the property that figures in explaining all that is systematic in her usage. Then for applications where her usage isn't systematic, Horwich adds, the meaning is indeterminate. (One could alternatively include in the "basic acceptance property" whatever properties explain the lack of systematicity in certain applications. It is then a further stipulation that in these applications the meaning is indeterminate—but this extension is, I recognize, quite natural.)

Whether Horwich's naturalistic theory of meaning is successful requires further investigation, as I'm sure he agrees. Is it really the case that each word has a basic acceptance property, a property that is the explanatory basis for its overall use? This, as I get the idea, is

12 In a note, Horwich speaks of empirically equivalent explanations of the phenomena that invoke different dispositions. "We'd again get the result that there's no objective answer to the question of which particular disposition is the explanatory basis for its overall use." But in my example, the explanatory basis for the Newtonian's usage is clear, whereas the meaning of her word 'mass' might be controversial. Again, the question of what her word means isn't just how to explain her usage—as Horwich assumes it is. It is how to engage her thinking.

a single property of the word as used that, for each instance of its use, joins other factors not peculiar to that word to explain the use in this instance. It's a bold hypothesis that explanations of the use of a word can always take such a form; the hypothesis may well be correct, but whether it is requires investigation I'm not in a position to pursue.

Suppose, though, that this hypothesis is right. That will be a great empirical triumph. Where will that leave the concept of meaning? It would leave us, I have been saying, with two different concepts, one purely naturalistic and one with normative ingredients. Why, though, wouldn't this way of proceeding extend to words more generally, giving us a plethora of normative meanings for every naturalistic term? What's special about meaning in this regard? It is, I say, the special systematicity of the tie of what sentences mean to when to accept them.

I'll end in a way that isn't thoroughly concessive: Horwich's own theory of meaning, I'll argue, supports my metatheory of meaning, not his. What do people actually mean by 'meaning'? One aspect of their usage we must explain is how they will respond if they think that, in an instance, Horwich meaning and Gibbard meaning come apart. Suppose they think that what causally best explains a speaker's usage of the word 'rectangle' is a prototype with unequal sides, but for the case of being that speaker, they plan to include squares, for the less complicated systematicity this brings. Which will they think is that speaker's meaning? Won't they go by the meaning that is tied to what to accept in and for the speaker's circumstance? If so, what causally best explains their usage is my metatheory, not Horwich's. So Horwich's theory and metatheory of meaning support the normative interpretation of the term.

Still, I would agree with this: Suppose Horwich's hypothesis is correct, his hypothesis that each word has a basic acceptance property, a property that joins other factors in explaining a subject's acceptance and rejection of sentences with that word, factors not tied to that word in particular. Then there will be much to be said for revising our language to mean by our word 'meaning' what Horwich claims it means. The revised meaning will suit discussions where it is taken for granted, in a way that puts it beyond needing discussion, that one's own sentences should be accepted according to their Horwich meaning.

All this leaves many of the fascinating things that Horwich says unaddressed, but I'll break off here. Again, I thank Paul Horwich profusely for his remarkably detailed scrutiny of my metatheory of meaning.

Jamie Dreier

My theory of normative judgment aspires to say, among other things, what it is to judge that torture is wrong. It doesn't say whether torture is indeed wrong. Dreier says that according to expressivism, this is the best we can get. I think he's wording this imprecisely: Expressivism doesn't say that this is the best we can get; it says that this is the best we can get purely drawing on a correct theory of what words like 'wrong' mean. As decent human beings, though, we can get something better: the conclusion that torture is greviously wrong. Few if

any expressivists will confine themselves in their beliefs to expressivism. I take it that Dreier doesn't intend his words to contradict what I am saying, but if he could be read as doing so, then as so read, it is worth disclaiming.

One way to see how my metanormative picture works is to see how it might combine with a particular normative theory. Consider universalistic hedonistic consequentialism as a theory of what being the morally right thing to do consists in, and consider a theory of what degrees of belief are rational that takes this form: it starts with a specific epistemic prior probability function ρ_0, a specification of what degrees of belief it is rational to have prior to any experience; given any total body E of evidence, it says that the rational degree of belief to have in any naturalistic thought S is $\rho_0(S/E)$. On my view, there is no such thing as a peculiarly normative property; there are just properties. The same property can be signified in normative terms and in naturalistic terms. Thus the property of maximizing prospective total happiness in the world can be signified in at least two ways: causally, as I have just done, or normatively as the property of being the right thing to do. According to the normative theory we are considering, the property of being the right thing to do is identical to the property of maximizing prospective total happiness—or as I shall say, being *maxihedonic*. This claim—that the property of being the right thing to do is identical to the property of being maxihedonic—is a normative claim. Asserting it amounts to issuing this imperative: On any possible occasion for action, do whatever is maxihedonic! (These formulations ignore the possibility of ties.) To believe that on any possible occasion for action an act is right iff it is the maxihedonic act is to permit oneself, for any possible occasion for action, all and only the acts that are maxihedonic.

Thus according to this normative theory, the property of being maxihedonic is identical to the property of being a right thing to do, of being okay to do. Whether an act open to an agent on an occasion is maxihedonic is a matter of its expected value as calculated with the probabilities $\rho_0(S/E)$ for every relevant naturalistic thought S.

The solution in Dreier's Sec. 5—the one he says is "the *only* solution to the problem" as he has posed it—is indeed what I maintain (448). There are no peculiarly normative facts, I insist; what there are is normative ways of signifying facts. I have never asked whether "meaning is normative", except as a rough way of indicating a contention that needs to be reformulated if it is to be plausible and intelligible. The claim I have tried to elucidate is that a *concept* is normative, the concept MEANING. What a word means can be disputed even when there is complete agreement on the facts as signified purely in naturalistic terms.

Dreier considers an example that I use, the issue that Kripke's Wittgenstein raises as to whether meaning is individualistic or social. Dreier doesn't think this issue is genuine; there's individualistic meaning and social meaning, and we just have to decide which to talk about. I claim, on the contrary, that there is a genuine issue, and that this issue is normative. To begin with, what a person means by a word, I say, is a matter of what sentences with that word to accept in that person's language if one is that person. (Not what sentences the person does accept, but what sentences *to* accept—what sentences one ought to accept.) Kripke's

thinker can therefore pose the issue of what the person means by '+' as follows, for the case of being that person (I put this in my own words.):

> Suppose my linguistic community means QUUS by '+'. Then what is 58 + 67? Ought I, under that supposition, to accept my sentence '58 + 67 = 5'? Or ought I, under that supposition, to accept instead my sentence '58 + 67 = 125'? The first sentence, '58 + 67 = 5', is what's correct under the interpretation of my words that optimally balances simplicity with closeness of fit to the linguistic dispositions of my community. The second sentence, '58 + 67 = 125', is what's correct under the interpretation of my words that optimally balances simplicity with closeness of fit to my own individual linguistic dispositions. Which of these sentences ought I, under that supposition, to accept?

Note that for simplicity, I am considering just two possible theories of what the person means by her words, even though there are many more possible theories. And the example is far from ideal; we should really focus on elementary steps like, when you carry in arithmetic, writing the ten's digit at the top of the next column. But suppose we do apply individualistic and social theories of meaning to elementary operations. Both these theories, in some sense, balance simplicity with closeness of fit to some set of linguistic dispositions—in one case, the linguistic dispositions of the community, and in the other, those of the individual.

So in Kripke's example or a better one, there's something genuine at issue between these alternatives. At issue is, for the case of being that person, which sentences in one's own language to accept. Perhaps, to be sure, one should accept neither, but these are distinct, intelligible answers to the question of what the person means.

Dreier, as I say, believes the issue of what the person means to be bogus. He then works to develop a different issue as genuine—not whether the concepts MEANS is normative, but whether meaning itself is normative. I'm not entirely clear how he ends up, but I read him as concluding there's no such issue to be found; there's no plausible sense in which meaning itself might be normative. True, as I would be the first to say, what people mean by their words is somehow a matter of how they are disposed to respond to thinking couched in those words, a matter of the linguistic rules that they implicitly accept. But this isn't a genuinely normative matter, he maintains. There's a difference between purported normativity, "a kind that is gripping only once one accepts a given framework," and "genuine normativity, which has its force in a more categorical way" (453). I myself am doubtful that meaning—meaning itself—fits either pattern he uses to illustrate the difference. Put stipulated meaning to the side, and think of meaning in a language that a person thinks in, or meaning in a language that the person converses in unselfconsciously. Yes, say I as observer-theoretician, what a word or a syntactic device means is somehow a matter of spontaneous linguistic reactions of the people involved. Is meaning therefore normatively constituted? It isn't like being a bishop, Dreier says, and I agree, though perhaps not for his reason. A piece's being a bishop, I would say, depends on players' thinking of themselves as playing chess, or presenting

themselves as playing chess. Thinking, or conversing at its least self-conscious, doesn't hinge on anything analogous. It doesn't ordinarily depend on regarding oneself as speaking in English or presenting oneself as speaking in English. Questions of what variant of English one is thinking in or presenting oneself as speaking in can arise, but that will be to try to sort out matters that have gone awry.

Alternatively, is meaning like theft? Not in a sense that concerns theft itself, as opposed to the concept of theft. Theft, says Dreier, "is normatively constituted; calling an act a theft is making a normative assertion" (452). This brings us back, though, to questions of signification. Calling something the meaning of a term is likewise making a normative assertion—that's what I've been experimenting with arguing. But this isn't a claim about meaning itself, but the concept of meaning.

I haven't, then, managed to make sense of Dreier's question of whether meaning itself is normative—and I'm not clear whether Dreier thinks there's sense to be made of it. In any case, I maintain that when we disagree as to what a person means by a term, our dispute needn't "go away in the face of full explication of our respective positions" (453). It may still, I conclude, qualify as normative.

6. Consequentialism

20. Connie Rosati (University of Arizona), "Gibbard on Reconciling Our Aims"
21. David Braddon-Mitchell (University of Sydney), "Freedom and Binding Consequentialism"

David Braddon-Mitchell

I seem to recall that I blundered in trying to cut my Berkeley Tanner Lectures down to size. I can't confirm this from what I find on my computer, but my shaky memory is that at some point or other, trying to conserve length, I cut out a paragraph saying that I didn't have direct utilitarianism in mind for what I was supporting. Rather I had in mind some kind of indirect utilitarianism in the spirit of Brandt's rule-utilitarianism. As a consequence of this blunder, David Braddon-Mitchell, Connie Rosati, and others have read me as advocating act-utilitarianism. That wasn't what I intended. But Braddon-Mitchell argues that I don't need to alter this, that understood the right way, act-utilitarianism can be a plausible moral theory. I'll be inquiring whether this is right.

Even though I kicked myself over the cut, I didn't believe that some version of indirect utilitarianism can be made precisely right. No formulation of indirect utilitarianism avoids dilemmas, I thought; some situations just do make for dilemmas. I didn't think that the dilemmas refute utilitarianism; they are ones that arise for any comprehensive moral view. Consider the voter's dilemma: if few people who would support a more equitable income distribution vote for a plan that would promote it, one may accomplish the most good by pursuing other goals. But if everyone voted for the plan, let's suppose, far more

good would ensue. More generally, when it would be best for everyone to join in a cooperative scheme but not everyone does, a lonely stance of being the only person who joins may not be what does the most good given that the others won't join in. A simple form of rule-utilitarianism says to play one's part anyway, even if you are the only one. But if you can't bring the others along, why should you take this lonely stance?

If Braddon-Mitchell succeeds in making act-utilitarianism acceptable without resorting to some sort of rule-utilitarianism, I ought of course very much to welcome that. He speaks of limiting one's alternatives, and as he recognizes, the big worry about his proposal must be what constitutes doing this. Direct utilitarianism for acting says that an act is right just in case, among the things one can do, it produces maximally good consequences. That's for being right in the objective sense; for the subjective sense, it's that an act is right just in case its prospects for good consequences are maximal in light of the agent's information. To apply either of these, we have to say what constitutes an act's being something the agent *can* do, and so what constitutes restricting one's options. Like Braddon-Mitchell, I maintain that, to the extent that the notion of what one can do has any precise sense, some version of compatibilism obtains. That I'm disposed not to do a thing isn't enough to establish that I can't do it. That one can do a thing means something like that if one tried one would succeed. Clearly there are ways of affecting what one will be able to do on a later occasion; burning one's bridges behind one is a stock example, ensuring that one won't be able to retreat. Do the kinds of things that Braddon-Mitchell has in mind, though, qualify as restricting one's options?

Whether they do or not, I agree with Braddon-Mitchell that often the crucial moral question is not what to do right now, but what dispositions to instill in oneself. I agree with him too that what matters morally is one's extended sequences of acts, and not the moral status of the individual acts that make them up. It's the extended sequence, I agree as well, that should be the focus of moral evaluation and criticism. I'm reminded of getting to know Tom Schwartz when I visited at Stanford in the fall of 1972. He told of Dr. Krankheit who never did wrong. He was supposed to anaesthetize his patient and operate, but did neither. "I was right not to anaesthetize him, because I wasn't going to operate anyway. And I was right not to operate, because the patient wasn't anaesthetized." His fault, Braddon-Mitchell might say, lies not in an act but a disposition: he was disposed not to operate even if the patient had been properly anaesthetized. Or better, as Braddon-Mitchell suggests, we can say that the fault consisted in performing a wrong temporally extended action, the one that began with not anaesthetizing.

Braddon-Mitchell advocates transforming yourself in the way that makes for best outcomes. He could maintain further that what's right is whatever one then will do, whether or not that is what act-utilitarianism prescribes. Whether this indeed fits what Braddon-Mitchell says depends still on what constitutes restricting one's options. But we could argue that what matters is not this but which moral theory captures the substance of being morally right, whether or not it counts as act-utilitarianism. (Perhaps it matters too how the moral theory is framed and motivated.)

Will transforming one's motivations help with the voter's quandary? Even if I have unlimited capacity to bind myself to a course of action, the dilemma seems to remain: What if the others don't bind themselves to vote? Shall I bind myself, and so transform myself into someone who will vote whatever others do. That seems futile. Of shall I transform myself into someone who will vote only if enough others do? Everyone has abided by this prescription if no one votes.

Perhaps, though, self-binding can help with this too. What would be needed is a self-binding that's conditional: I bind myself to play my part if enough others likewise bind themselves. I hope that something along this line can be made to work, though it probably can't resolve all the moral dilemmas we face. It might well work if one could bind oneself to cooperate conditionally on everyone's being in fact so bound—but making action conditional depends not just on the condition's being met, but on everyone's knowing that it is met, and so we face an assurance problem. Whether the kind of self-binding that Braddon-Mitchell proposes can solve this problem needs further investigation. The main thing to say, familiarly, is that situations like this are a chief sort of thing that makes institutions necessary as mechanisms of coordination. But we need kinds of coordination as well that couldn't be formally instituted. Also, among the things that Braddon-Mitchell stresses, self-binding can solidify one's adherence to beneficial institutions and practices, where absent self-binding, even if the best thing to do is to act in conformity, the motivations are likely not to be strong enough.

It would be amazing if Braddon-Mitchell's proposal accomplished everything, and whether or not it does, it accomplishes a lot, as he indicates. I'm not sure whether what I've been saying vindicates Braddon-Mitchell with act-utilitarianism as he shows it to be, or supports a different moral standard. I still find all this puzzling.[13]

Connie Rosati

We can characterize my substantive moral views on morality and justice, Rosati suggests, "as *methodologically contractarian*, but substantively utilitarian in its upshot" (469). She is right—and this, I maintain, is essentially how Rawls's own views work. Rosati is skeptical that what I draw from Harsanyi in my lecture "need be convincing to those not already persuaded of utilitarianism", but if I'm right, if we should be convinced by Rawls, we should likewise find Harsanyi convincing (472). Rosati also, though, doubts some of the arguments that Rawls takes to support an alternative to a utilitarian approach, an alternative based on fair reciprocity with society viewed a cooperative scheme for mutual advantage. Let me start, though, with Rawls and utilitarianism.

Rawls highlights his aim as to devise an appealing alternative to utilitarianism, but the utilitarianism he rejects isn't indirect utilitarianism of the kind I would have advocated.

13 My PhD thesis *Utilitarianisms and Coordination* was on issues like this, and was later published, obscurely, by Garland Publishing. I thought that this publication would make it available to anyone who really wanted it, but it seems to be thoroughly out of print.

Indeed Rawls explicitly didn't argue against a utilitarianism that's indirect; he says that it isn't what he means by "utilitarianism". "Utilitarianism, as I have defined it," he specifies, "is the view that the principle of utility is the correct principle for society's public conception of justice"[14]. For what the valid principles of justice are and why, he also explicitly doesn't reject the possibility that his Two Principles of Justice are what an indirect utilitarianism would endorse, and endorse for the very reasons that Rawls puts forth.

As Rosati suggests for me, Rawls is of course methodologically contractarian. Where he differs sharply from traditional utilitarianism is on the motivation to be just. Traditional utilitarianism appealed to sympathy, whereas Rawls dismisses sympathy as too weak a motive to render us just. Instead, he invokes reciprocity in a fair scheme of social cooperation. I think that Rawls was quite right on this, and this is one of the deepest things he taught me. In addition, Rawls rejects a widespread rationale for utilitarianism, direct or indirect. Frankena said at one point that morality is made for humanity, not humanity for morality; that might suggest that the basic point of morality is to foster the general good for people—as Sidgwick, Moore, and many other philosophers have assumed. As I have been saying, Rawls took the basic rationale to be fair reciprocity.

This yields an answer to the voter's quandary and like moral quandaries: in those cases, there's nothing to reciprocate, and you don't yourself owe compliance to a moral standard. This differentiates his view from utilitarianism, which says to produce as much good as you can even if no one else will join in. What, though, if others somewhat will and somewhat won't comply with a scheme? Rawls addresses this as the case of "partial compliance": he was rightly greatly impressed with Martin Luther King, and supported civil disobedience for cases where others comply with demands of justice only partially.

Rawls's "Original Position" is similar in important ways to the criterion Harsanyi had embraced for moral preferences: that one not know who one is in society. Rawls works hard to distinguish his Original Position from Harsanyi's basis, among other things, rejecting probabilistic reasoning. First, though, a word on the history of Rawls's ideas. It's clear to me that when Rawls devised his Original Position, he wasn't aware of Harsanyi. The name 'Harsanyi' doesn't occur in his 1958 article "Justice as Fairness" nor in his 1967 article "Distributive Justice". Clearly, though, Rawls did become vividly aware of Harsanyi substantially before *A Theory of Justice* appeared in 1971. I know this because in my last year in graduate school, fall 1968, Kenneth Arrow came to Harvard. and I had the amazing opportunity to take a joint seminar that Rawls and Arrow gave with a brilliant young economist from India I had never heard of, Amartya Sen. Of course in that seminar, there was copious discussion of Harsanyi. Rawls was thus immersed in discussions of Harsanyi in the three years before *A Theory of Justice* was published—and in his Preface he acknowledges correspondence with Harsanyi.

Rawls differs from traditional utilitarianism in important ways: on the rationale for justice and the motive to be just. But on which principles of justice are valid and why, Rawls and

14 Rawls, John (1999) *A Theory of Justice*, 158.

an indirect utilitarian may not differ, at least for cases of full compliance. Look to how Rawls supports the Difference Principle as giving what distributive justice consists in—maximin in the sense of making the worst off best off, or more specifically, making the economic prospects of those in the worst starting positions as good as possible. A party to the Original Position, he says, in choosing a public conception of economic justice, "cares very little, if anything, for what he might gain above the minimum stipend that he can, in fact be sure of by following the maximin rule" (134). On the other hand, "the rejected alternatives have outcomes that one can hardly accept" (134). In economists' terms, this amounts to saying that one's expected utility as a function of income and wealth is sharply declining. This, utilitarians have long recognized, gives strong support for economic equality. If Rawls is right on what parties to the Original Position care about, then they choose what an indirect utilitarianism favors: a high degree of equality in economic prospects.

Still, if one could raise the prospects of the best off at small cost to the worst off, won't that increase average utility? Perhaps, but Rawls has reasons not to favor that, reasons that, if they are correct as Rawls sets things up, an indirect utilitarian will share. Remember first that Rawls is considering candidates to serve as a "a conception of right"—that is to say, "a set of principles, general in form and universal in application, that is to be publicly recognized as a final court of appeal for ordering the conflicting claims of moral persons" (117). Crucial, then, is public recognition; parties to the Original Position look to how public recognition of candidate sets of principles might affect how they fare in society. Thus the principles can't be too complicated for public recognition to be effective, and the parties are to consider, among other things, what makes for stability: for the conception to generate its own support. Sometimes he does say that the principles are chosen in view of the consequences of everyone's complying, but even if initially by fiat everyone complies, parties must look to the long run. Stability is a matter of whether, if the principles are accepted and implemented initially, their effective compliance will last. If Rawls is right that his Two Principles will be chosen by the parties because they are stable in this way and the alternatives aren't, an indirect utilitarian too will find this a strong reason to choose the Two Principles.

Remember too that Rawls doesn't go for strict maximin. As concerns economic distribution, he favors maximizing the prospects of those in the broadly worst starting positions—perhaps, for example, those who start out with half or less of the median income. For all Rawls purports to establish, this might be realized in a fully socialistic command economy or, alternatively, in a mitigated market system. It might be that compared to a command economy, a system of economic incentives with an income floor and strongly progressive taxation would tend to benefit all. Alternative systems, remember, are to be evaluated against each other by the life prospects of those in the worst starting positions. (In 1970, many of us thought that although our economic system was far from satisfying the difference principle, things were heading in the right direction. How wrong we were on that! But on a world scale, extreme poverty diminished to a degree that many of us likewise didn't expect—before the coronavirus pandemic, at least.)

Rosati says of me, "We also, he says, can't use something like Rawls's 'primary goods' in our moral thinking; Gibbard maintains, for reasons that he doesn't provide in the lectures, that Rawls did not supply 'an adequate, defensible rationale for this solution'" (474). So I should try to explain what I had in mind—though I won't be able to analyze the issues at all fully. Rawls thought that with his qualified maximin, he could avoid interpersonal comparisons of utility by reckoning prospects in terms of income and wealth rather than anything like utility. But that won't work, we should be able to see. In any economic system that includes markets, how one ends up in life depends to a considerable degree on luck. There's a social decision to be made, then, about how amply to insure people against bad economic luck. Suppose, for instance, that of those who start life among the worst off, half find great success and end up living in commodious prosperity, and half end up living in destitution. The monetary average of this would allow a life that's not so bad. In terms of what prospects to prefer for oneself, though, the chance of living prosperously doesn't come close to making up for the risk of destitution. We need to think in terms of what it makes sense to want for oneself and how strongly, not just in terms of how, for those who start out among the worst off, the chancy prospects for income and wealth come out as a monetary average. Doing this amounts to thinking in terms of something broadly like utility, and not just the primary goods of income and wealth.

Apart from trying to replace utility by money in reckoning prospects for those in the worst-off starting positions, I conclude that for the case of full compliance, Rawls has no disagreement with an indirect utilitarian as to what economic justice demands. Indeed, Rawls acknowledges explicitly the possibility that public acceptance of the two principles of justice would maximize average utility. "This may conceivably be the case." He responds that this sort of indirect view isn't what he means by 'utilitarianism' (158). It thus isn't indirect utilitarianism that he is rejecting.

Once we bring in something like utility in deciding how much social insurance against bad economic luck is adequate and how much would be excessive, I think that Rawls's arguments for his moderated maximin are powerful and well worth considering. Holly Smith (then writing as Holly Goldman) has an excellent discussion of all this; she and I had illuminating discussions as she was writing her piece. I should confess how insecure I had been earlier as an Assistant Professor at Chicago and an old Rawls student: Warner Wick as editor of the journal *Ethics* had solicited a review of *A Theory of Justice* from me, and I got to the point of presenting my draft in a lecture at Princeton. But I never dared to let go of it for publication. (I hope that this confession can be of help to philosophers at the beginning of their careers; you can imagine how much I sympathize with anyone responding, as I was, to early career terrors.)

Turn, then, to some of Rosati's criticisms. "As I shall explain," she says, "Gibbard's account doesn't seem to capture either the phenomenology or the normativity of moral inquiry" (460). In different ways, though, it may capture both. The phenomenology will arise within the ethos that is publicly accepted; it is the phenomenology that parties to the Original

Position choose to foster. As for the normativity, it is what I try to explain in the first of my Berkeley Tanner Lectures: The normative questions are what to do and how to feel about things people do or might do. I engage in evolutionary speculations, rooting what I say in selection pressures favoring coordination.

All this, Rosati says, "does not yet explain why the moral emotions would be guided by judgments of fairness, rather than by judgments of how one's feelings and resultant actions might affect one's social status," whether, say, one would be thought a jerk to have those feelings (462). Once we distinguish sharply how it makes sense to feel about things from how it makes sense to *want* to feel about them, this question may amount to why we are concerned substantially with how it makes sense to feel. That's an important challenge, and I can't respond in any way that's definitive. First, though, let me observe that norm acceptance is at least one way to coordinate—and it is a way we do have, if I am right. Would evolutionary selection pressures favor this way over alternatives? Would they favor coordinating through converging in the norms we accept and being guided by those norms? That's a bigger question than I know how to address, and I hope that others can delve into it. Rosati is correct, I take it, that one of the ways we are is to be highly responsive to cues concerning social status. So our question can't be why norm acceptance is the only factor we respond to; it isn't. It's more why selection pressures might have favored this way of tending toward consensus and being responsive to it. As I have indicated, I'm convinced that these questions demand profound inquiry.

In my Berkeley Tanner Lectures, I pay serious heed to Scanlon's skepticism that there's a concept of a person's good or benefit that can play the all the role philosophers have ascribed to it. Rosati responds that the question of whether there's such a thing as a person's good answers itself directly (476). For this

> surely is what we would expect, given the scientific picture of us biological creatures. Reflect on the nurturing required for human development and the sustenance required for continued life and activity. Consult your own experiences of pleasure and pain. Consider the importance in your own life of loving, interpersonal connections and gratifying pursuits.

Scanlon's challenge, though, is whether the notion of a person's good picks out something definite, something that, among other things, can be added up for different people in the sort of way that utilitarianism demands. This challenge surely needs answering. All the things that Rosati adduces in this passage do pertain to a person and are well worth wanting. But the history of twentieth century ethics is full of disputes as to what intrinsic good consists in, and clearly these debates would carry over to the good of each person, to the question of what it is to be to a person's benefit. Pleasure and pain aren't in dispute as parts of a person's good or ill, I agree. But as for gratifying pursuits, what if the gratification rests on a mistake? One is gratified, say, in pursuit of immortality, or in pursuit of confirmation of a thrilling discovery one thinks one has made; what if in fact one is mistaken? As for life and activity,

life is good except when it isn't. Activity can stem from terror; does such activity benefit one in itself, or only if it wasn't futile, or what? Does the benefit of loving ties with others depend in itself on whether the others' apparent love is genuine? If one ought for its own sake to want to possess integrity, is having integrity to one's good? If you successfully urge integrity on me, is that for my benefit? If I help your children after your death, is that to your benefit? These are questions philosophers have disputed; is it really obvious what's at issue in these disputes? Scanlon doubts that these questions have genuine content, and his doubts seems to me to be worth addressing.

Biology by itself doesn't answer these questions. Biology tells us that the genetic "plan" for the human psyche was selected for reproduction in ancestral conditions. But bare reproduction isn't, I would have thought, a benefit in itself. Perhaps it should count as part of a person's good if the descendants live fulfilling lives, but whether it should isn't anything that biology by itself addresses.

Rosati suggests that general capacities to reason correctly will explain how we might tend to come to correct moral conclusions (464).

> Our evolved capacities for gaining knowledge about our environment are such that, once we have them, we have acquired general abilities to learn and reason and so to learn much about the world that may have no particular bearing on gene proliferation, at least not in any direct way, such as abstract mathematics and theoretical physics. The same abilities that explain how we can acquire the latter sort of knowledge might well explain how we can acquire moral knowledge.

I wouldn't think that science can explain any tendency to get things right across the whole domain of things we think about. Rather, we can get somewhere in explaining various particular kinds of knowledge: knowledge of logic, empirical knowledge, and various other particular kinds of knowledge. Our logical abilities can explain why if we get our axioms right, our derived conclusions will be right. How, though, do we manage to distinguish valid axioms from spurious ones? Philosophy of science addresses questions of how we can gain empirical knowledge through observation, induction, hypothesis testing, and the like. How, though, will these abilities enable us to get right non-empirical starting points for ethical thinking. As for mathematics, I concede that Scanlon has powerful arguments that some of the bases of good mathematical thinking share characteristics with good ethical thinking. Perhaps, then, what Rosati is suggesting can be vindicated across a domain that includes the bases of mathematics and of ethics. Insofar as mathematical thinking rests on requirements of logical consistency, those requirements won't by themselves resolve ethical issues, since terrible putative systems of ethics can be made logically consistent. To my mind, it's a genuine puzzle how we can tend to get foundational matters in mathematics right when doing so requires more than logic. If the answer is the same for mathematics and for ethics, this isn't, I think, an answer that we have managed to give and establish.

In short, I myself maintain that moral questions are questions of what to do and how to feel about doings. Rosati might possibly be suggesting that moral questions don't have this sort of distinctive nature, that what characterizes them is the realm of properties they concern, the moral realm. Over such a view, my program has the advantage that it offers an explanation of why moral questions matter—I don't see that the more traditional view of things that Rosati may be favoring has this explanatory virtue. Join to this the way that, as Nye lays out, my kind of program can explain why in general not to do things that are morally wrong, and I conclude that my general way of doing things has the balance of advantage.

Rosati says much more that demands inquiry, but I should leave my response at that. So far as I know, Rosati is the only writer to address my Berkeley Tanner lectures at length, and I hope that others will take up the issues she raises.

This concludes my responses to the commentaries. To all the commentators I reiterate my immense gratitude. It is a high privilege that superb philosophers like these scrutinize my efforts so assiduously and insightfully. I have tried—perhaps in vain—not to abuse the privilege.