OXFORD

Oxford Studies in Metaethics Volume 3

OXFORD STUDIES IN METAETHICS

This page intentionally left blank

Oxford Studies in Metaethics

VOLUME 3

Edited by

RUSS SHAFER-LANDAU

OXFORD UNIVERSITY PRESS · OXFORD

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford OX2 6DP

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide in

Oxford New York

Auckland Cape Town Dar es Salaam Hong Kong Karachi
Kuala Lumpur Madrid Melbourne Mexico City Nairobi
New Delhi Shanghai Taipei Toronto

With offices in

Argentina Austria Brazil Chile Czech Republic France Greece
Guatemala Hungary Italy Japan Poland Portugal Singapore
South Korea Switzerland Thailand Turkey Ukraine Vietnam

Oxford is a registered trade mark of Oxford University Press
in the UK and in certain other countries

Published in the United States
by Oxford University Press Inc., New York

First published 2008

British Library Cataloguing in Publication Data
Data available

Library of Congress Cataloging in Publication Data
Data available

Typeset by Laserwords Private Limited, Chennai, India
Printed in Great Britain
on acid-free paper by
Biddles Ltd., King's Lynn, Norfolk

ISBN 978–0–19–954206–2
ISBN 978–0–19–954207–9 (Pbk)

1 3 5 7 9 10 8 6 4 2

Contents

Notes on Contributors

Thomas Baldwin is Professor of Philosophy, York University

David O. Brink is Professor of Philosophy, University of California, San Diego

William J. FitzPatrick is Associate Professor of Philosophy, Virginia Polytechnic Institute and State University

Matthew Hanser is Associate Professor of Philosophy, University of California, Santa Barbara

Chris Heathwood is Assistant Professor of Philosophy, University of Colorado, Boulder

Frank Jackson is Regular Visiting Professor of Philosophy at Princeton University, and Fractional Research Professor at La Trobe University

Sarah McGrath is Assistant Professor of Philosophy, Princeton University

Geoffrey Sayre-McCord is Professor of Philosophy, University of North Carolina, Chapel Hill

Caj Strandberg is Lecturer in Practical Philosophy, Gothenburg University Lund University

Sharon Street is Assistant Professor of Philosophy, New York University

Nick Zangwill is Professor of Philosophy, Durham University

Introduction

Russ Shafer-Landau

Oxford Studies in Metaethics is designed to collect, on an annual basis, some of the best new work being done in the field of metaethics. I'm very pleased to be able to present this third volume, one that has managed so successfully to fulfill the aims envisioned for the series.

David Brink's contribution, "The Significance of Desire," opens the collection. Brink offers an extended critical examination of a variety of *desiderative* conceptions of practical reason and personal welfare. These conceptions are each based on the idea that our actual or hypothetical desires play a central role in determining what we have reason to do, and where our own good lies. Brink is not sanguine about the prospects of these theories, a pessimism shared by our second author, Chris Heathwood. In "Fitting Attitudes and Welfare," Heathwood argues directly against what he calls *fitting-attitude* analyses of personal welfare, according to which one's welfare is identical to what we have reason to want for our own sake. He claims that anyone committed to a fitting-attitude analysis of intrinsic value should be committed to a similar analysis of personal welfare, and so uses his rejection of the latter to undermine the former.

Up next: two articles about the metaethical relevance of moral disagreement. In "The Argument from the Persistence of Moral Disagreement," Frank Jackson launches a sustained critique of a classic metaethical argument, one that begins by noting the breadth and intractability of moral disagreement, and concludes by embracing an expressivist analysis of moral discourse. Jackson thinks that the argument fails, because the conception of disagreement that the expressivist must accept leaves cognitivists equally able to diagnose the cause and frequency of moral disagreement.

The focus remains on moral disagreement in Sarah McGrath's contribution, "Moral Disagreement and Moral Expertise." However, McGrath is not so much concerned with moral metaphysics, or semantics, as she is with the epistemic consequences of finding oneself possessed of seriously controversial moral beliefs. Taking a cue from Sidgwick, she argues that if you realize that your beliefs are disputed, and have no more reason to suspect your interlocutor of error than yourself, then you ought to suspend judgment about the contested matter. Since we find ourselves in many such situations when it comes to our moral beliefs, the persistence of moral disagreement under these conditions yields skeptical results.

Nick Zangwill next offers us the first of a pair of articles on the ways in which the moral depends on the nonmoral. Zangwill distinguishes supervenience relations, which have been the subject of much discussion in metaethics over the past three decades, from dependence relations, which isolate just the nonmoral features that are responsible for the instantiation of a moral property. Supervenience relations will include all that is relevant to the instantiation of a property, and can do this without isolating those features on which a property's instantiation depends. Zangwill seeks to explicate this latter notion, and concludes that we may well need a *sui generis* conception of it to discover precisely how moral properties depend on their non-moral bases.

Caj Strandberg's paper, "Particularism and Supervenience," very nicely complements Zangwill's contribution. Strandberg is concerned to defend traditional conceptions of supervenience against the sort of concerns raised by Zangwill and, earlier, by Jonathan Dancy in work published elsewhere. The deep questions at the heart of supervenience discussions—just how is the moral related to the nonmoral; what is the sense in which something has a certain moral feature *just because* it has a nonmoral one?—can, says Strandberg, be answered by reference to familiar conceptions of the supervenience relations. This despite the challenge levelled by particularists to the effect that nonmoral properties are always of variable moral relevance.

William FitzPatrick's offering, "Robust Realism, Non-Naturalism, and Normativity," is an ambitious exploration of the merits of ethical realism. He finds such a view highly congenial, and offers a battery of considerations that explain the attractions of such a position. He does not take himself to have refuted anti-realists, but rather to have identified the features that have persuaded many philosophers to join the realist ranks. There is deep division within those ranks, however, between naturalists, who seek to make morality of a piece with the natural sciences, and non-naturalists, who resist this assimilation. FitzPatrick is firmly on the side of the non-naturalists, and argues that some of his fellow travelers (myself included) have not gone far enough in resisting naturalistic temptations.

Sharon Street is no ethical realist, and her "Constructivism about Reasons" presents a wide-ranging and provocative defense of the titular theme. As she sees it, there are no normative truths that hold independently of our evaluative attitudes. Ultimately, things are valuable, and provide us with reasons to act, only because we invest them with significance. Street rightly notes that constructivism has gained increasing attention in the metaethical literature, but, surprisingly, has received only a very few direct, extended elaborations. She aims to remedy this gap in her wide-ranging and important essay.

In "Rawls and Moral Psychology," Thomas Baldwin focuses on a set of issues that have been relatively little discussed in John Rawls's work—namely, his account of moral psychology, and its relations to other aspects of his work. Rawls never managed to fully articulate his ideas in this area; the reader can find his intringuing remarks at various places within his corpus. Baldwin does us the service of drawing our attention to these scattered writings, and determining whether there is a coherent view to be extrapolated from them. He thinks that there is, and proceeds to explore how this view is interestingly related to some of the major philosophical themes of Rawls's work.

Matthew Hanser next gives us his paper, "Actions, Acting, and Acting Well." As he rightly notes, the nature of moral judgments has long been a central topic in metaethics. Hanser doesn't propose to give us yet another take on the nature of such judgments; rather, he asks a simple but largely ignored question: what are moral judgments *of*? What are they about? He seeks to show that the easy answer—that they concern actions, or types of action—is mistaken. Blending action theory and ethical theory, Hanser treats us to a nuanced exploration of the subject matters of our ethical judgments.

The volume concludes with Geoffrey Sayre-McCord's contribution, "Hume on Practical Morality and Inert Reason." Sayre-McCord claims that the standard readings of Hume's views of practical reason are mistaken, and, in particular, that the motivational internalism and noncognitivism often attributed to Hume are impositions not warranted by his actual writings. The rationalism that was his main target is, argues Sayre-McCord, genuinely vulnerable to the arguments that Hume mustered against them. The reconstructed arguments that Sayre-McCord places at the tip of Hume's quill are deep and powerful. Whether they are enough to undermine the rationalism that Hume so opposed is a matter best left for the reader's consideration.

Most of the articles included in this volume are significantly revised versions of papers given at the third annual Metaethics Workshop in Madison, Wisconsin, in September 2006. My thanks to the University of Wisconsin Anonymous Fund, whose generosity underwrote the costs associated with the conference. I'd like to extend my sincere thanks as well to the eminent philosophers who comprised the workshop program selection committee, and so served as de facto referees for the present collection: Michael Ridge, Connie Rosati, David Sobel, Nicholas Sturgeon, and Mark Timmons. Mark van Roojen, and another philosopher who prefers anonymity, were commissioned by the Press as external referees for this volume, and offered thorough, constructive, incisive criticism of the first order. The excellence of the present volume reflects the hard work undertaken by each of these fine philosophers, and I'm very grateful to have had their significant assistance. Last, but not least, my thanks, as always, to Peter Momtchiloff, who has served as such an excellent companion on this venture.

1

The Significance of Desire

David O. Brink

There is a venerable tradition of treating practical reason and theories of the good, especially the agent's own good, as grounded ultimately in facts about the responses that an agent does or would have to various situations and options upon suitable reflection. These are *response-dependent* conceptions of practical reason and the good. An important form of response-dependence is a *reductive* form that aims to reduce facts about reasons and the good to facts about desire. Such *desiderative* conceptions of response-dependence treat practical reason and the good as consisting in facts about what an agent would desire to care about and pursue upon suitable reflection. Even those who deny that all reasons or intrinsic goods are grounded in desire often assume that some are desire-dependent. Though I will address the more modest claim that some aspects of practical reason or the good are desire-dependent, it will be easier to begin with *pure* desiderative conceptions. One possible focus is desiderative conceptions of practical reason. But many of the same issues arise for desiderative conceptions of the good as well, and it will be useful to discuss these at points. Indeed, it may be most plausible to assign desire an ultimate role when we turn our attention from practical reason or the good, as such, to the narrower topic of a person's good or well-being.

This material was initially presented at an invited session at the Eastern Division Meetings of the APA in December 2003. Stephen Darwall provided extremely useful comments on that occasion. Since then, I have presented this material in a graduate seminar at UCSD in 2004, at a colloquium talk at Rice University in 2006, at a keynote talk at the third annual Metaethics Workshop at the University of Madison in 2006, and at a 2007 Kline Conference at the University of Missouri at Columbia. For helpful feedback I would like to thank Richard Arneson, Daniel Attas, Tom Baldwin, Sarah Buss, Eric Campbell, Thomas Carson, Dale Dorsey, Shelly Kagan, Brian Leiter, Don Loeb, Alistair Norcross, Luke Robinson, Geoff Sayre-McCord, Russ Shafer-Landau, George Sher, Jeff Stedman, William Tollhurst, Peter Vallentyne, Peter Vranas, Nick Zangwill, other members of those audiences, and two anonymous referees.

There are many possible reasons for focusing on desiderative conceptions of practical reason or the good. I will focus on three apparently independent rationales that I believe to be central and to have been influential. Desiderative conceptions fit with the Humean idea that reason can only be *instrumental*. They also promise to explain the way in which recognizing something as reasonable or as beneficial tends to *resonate* with agents or exert a motivational pull on them. Finally, desiderative conceptions promise to explain the *diversity* of reasons and good lives that most of us recognize. By way of explaining the appeal of desiderative conceptions, I will elaborate these three rationales.

However, despite these sources of potential appeal, desiderative conceptions ultimately prove problematic. Their most serious problem is an inadequate account of the normativity of practical reason and the good. In particular, we lack an adequate account of the *normative authority* of desire. An adequate conception of practical reason or the good must not only provide a decent *fit* with our reflective beliefs about what is or could be reasonable or valuable but must also be able to explain *why* we should care about conformity to its demands. Conceptions of practical reason and the good in which desire plays a genuinely foundational role are problematic along both dimensions. Herein lies the appeal of non-desiderative conceptions of practical reason and the good, especially those that are grounded in agency or other values. I try to explain the special appeal of *perfectionist* conceptions that appeal to rational nature or agency.

The adequacy of this sort of perfectionist conception of practical reason and the good depends, in part, upon its ability to respond persuasively to the considerations underlying the three rationales for desiderative conceptions. The resonance constraint appears to favor desiderative conceptions of practical reason insofar as we assume that motivation involves desire and that motivational pull must be found in antecedently held desires. But if desire can be responsive to reason, rather than its master, then desire and, hence, motivation can be consequent upon recognizing reasons or values. Rejecting the Humean dictum that reason can only be the slave of the passions is the key to accommodating the resonance condition without resort to the problematic commitment to desire-dependence. Moreover, the perfectionist appeal to rational nature or agency allows us to explain the commitment to diversity or pluralism about the content reasons and value without the problematic desiderative commitment to content-neutrality.

For all the problems that desiderative conceptions face, they provide an easy explanation of the evident fact that something's being the object of an agent's desire is normally, if not always, a good reason for the agent, if not others, to care about or pursue that thing. It is a problem for perfectionism if it cannot explain this evident fact. The perfectionist should locate rational and evaluative significance in choice or rational endorsement, rather than desire, per se. Desire inherits significance insofar as it can be seen as the product of reasoned choice or endorsement. But rational nature imparts significance not just to the fact of choice or endorsement but to the content of choice or endorsement as well. This raises a question about what attitude the perfectionist should take toward choice of inappropriate ends. I conclude by exploring different models of how to relate these two aspects of the significance of choice.

1. PRACTICAL REASON, THE GOOD, AND WELL-BEING

I am sympathetic with those who take practical reason to be the ultimate currency of normative inquiry.[1] For this reason, I am especially interested in response-dependent and, in particular, desire-dependent conceptions of practical reason. Such conceptions can be motivated, we will see, by familiar assumptions about the nature, limits, and upshot of practical reason. But the primacy of practical reason within normative inquiry is a contestable position. Others take evaluative categories of the good or the good for a person to be primary. Whether we take practical reason or the good to be primary, many of the same issues that arise for practical reason can arise for value. In particular, there are comparable motivations for response-dependent and specifically desire-dependent conceptions of the good and the personal good.

Indeed, this parallelism should come as no surprise if we can treat reasons and values as interdependent. On one such view, we could treat the good as whatever is a legitimate object of rational concern.

> Something is (intrinsically) good just in case it is (intrinsically) rational to care about or pursue it.

We might call this the *Reason–Value Link*.[2] To accept that the good and practical reason are linked in this way does not prejudge the question of which notion, if either, is explanatorily primary. The biconditional relationship is compatible with the good being prior in explanation and

[1] See e.g. T. M. Scanlon, *What We Owe to Each Other* (Cambridge, MA: Harvard University Press, 1998), esp. chs. 1–2, and Stephen Darwall, *Welfare and Rational Care* (Princeton: Princeton University Press, 2002).

[2] Cf. Franz Brentano, *The Origin of our Knowledge of Right and Wrong* [originally published 1889] (London: Routledge, 1969), 18, and C. D. Broad, *Five Types of Ethical Theory* (London: Routledge, 1930), 283.

with practical reason being prior in explanation. This debate may be relevant later. But present purposes do not require taking sides.

This allows us to link practical reason and the good. It does not yet tell us about the evaluative notion of the good for a person. We can equate a person's good with her welfare or well-being, her self-interest, her quality of life, and, on some views, with her happiness.[3] We might link these evaluative notions with rational concern as follows.

> Something is (intrinsically) good for X just in case it is (intrinsically) rational to care about or pursue it for X's own sake.

Call this the *Reason–Well-being Link*.[4] As with the Reason–Value Link, this link does not prejudge which relatum, if either, is explanatorily prior.

Notice that the Reason–Value and Reason–Well-being Links are agnostic about the relationship between the good and the personal good or well-being. Some extreme views eliminate one evaluative concept in favor of the other—denying the existence of the good while recognizing the existence of the personal good, or denying the existence of the personal good while recognizing the existence of the good. For instance, G. E. Moore famously thought that the notion of a personal or relational good is incoherent.[5] Other views are not eliminativist, but reductive; they purport to explain the good in terms of the personal good, or vice versa. For instance, the classical utilitarians, such as Jeremy Bentham, John Stuart Mill, and Henry Sidgwick, all seem to have thought that for something to be good is simply for it to be good for someone and that something's goodness was proportional to how much well-being it advanced.[6] But we can also imagine alternatives to these eliminativist and reductive extremes. For instance, one might recognize goods for persons and believe that things can be regarded as good (simpliciter) insofar as they are good for people or contribute to their well-being and still recognize some things as good independently of

[3] One potential obstacle to equating happiness with these other concepts (personal good, well-being, self-interest, and quality of life) is that, whereas it is comparatively easy to formulate objective conceptions of these other concepts, some people assume that happiness is inherently subjective and does not admit of objective conceptions. For an effective reply that defends the coherence of objective conceptions of happiness, see Richard Kraut, "Two Conceptions of Happiness" *Philosophical Review* 88 (1979), 176–96.

[4] Cf. Darwall, *Welfare and Rational Care*, 8–9.

[5] G. E. Moore, *Principia Ethica* (Cambridge: Cambridge University Press, 1903), 97–105.

[6] See Jeremy Bentham, *An Introduction to the Principles of Morals and Legislation* [originally published 1823] (London: Athlone Press, 1970), Ch. I, §§ iii–v; John Stuart Mill, *Utilitarianism* [originally published 1861] (Indianapolis: Hackett, 1979); and Henry Sidgwick, *The Methods of Ethics* [originally published 1874], 7th edn (London: Macmillan, 1907).

their being good for people. I am not an eliminativist about the personal good, and I think that that the Reason–Well-being Link provides one natural way to approach issues about the personal good. But I will otherwise remain largely agnostic about how best to understand the relation between the good and the personal good.

The Reason–Value and Reason–Well-being Links do not settle substantive questions about either practical reason or the good but they should allow us to move between claims about practical reason, the good, and well-being and to formulate desiderative conceptions of any of them.

2. SKEPTICISM AND INSTRUMENTALISM ABOUT PRACTICAL REASON

In *The Treatise of Human Nature* David Hume famously claims that "Reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them."[7] It is natural to interpret this and other remarks Hume makes as implying skepticism about practical reason. In particular, Hume understands reason as a faculty that allows us to judge of the truth or falsity of ideas (III.i.1/458). Ideas are representations or copies. Actions and passions, as such, are real existences, not ideas. It follows that neither actions nor passions and desires, as such, can be in conformity with or contrary to reason.[8] However, Hume does allow that actions and passions can be contrary to reason but only so far as they are dependent beliefs about matters of fact or relations of ideas. Many actions and desires are so dependent. In particular, desires and ultimately actions are often the product of other desires and beliefs about the means or necessary conditions to satisfying those antecedent desires. As Hume writes in his *Enquiry Concerning the Principles of Morals*,

> Ask a man *why he uses exercise*; he will answer, *because he desires to keep his health*. If you then enquire, *why he desires health*, he will readily reply, *because sickness is painful*. If you push your enquiries farther, and desire a reason *why he hates pain*, it is impossible that he can ever give any. This is an ultimate end, and is never referred to any other object.[9]

[7] See David Hume, *A Treatise of Human Nature* [originally published 1739], ed. P. H. Nidditch (Oxford: Clarendon Press, 1978), book II.part iii.section 3/page 415.

[8] Cf. Stephen Darwall, *Impartial Reason* (Ithaca, NY: Cornell University Press, 1983), 53.

[9] David Hume, *An Enquiry Concerning the Principles of Morals* [originally published 1751], ed. P. H. Nidditch (Oxford: Clarendon Press, 1975), appendix I, section v.

One can often trace an agent's actions to desires that are *derived* from other desires and the agent's beliefs. And these desires may themselves be derived desires. But ultimately one must trace back through derived desires to some *ultimate* desire that is not derived from others. Derived desires and the actions that are based on them can be unreasonable, Hume claims, in the sense that they can be based on false beliefs about the causal means or necessary conditions to satisfying other desires—false beliefs about what we might call instrumental relations. But, he seems to assume, actions or desires that are not based on false beliefs about instrumental relations cannot be contrary to reason. It follows, as Hume infamously claims, that

> 'Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. 'Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an *Indian* or person wholly unknown to me. 'Tis as little contrary to reason to prefer even my own acknowledg'd lesser good to my greater, and to have a more ardent affection for the former than the latter. (II.iii.3/416)

Of course, gross solipsism and imprudence can be, and typically will be, contrary to reason in the sense that they will frustrate the satisfaction of other ultimate desires we have that presuppose the continued existence of ourselves and the world. Hume's claim in this passage is presumably that such behavior and preferences are not inherently contrary to reason and are not, considered in themselves, contrary to reason.

Whereas Hume does claim that derived desires based on false beliefs can be contrary to reason, he denies that ultimate desires can be reasonable and that actions or derived desires are rational when they are based on true beliefs about instrumental relations. This asymmetry between ascriptions of irrationality and ascriptions of rationality implies that Hume is best interpreted as a skeptic about practical reason. Not only are no actions or desires inherently contrary to reason but also no actions or desires are rational. The crucial questions in assessing Humean skepticism are why we should accept this asymmetry and why we should think that reason can only judge of the truth or falsity of ideas or beliefs.

Some modern-day conceptions of practical reason and the good appeal to Hume's claims but draw different conclusions. *Instrumentalism* about practical reason accepts Hume's claim that reason can only be the slave of the passions or appetites. Derived desires can be criticized as based on false beliefs about instrumental relations, and so can actions based on such derived desires. But actions and desires are not otherwise criticizable and, in particular, ultimate desires or ends are not rationally criticizable. But, unlike Hume, the instrumentalist does assume that practical reason endorses desires or actions that contribute to the satisfaction of one's desires, provided these desires are not themselves based on false beliefs about instrumental relations. The instrumentalist rejects Hume's asymmetry about ascriptions of rationality and irrationality. Like Hume, the instrumentalist maintains that ultimate ends are neither reasonable nor unreasonable, but she rejects the conclusion that desires and actions conducive to satisfying ultimate ends are not rational. Because ultimate ends are immune to rational criticism, and because all derived desires relate ultimately to ultimate ends, instrumental rationality can be defined in terms of promoting one's ultimate ends or desires. Instrumental rationality, on this view, is a matter of adopting means and necessary conditions to the promotion of one's ultimate ends. One's ultimate ends can change over time. So presumably instrumental rationality must be temporally relative, relativizing one's reasons for action to one's ultimate ends at the time of action. A great many people recognize instrumental rationality, so construed, as one aspect of practical reason. But if we accept the Humean claim that reason can only be the slave of the passions, then it appears that there could be nothing more to practical reason than instrumental rationality.

Though instrumentalism is typically formulated as a claim about practical reason, related claims can be formulated about the good. Indeed, if we accept the Reason–Value Link, then a purely instrumental conception of practical reason yields a conception of the good that makes something's goodness consist in its conduciveness to satisfying one's ultimate desires.

Though Hume himself draws largely skeptical conclusions from his assumption that reason can only be the slave of the passions, the instrumentalist draws a more constructive conclusion. Because of the basis of instrumentalism in some Humean claims, instrumentalists are often viewed as Humeans. We do no serious harm by calling instrumentalists Humeans, provided that we remember that Hume was no Humean.

3. RESONANCE AND INTERNALISM

Another influential rationale for response-dependent and specifically desiderative conceptions of practical reason and the good is the thought that normative notions, such as practical reason and the good, should not leave the agent indifferent but should *resonate* with her. Resonance requires that normative claims be capable of motivating agents. But motivation is a matter of having suitable pro-attitudes or desires. Hence, normative claims must be grounded in an agent's desires in some way.

We can clarify this rationale by looking at Bernard Williams's influential defense of internal reasons.[10] Williams focuses on reasons for action and identifies internal reasons as ones that are relative to the agent's "subjective motivational set" (pp. 101–2). External reasons, by contrast, would not depend on the agent's motivational set. Williams clearly identifies the relevant elements of a person's motivational set with her desires in a broad sense that encompasses various kinds of pro-attitudes (pp. 101, 105). He is not explicit about the reasons for focusing on desires. Presumably, he is attracted to the familiar view of intentional action as the product of representational states, such as belief, and pro-attitudes, such as desire. On this reconstruction, we can distinguish, at least in principle, between the internalist constraint on practical reason that reasons for action be capable of motivating the agent and a specifically desiderative conception of practical reason that grounds reasons for action in the agent's desires. Because Williams believes that motivational states involve desires, he concludes that only a desiderative conception of practical reason can satisfy the internalist constraint.

Williams makes clear that his preferred desiderative conception of internalism will not simply appeal to an agent's actual desires but will instead recognize idealizations of her desires. An agent does not have an internal reason, according to Williams, to satisfy derived desires that are based on false beliefs about the instrumental means to and necessary conditions of satisfying her more ultimate desires (pp. 102–3). Because an agent may be mistaken about what will be most conducive to satisfying her ultimate desires, she can be mistaken about what her internal reasons are (p. 103). Williams is willing to countenance internal reasons that are relative to the desires that an agent would have after suitable deliberation on and from her initial (pre-deliberative) desires (pp. 104–5).

Unfortunately, Williams is frustratingly vague about what he will count as suitable deliberation (pp. 105, 110). If internalism is to avoid vacuity, then motivation and desire must play the ultimate role in the justification of action. But this precludes appeal to desires that are produced by forms of deliberation that track truths about practical reason or the good. For if the new desires depend upon deliberation about practical reason or the good, the agent would have them regardless of the desires with which she began. But this would violate the demand that practical reason be traceable to the agent's initial motivational set. Presumably, Williams has in mind content-neutral forms of deliberation, such as means–ends reasoning and imaginative and vivid appreciation of the causes, nature, and consequences of one's alternatives.

[10] See Bernard Williams, "Internal and External Reasons" reprinted in his *Moral Luck* (Cambridge: Cambridge University Press, 1981).

This gives us a better idea of how Williams understands his preferred desiderative conception of internalism. But why should we accept such an account of practical reason? Williams appeals to connections between motivation and possible explanation.

> If something can be a reason for action, then it could be someone's reason for acting on a particular occasion, and it would then figure in an explanation of that action. Now no external reason statement could *by itself* offer an explanation of anyone's action. ... The whole point of external reasons statements is that they can be true independently of the agent's motivations. But nothing can explain an agent's (intentional) actions except something that motivates him to act. (pp. 106–7)

But this appeal to explanation is problematic. We can put the problem as a dilemma.

On the one hand, it cannot be that reasons for action must actually motivate and explain the agent's actual behavior. Conceptions of practical reason are concerned with reasons that would *justify*, rather than explain, action. So we want to allow that an agent's justifying reasons—what she ought to do—may not be the reasons that motivate her or explain her behavior. Moreover, the idealization contained in Williams's own desiderative conception means that internal reasons often fail to motivate and explain an agent's actions. If my desire to drink the substance in this glass, which is petrol, is based on the false belief that it is gin, then Williams thinks that the internalist should recognize no reason to drink the stuff in the glass and a reason not to drink it. But then the agent's internal reason not to drink the stuff in the glass will not explain his actual drinking of the stuff in the glass.

On the other hand, we might loosen the link between reasons for action and motivation and explanation, requiring only that an agent's practical reasons must be *potentially explanatory*. One way to see an agent's reasons for action as potentially explanatory is to recognize that her reasons explain her action just insofar as she is behaving rationally. But this threatens to become a trivial or vacuous requirement. For any conceivable standard of behavior X, no matter how peculiar, it will be true that X explains an agent's actions just insofar as she is behaving X-ly. But that means that this looser version of the explanatory rationale provides no constraint at all on the content of reasons for action.

The problem is that it is not clear that we can motivate and articulate the internalist requirement in a sensible way by appeal to explanation, actual or possible. A more promising interpretation focuses on the potential for *alienation* in externalist conceptions of practical reason. In his earlier influential criticism of utilitarianism, Williams identifies the unreasonable character of utilitarian demands with the way in which they alienates agents from their projects and attitudes.

> It is absurd to demand from … a man, when the sums come in from the utility network which the projects of others have in part determined, that he should just step aside from his own projects and decision and acknowledge the decision which utilitarian calculation requires. It is to alienate him in a real sense from his actions and the source of his actions in his own convictions. It is to make him into a channel between the input of everyone's projects, including his own, and an output of optimific decision; but this is to neglect the extent to which *his* actions and *his* decisions have to be seen as the actions and decisions with which he is most closely identified.[11]

In "Persons, Character, and Morality" Williams generalizes this concern about alienation from utilitarianism to Kantian and other impartial moral theories.[12] We might then interpret Williams's defense of internal reasons as articulating the conception of practical reason underlying these worries about utilitarianism and other impartial moral theories. On this reading, Williams is appealing to what might be called a *resonance* constraint—an agent's reasons for action, at least when recognized as such, must be capable of commanding and sustaining her emotional allegiance and motivational engagement. Internalist conceptions of practical reason, which relativize an agent's reasons to her motivational capacities, meet this resonance constraint. By contrast, externalist conceptions of practical reason, which do not relativize an agent's reasons to her motivational capacities, appear unable to satisfy the resonance constraint. If, as Williams believes, something is capable of motivating someone in the relevant way only if it is conducive to satisfying her actual desires or the desires she would have were she to follow the right deliberative procedures, then it follows that his desiderative conception of practical reason is the best way of satisfying the resonance constraint.

We might extend this resonance constraint from conceptions of practical reason to conceptions of the good. We are forced to do this if we accept the Reason–Value Link. Intuitionists, such as Moore, advanced theories of the good that treat the good as independent of and prior to the good for a person. Indeed, Moore found the latter notion incoherent. He recognized various

[11] Bernard Williams, "A Critique of Utilitarianism" in J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1973), 116–17. Evan Tiffany helped me see the relevance of Williams's critique of utilitarianism to understanding his defense of internal reasons. See Evan Tiffany, "Alienation and Internal Reasons for Action" *Social Theory and Practice* 29 (2003), 387–418. However, Tiffany's interpretation of Williams seems to distinguish the appeals to a non-alienation constraint and to a motivational constraint. On my view, the motivational constraint is best interpreted as following from a non-alienation or resonance constraint.

[12] Bernard Williams, "Persons, Character, and Morality" reprinted in his *Moral Luck*, esp. 14.

things as intrinsically good—including beauty itself—independently of any contribution that such goods make to a person's good.[13] But we might well doubt whether Moore's intrinsic goods, understood as impersonal goods, would satisfy the resonance condition.[14] They certainly would be correlated with external, rather than internal, reasons. Indeed, this worry for Moore might extend to any conception of an impersonal good. Why should any conception of the good, which is in no way relative to the interests of persons, resonate with agents?

It is easier to see how a conception of the good for a person or well-being might satisfy a resonance constraint, precisely because an account of the personal good can be internalist and desiderative. Peter Railton appeals to something like a resonance constraint in motivating his own desiderative conception of well-being.

> It does seem to me to capture an important feature of the concept of intrinsic value to say that what is intrinsically valuable for a person must have a connection with what he would find in some degree compelling or attractive, at least if he were rational and aware. It would be an intolerably alienated conception of someone's good to imagine that it might fail in any such way to engage him.[15]

If we assume that such engagement requires the potential to motivate and that motivation requires suitable desires, then resonance leads us to a response-dependent and specifically desiderative conception of well-being.

Desiderative conceptions of well-being have a distinguished pedigree. In *Utilitarianism* Mill at least suggests an idealized desire conception of happiness when he explains the intrinsic, and not just instrumental, superiority of higher pleasures by appeal to the preferences of a competent judge.

> If I am asked what I mean by difference of quality in pleasures, or what makes one pleasure more valuable than another, merely as a pleasure, except its being greater in amount, there is but one possible answer. If one of the two is, by those who are competently acquainted with both, placed so far above the other that they prefer it, even though knowing it to be attended with a greater amount of discontent,

[13] *Principia Ethica*, 83–5. Cf. W. D. Ross, *The Right and the Good* (Oxford: Clarendon Press, 1930), ch. 5.

[14] Some claim that the real legacy of Moore's open question argument is recognition of the normativity of ethics and, in particular, the good. See e.g. Stephen Darwall, Allan Gibbard, and Peter Railton, "Toward *Fin de Siècle* Ethics: Some Trends" *Philosophical Review* 101 (1992), 115–89, and Connie Rosati, "Naturalism, Normativity, and the Open Question Argument" *Noûs* 29 (1995), 46–70. If normativity is articulated in such a way as to yield an internalist constraint, then Moore's own conception of the good threatens to run afoul of the open question argument.

[15] Peter Railton, "Facts and Values" *Philosophical Topics* 14 (1986), 9. See also Connie Rosati, "Internalism and the Good for a Person" *Ethics* 106 (1996), 297–326.

and would not resign it for any quantity of the other pleasure which their nature is capable of, we are justified in ascribing to the preferred enjoyment a superiority in quality so far outweighing quantity as to render it, in comparison, of small account. (*Utilitarianism* ii.5)

At one point in *The Methods of Ethics* Sidgwick proposes that we understand a person's overall good to consist in "What he would now desire and seek on the whole if all the consequences of all the different lines of conduct open to him were accurately foreseen and adequately realized in imagination at the present point in time" (*Methods* 111–12). In *A Theory of Justice* John Rawls adapts Sidgwick's proposal and identifies a person's good with a rational plan of life. "It is the plan that would be decided upon as the outcome of careful reflection in which the agent reviewed, in light of all the relevant facts, what it would be like to carry out all of these plans and thereby ascertained the course of action that would best realize his more fundamental desires."[16] In *A Theory of the Good and the Right* Richard Brandt identifies a person's well-being with what it would be rational for her to desire, and he understands rational desire as desire that would survive a process of cognitive psychotherapy that requires full and vivid exposure to logic and the relevant facts.[17]

However, appeal to resonance suggests some modifications in the classical informed desire theory of well-being. Recognizing that even in a more idealized state we might have desires that we do not endorse or identify with, David Lewis proposes that something is good just in case one would, under conditions of full imaginative acquaintance with the alternatives, desire to desire it.[18] Railton notices that an ideal appraiser is likely to be very different from the actual self that it idealizes and that, consequently, what my idealized self may want for himself may not be appropriate for me. For instance, education appears to be a good for my actual self, but because my idealized self is already fully informed,

[16] John Rawls, *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971), 417.

[17] Richard Brandt, *A Theory of the Good and the Right* (Oxford: Clarendon Press, 1979), esp. chs. 4–8.

[18] David Lewis, "Dispositional Theories of Value" *Proceedings of the Aristotelian Society*, suppl. vol. (1989), 113–37. However, the merits of idealizing to second-order or aspirational desire is open to question. Some appeal to aspirational desires to reveal an agent's "true" self or values. But I see no reason to privilege aspirational desires in this way. If the unwanted first-order desires systematically regulate the agent's deliberations and actions and contrary aspirational desires express themselves only occasionally and ineffectually, as in so many New Year's resolutions, then it's hard to treat the aspirational desires as reflecting the agent's true self or values. It is also hard to see how in such a case reasons or values grounded in merely aspirational desires could be more resonant than those grounded in central first-order desires.

he may not desire (or desire to desire) to get an education. To remedy this source of potential alienation, Railton proposes that we appeal to what the ideal appraiser would want his actual self to want—in effect, what A+ would want A to want. "[A]n individual's good consists in what he would want himself to want, were he to contemplate his present situation from a standpoint fully and vividly informed about himself and his circumstances, and entirely free of cognitive error or lapses of instrumental rationality."[19] Railton's Ideal Advisor theory is perhaps the most sophisticated articulation of the informed desire theory of well-being, and it will be useful at points to focus on it.[20] Railton's theory illustrates well how appeal to resonance lends support to desiderative conceptions of well-being.

4. PLURALISM ABOUT PRACTICAL REASON AND THE GOOD

A final rationale for desiderative conceptions of practical reason and the good is its ability to explain the apparent diversity of rational plans and goods, especially good lives. It is common to think that there is typically more than one reasonable course of action in a given situation. Even where there is a uniquely reasonable course of action for an agent to take in a given situation, that path is typically uniquely reasonable relative to an agent's larger plan of life. But it also seems evident that there are many different equally or comparably reasonable plans of life. What is evident about practical reason is also evident about the good, especially well-being. Indeed, given the Reason–Well-being Link, the diversity of possible objects of rational concern insofar as one is concerned about someone for his own sake implies the diversity of well-being. Typically, at any one point in a person's life, there are many different activities, projects, and commitments that would contribute constitutively to an agent's good. Even where one activity, project, or commitment is uniquely valuable, such goods are typically uniquely beneficial relative to some previous and larger activity, project, or commitment. But there surely is a plurality of diverse projects

[19] "Facts and Values" 16.

[20] Also see Thomas Carson, *Value and the Good Life* (Notre Dame IN: University of Notre Dame Press, 2000). In "Internalism and the Good for a Person" Rosati suggests that to avoid alienation Railton needs to add that one's actual self (A) be prepared to care about the way in which one's ideal self (A+) is different from one's actual self. However, idealization is a normative notion. If A+ is better situated epistemically than A, then A ought to care about A+'s advice for A. +, after all, is essentially desirable.

and lifestyles that are equally or comparably good for the person whose life it is.[21]

Desiderative conceptions appear well positioned to explain this kind of pluralism about the reasonable and the good. Desiderative conceptions are *subjective* insofar as they ground practical reason and the good in an agent's contingent and variable psychological states. Because of this subjectivity, desiderative views appear to underwrite pluralism. Now it should be noted that the most interesting desiderative conceptions do not appeal to actual desires, but rather to *idealized* desire. While it is quite evident that people do differ in their actual desires, it is less clear that they will differ in their idealized desires. This will depend in part upon the sort of idealization in question. For instance, if the relevant idealization simply incorporated certain rational concerns or values, then there would be no reason to expect a diversity of idealized desires. But, in discussing Williams, we saw that any such conception of the process of idealization would no longer assign desire a foundational role. Desire would not explain reason or value, because the relevant desires would presuppose prior reasons or values. What a genuinely desiderative conception of practical reason or the good requires is a conception of idealization that is *content-neutral*. This, I suggested, is a constraint that Williams has reason to recognize on the form that deliberation may take within an internalist view. Moreover, this is a constraint that appears to be observed by all those advancing desiderative conceptions of well-being, certainly by Rawls, Brandt, Lewis, and Railton. Provided the relevant kind of idealization is content-neutral, desiderative conceptions must allow for the possibility of diverse objects of desire both for a given agent and for different agents.

The subjectivity of desiderative conceptions contrasts with more objective conceptions of practical reason and the good. In fact, we could just equate objective and non-desiderative conceptions. On this view, a conception of practical reason or the good is objective just in case it identifies things as reasonable or valuable independently of being the object of the agent's actual or informed desire. For instance, external reasons would be objective in this sense. If there is a categorical reason to be concerned about one's own good or the good of others, whose authority is independent of one's caring about these things, then practical reason will be objective. Moreover, one might understand a person's good in objective terms as consisting, for example, in the *perfection* of one's essential (e.g. rational or deliberative) capacities or in some *list* of disparate objective goods (e.g. knowledge, beauty, achievement, friendship or equality). The invariant character of

[21] By comparable value I have in mind something like the notion of parity defended in Ruth Chang, "The Possibility of Parity" *Ethics* 112 (2002), 659–88.

objective reasons and goods appears to restrict severely the diversity rational plans and good lives.

5. THE REDUCTIVE CHARACTER OF DESIDERATIVE CONCEPTIONS

Desiderative conceptions of practical reason and the good identify the reasonableness or value of something with its tendency to produce a certain sort of response in an agent or appraiser. As such, desiderative conceptions represent a kind of *dispositional* and *response-dependent* approach to practical reason and value. It is important to notice, however, that desiderative conceptions involve a *reductive* form of dispositionalism and response-dependence. In particular, desiderative views reduce normative notions of reasonableness or value to non-normative facts about desire.[22]

We might contrast desiderative conceptions with two different kinds of *non-reductive* dispositionalism. One form of dispositionalism is *overtly* non-reductive, because it expressly invokes normative notions into the dispositional analysis of normative notions. One way for normative notions to figure overtly within a dispositional analysis of normative notions is for it to focus on responses that involve normative belief. For example, an attempt to analyze the good in terms of things that an appraiser is disposed to judge valuable would clearly be non-reductive.[23] Alternatively, the idealization, rather than the response itself, may be overtly normative. For example, John McDowell proposes that something is valuable just in case it is such as to *merit* approval.[24] Other forms of dispositionalism, while not overtly non-reductive, are nonetheless *implicitly* non-reductive. This will be so when either the response itself or the idealization is implicitly normative. For example, if we were to analyze something's value in terms of its tendency to elicit certain kinds of emotional responses, such as pride or resentment,

[22] Brandt is clear that his concept of rationality "does not import any substantive value judgements" (*A Theory of the Good and the Right*, 13). Lewis explicitly acknowledges the reductive character of his dispositional conception of value ("Dispositional Theories of Value," 113). Railton comes close ("Facts and Values," 9). Though proponents of desiderative conceptions do not always explicitly acknowledge the reductive character of their views, I don't think that this aspect of their views is in dispute.

[23] For instance, Firth resists Ideal Observer Theories that analyze the rightness of conduct in terms of it tendency to elicit beliefs that it is right. See Roderick Firth, "Ethical Absolutism and the Ideal Observer" *Philosophy and Phenomenological Research* 12 (1952), esp. 325–9.

[24] John McDowell, "Values and Secondary Qualities" in *Morality and Objectivity*, ed. T. Honderich (Boston: Routledge & Kegan Paul, 1985). Cf. Darwall, *Welfare and Rational Care*.

under certain conditions, then our view would be non-reductive, insofar as these emotional responses involve constitutive normative judgments about something being valuable or involving wrongdoing. Alternatively, our idealization might be implicitly normative. For example, David Wiggins proposes that something is valuable just in case it is such as to produce approval in the *appropriate* sort of appraiser.[25] Though one could have a reductive conception of an appropriate appraiser, Wiggins makes clear that he thinks that an appropriate appraiser is a good judge and that a good judge is one who is apt to get things right.[26] There is a final way in which a dispositional or response-dependent conception might be implicitly non-reductive. A dispositional view might analyze normative notions of reasonableness or value in terms of tendencies to elicit psychological responses that do not themselves involve normative judgment, but it will still be implicitly non-reductive if the rationale for focusing on those particular responses or responses formed in that particular way is the desire to constrain the results in ways that meet some independent normative criteria. For instance, if we understand appeal to an ideal appraiser or advisor as an impartial and sympathetic appraiser whose desires are formed on the basis of an equally sympathetic identification with the interests of all affected parties, then our conception of idealization is not content-neutral; it stacks the deck in favor of some normative outcomes.[27] Such a view would not be genuinely reductive, because it explains normative notions in terms of a class of psychological states that has been selected on normative grounds.

Though I believe that the reductive character of desiderative conceptions of practical reason and the good ultimately poses problems for the normative adequacy of such conceptions, their reductive character looks like a virtue in a dispositional analysis. Such conceptions present an informative dispositional analysis of normative notions in which the appraiser's or advisor's response does genuine explanatory work. By contrast, non-reductive forms of response-dependence threaten to provide analyses that are circular, in which the responses do no real explanatory work, or that are comparatively uninformative.

For instance, someone might analyze goodness as a property of objects that tends to elicit in ideal conditions and appraisers the judgment that

[25] David Wiggins, "A Sensible Subjectivism?" in his *Needs, Values, and Truth* (Oxford: Blackwell, 1987).

[26] Ibid. 194–5.

[27] Smith's dispositional analysis of rightness is non-reductive in this way insofar as he places substantive constraints on the kinds of acts that a fully rational person would desire to perform. See Michael Smith, *The Moral Problem* (Oxford: Blackwell, 1994), chs. 5–6, esp. 184.

it is good or valuable. Here we invoke the very value we are analyzing in our analysans. It is true that, on this view, we analyze X, not in terms of X, but in terms of beliefs about X. But if we accept the not unreasonable assumption that any story about what makes a belief a belief about X must eventually advert to X, then it appears that this sort of analysis is ultimately circular.[28]

Even non-reductive conceptions that are not strictly circular may deprive the appraiser's response of genuine explanatory value. Any conception of ideal conditions, the ideal appraiser, or her responses that is not content-neutral threatens to make the appeal to her responses otiose. We could apparently bypass her responses and appeal directly to the normative criteria that inform the selection of specific kinds of idealization or sensibilities. Just as a rigged election means that the voting itself does not explain the outcome, so too a content-specific conception of ideal conditions, the ideal appraiser, or her responses threatens to make the appraiser's responses an idle wheel.[29]

Finally, even if the non-reductive analysis is not strictly circular and the response is not explanatorily idle, the analysis is likely to be comparatively uninformative. Consider the Reason–Well-being Link, which could be used to analyze well-being in terms of what it would be rational to care about for someone for his own sake. This might be put forward as a non-reductive form of response-dependence about well-being that is not circular and in which concern plays an important explanatory role. Even if this is true, the view is comparatively uninformative about what well-being consists in. An important measure of content or information is what possibilities are ruled out. But the Reason–Well-being Link places no substantive constraints on well-being. So even if it is true, it is comparatively uninformative.

[28] Here, I've been influenced by Jonathan Cohen. It is a somewhat open question just what conclusion to draw from the circularity of some non-reductive forms of dispositionalism. Wiggins is happy to concede the circularity of his form of dispositionalism. He views circularity as a defect in a definition or analysis, but not in the sort of commentary or elucidation that he claims to offer. All he cares about is whether the biconditional is true ("A Sensible Subjectivism?" 188–9). I have some sympathy with Wiggins's more modest methodological aspirations. However, I think that the capacity of this sort of circular elucidation to illuminate is limited.

[29] Stephen Darwall notes that I tend to equate reduction and content-neutrality and suggests that some forms of dispositionalism might be non-reductive but content-neutral. One conception of well-being that might be like this results from accepting the Reason–Well-being Link but treating reasons for concern as explanatorily prior to well-being. I am sympathetic to this view, but it strikes me as a conceptual proposal about how to understand the interdependence of reason and value, rather than a substantive conception of well-being. Moreover, insofar as it grounds well-being in rational concern, I doubt that desire plays any significant explanatory role in this view.

6. THE NORMATIVE ADEQUACY OF DESIDERATIVE CONCEPTIONS

We have examined three rationales for desiderative conceptions of practical reason and the good. We now need to ask about the normative adequacy of such conceptions. We might begin by noticing the way in which desiderative conceptions promise to reconcile two distinct, and potentially conflicting, dimensions of normativity. Normative considerations purport to guide conduct and concern and to provide reasons for conduct and concern. This may lead us to think that normative considerations ought to be capable of motivating agents to conform to their guidance. We interpreted this idea as imposing a resonance constraint and saw how grounding practical reason or the good in an agent's actual or idealized desire promises to satisfy this constraint. But the need for normative guidance presupposes the possibility that one's actual ends or desires are mistaken or defective in some way. In practical deliberation, we are interested not just in discovering what we already want, but also what we should want. Normativity presupposes *fallibility*. Simple desire-satisfaction conceptions of practical reason or well-being are poorly placed to recognize robust forms of fallibility. But idealized desire conceptions promise to recognize ways in which an agent's actual goals can be mistaken and criticizable while maintaining the connection with desire apparently necessary to secure resonance.

In assessing the normative adequacy of any conception of practical reason or the good, we must bear in mind two issues. One aspect of normative adequacy is how plausible we find the actual and potential guidance that the conception offers. How well does it accommodate what we are prepared, on reflection, to think about the normative valence of various actual and hypothetical situations? Call this dimension of normative adequacy reflective *accommodation*. No conception is likely to be a perfect match with our reflective judgments, if only because our reflective judgments about various actual and possible situations are likely not to be perfectly consistent. If so, perfect accommodation is impossible and any acceptable conception of practical reason or the good will be revisionary to some extent. But we should be skeptical of conceptions that are highly normatively revisionary, especially if the view has no compensating theoretical virtues. All else being equal, we should prefer a conception that provides greater accommodation of our independent beliefs about practical reason or the good to one that provides less accommodation.

A second aspect of normative adequacy is how well a conception of practical reason or the good explains the normative *authority* of whatever it takes to be fundamental. If a conception of practical reason or the good is to supply normative guidance about what agents should care about or how they should act, it ought to be able to explain why we should care about whatever it takes to ground reasons for action or value. Any adequate conception must provide a *rationale* for the normative authority of its demands.

Despite various kinds of potential appeal, desiderative conceptions of practical reason and the good are problematic along both of these dimensions of normative adequacy.[30] They provide poor accommodation and lack an adequate rationale. We have identified idealized desire conceptions as the normatively most adequate version of the desiderative approach, but it will be useful to begin with difficulties for desiderative conceptions that involve less idealization and recognize fewer kinds of fallibility.

We might begin with the basic desire-dependent conception of practical reason and its failures of accommodation. Some of its problems are precisely those most obviously corrected by idealization. It attaches normative significance to satisfying desires that are based on mistaken factual beliefs, for instance, about the instrumental means to satisfying other desires or that are based on faulty inferences. But there are other problems. Agents can fail to have desires to do things that they appear to have reason to do.

Most of us recognize other-regarding moral duties of justice, fidelity, forbearance, and aid, and many would think that these moral duties generate at least pro tanto reasons for action, such that noncompliance is at least to that extent contrary to practical reason and open to rational criticism. But it seems quite possible for someone to be indifferent to such duties, if not as a matter of principle, then at least in particular cases. Perhaps depression or some more systematic neurological dysfunction underlies the indifference. In such cases, the basic desiderative model fails to recognize reasons that many of us would.

Another problem concerns time preferences. It is a common view that practical reason requires a temporally neutral concern with the way in which goods and bads are distributed within lives. Various forms of temporal bias are among our paradigms of irrationality. For instance, the long-term benefits of regular, routine preventive and corrective dental care make such treatment rational, even if it involves more short-term discomfort than ignoring one's dental health. Similarly, the long-term benefits of good

[30] My claims here merit comparison with those of Richard Kraut, "Desire and the Human Good" *Proceedings and Addresses of the American Philosophical Association* 68/2 (1994), 39–54, and Richard Arneson, "Human Flourishing Versus Desire Satisfaction" *Social Philosophy & Policy* 16 (1999), 113–42.

grades and a good education justify the short-term sacrifices involved in doing one's homework and studying hard for major exams. But it is also a familiar, if unfortunate, fact that many people are temporally biased, investing proximate goods and harms with significance out of proportion to their actual magnitude. But if the temporal bias or discounting is strong enough, then the basic desiderative model must endorse its rationality and condemn temporal neutrality. This fails to account for what many would regard as the unconditional irrationality of temporal bias.[31]

These problems with the basic desiderative conception of practical reason might lead us to explore its plausibility as a conception of the narrower concept of personal good or well-being. As the Reason–Well-being Link implies, we need here to ask whether the satisfaction of desire, whether actual or idealized, is what guides what we care about when we are concerned for someone's own sake. But the implications of the desiderative model are not much better here. Some problems carry over. The basic model implausibly attaches significance to desires that are based on mistaken factual beliefs and faulty inferences. Moreover, temporal neutrality is at least as plausible a constraint on an agent's overall good as it is on practical reason, as such. But then the basic model must condition the rationality of temporal bias on the psychological fact of temporal bias. But this ignores what appears to be the unconditional irrationality of temporal bias within a conception of someone's good.

Another problem for the basic desiderative model of well-being is that it attaches significance to satisfying desire without in any way constraining the content of desire. But most of us think that people can be satisfying their deepest desires and yet lead impoverished lives if their desires are for unimportant or inappropriate things. For instance, we are unlikely to view the life of someone devoted to collecting lint as a richly valuable life, no matter how successful a lint collector he is.[32] What I would want for my son for his own sake is not content-neutral in this way.

Moreover, desire-satisfaction would seem to counsel *adapting* our desires to fit our circumstances, for by adapting our desires, we increase the probability of achieving our aims. Such adaptive views of happiness are familiar from Plato's *Gorgias* and Epicurean ethics. No doubt there is an element of truth in this view, insofar as it often seems advisable to maintain some degree of realism in one's aspirations and ambitions. But there are many ways to

[31] For a partial defense of temporal neutrality, see David O. Brink, "Prudence and Authenticity: Intrapersonal Conflicts of Value" *Philosophical Review* 113 (2004), 215–45.

[32] Cf. Rawls's discussion of a person whose chief desire is to spend his life counting the blades of grass in the fields around him (*A Theory of Justice*, 432).

explain the importance of realism in one's aims. The basic desire-satisfaction model seems committed to unrestricted adaptation. The extreme adaptive approach to happiness is effectively criticized in Aldous Huxley's dystopia *Brave New World* in which Deltas and Epsilons form the working classes who are genetically engineered and psychologically programmed to acquiesce in and indeed embrace intellectually and emotionally limited lives that are liberally seasoned with mood-altering drugs.[33] Deltas and Epsilons lead contented lives precisely because they are satisfying their chief desires. They've got what they want. It's their desires that are frightening. We do not (in general) increase the value of our lives by lowering our sights, even if by doing so we increase the frequency of our successes.[34]

Furthermore, we may wonder whether the basic-desire satisfaction conception of well-being doesn't confuse what is in our interests and what interests us.[35] For it is not clear that everything that one might desire, even reasonably desire, would contribute to one's good. Satisfying my desire for personal achievement or friendship might be good for me. But it is not at all clear that the satisfaction of my desire that a cure for AIDS be discovered or that world hunger be relieved contributes to my well-being (assuming that I do not suffer from AIDS or hunger). Without further argument, it is hard to believe that the satisfaction of these desires, however admirable, contributes to my own good.[36]

One might try to respond to this worry by focusing, for purposes of well-being, on a narrower class of desires. For example, one might focus, as the Reason–Well-being Link also does, on desires one has for

[33] Aldous Huxley, *Brave New World*, 2nd edn (New York: Harper & Row, 1946). I take Huxley's Brave New World to be not merely a dystopia but an allegory for certain aspects of modern life. Interestingly, Huxley suggests that the proper lesson to be drawn from such a dystopia is recognition of a higher (perfectionist) form of utilitarianism (ibid., pp. viii–ix).

[34] This reflects the tension between control and completeness constraints in ancient discussions of *eudaimonia*. *Eudaimonia* can only be fully within the agent's control if we sacrifice completeness. Callicles implicitly recognizes this when he replies to Socrates's adaptive conception of happiness by saying that Socratic happiness is fit only for a stone or corpse (492e5). Cf. "Better to have loved and lost than never to have loved at all."

[35] See Kraut, "Desire and the Human Good," 40–1 and Stephen Darwall, "Self-Interest and Self-Concern" *Social Philosophy & Policy* 14 (1997), 158–78.

[36] Though we don't want to identify what interests us with what is in our interest, the two can be interdependent. If, for instance, I make a life's project out of pursuing a cure for AIDS or fighting poverty, then it is more plausible to treat the satisfaction of such projects as contributing to my own well-being. Scanlon makes a similar point by distinguishing between informed desires and rational aims, and using the latter, rather than the former, to inform his conception of well-being. See Scanlon, *What We Owe to Each Other*, 120–6. This difference between the role of desires and projects within an account of well-being can be explained, I believe, by the sort of perfectionist conception I defend below.

someone's own sake. Presumably, the basic desire model would explain X's well-being in terms of the satisfaction of desires that X has for her own sake. The problem with this proposal is that we can't understand the focus of such desire—one's own sake—independently of well-being. But then if the basic model is restricted in this otherwise natural way, it ceases to be reductive and so loses a principal virtue of the desiderative form of response-dependence.

So the basic desiderative model of practical reason and well-being does not accommodate many of our intuitions about reasons and value. But also it fails to explain the normative authority of desire. Though it may be commonly assumed that our desires always provide reason for action or that their satisfaction contributes to our good, it is not at all clear why we should care about the satisfaction of desires independently of the way in which they were formed or of their content. There is no apparent rationale for the normative authority of desire.

It might seem that we could answer some of these doubts about the normative significance of desire by appeal to idealized desire, which is precisely the approach to desire contained in all of the major desiderative conceptions that I surveyed earlier. For we might expect inappropriate and unimportant desires to wash out when we launder preferences through an ideal advisor who represents all aspects of all possibilities fully and vividly in her imagination and makes no mistakes of fact or inference. Moreover, idealization appears to be a normative notion. So even if actual desire lacks normative authority, idealized desire appears to possess it.

Unfortunately, I think that laundering preferences in this way does not help. For one thing, it introduces new problems, not afflicting the basic desiderative model. For all of the idealized desire conceptions appeal to the idea of an appraiser who is fully informed about all of his opportunities and vividly represents their various features, so that he is omniscient with respect to all the experiential and non-experiential aspects of the options available to him. But there are serious questions about the coherence and normative significance of an ideal of omniscient and vivid representation.

An ideal appraiser must evaluate different possible lives. But one question is whether it is possible to combine wildly disparate lives and perspectives into one overall evaluative perspective.[37] The conditions that make a vivid appreciation of one perspective accessible may make a vivid appreciation of a very different perspective inaccessible. For example, the conditions that

[37] See David Sobel, "Full Information Accounts of Well-Being" *Ethics* 104 (1994), 784–810, and Connie Rosati, "Persons, Perspectives, and Full Information Accounts of the Good" *Ethics* 105 (1995), 296–325.

make a naïve or insular perspective accessible, such that one can appreciate its attractions, may make a cosmopolitan perspective inaccessible, and vice versa.

Furthermore, even where diverse possibilities are jointly accessible from a common perspective that does each phenomenological justice, we may wonder whether the effect of vivid representation is normatively significant. One can't rule out the possibility that full confrontation with the facts wouldn't extinguish desire or shape it in ways that one would pretheoretically identify as pathological.[38] Perhaps the weakness of altruistic impulses is typically due to an inadequate appreciation of the suffering of others. But vivid exposure to the enormity of suffering involved in world hunger may overwhelm or de-sensitize appraisers so as to suppress, rather than elicit, sympathetic response. Here, vivid representation produces what are intuitively exactly the wrong normative results.

Moreover, the old problems about normative accommodation that plague the basic desiderative model also apply to idealized desire models. The basic worry, fueled by adaptive considerations, is that desiderative conceptions cannot explain what is wrong with shallow and undemanding lives provided that they are successful in their own terms. While full and vivid information about one's alternatives might extinguish preferences for such lives, it is hard to see how idealization can guarantee this. We can articulate this problem in terms of a dilemma that the ideal appraiser or advisor theory faces.

To be a genuinely desiderative conception of well-being, the ideal advisor theory must take the form of a reductive brand of dispositionalism. But for the dispositionalism to be reductive, the process of idealization must be purely formal or content-neutral. But if the idealization in question is purely formal or content-neutral, then it must remain a brute and contingent psychological fact whether suitably idealized appraisers would care about things we are prepared, on reflection, to think valuable. But this is inadequate inasmuch as we regard intellectually and emotionally rich lives as unconditionally good and intellectually and emotionally shallow lives as unconditionally bad. That is, for a person with the normal range of intellectual, emotional, and physical capacities it is a very bad thing to lead a simple and one-dimensional life with no opportunities for intellectual, emotional, and physical challenge or growth. One's life is made worse, not better, if, after informed and ideal deliberation, that is the sort of life to which one aspires.

Alternatively, we might conclude that anyone who would endorse shallow and undemanding lives simply could not count as an ideal appraiser or

[38] See Allan Gibbard, *Wise Choices, Apt Feelings* (Cambridge, MA: Harvard University Press, 1990), 20.

advisor. Consider, in this context, some of Mill's claims in his defense of the intrinsic superiority of higher pleasures or pursuits over lower ones. He claims that competent judges categorically prefer higher pleasures. But he sees the need to explain this categorical preference for modes of existence that employ their higher faculties, which he does by appeal to a competent judge's sense of his own dignity.

> We may give what explanation we please of this unwillingness [on the part of a competent judge ever to sink into what he feels to be a lower grade of existence] ... but its most appropriate appellation is a sense of dignity, which all human beings possess in one form or other, and in some, though by no means in exact, proportion to their higher faculties. (*Utilitarianism* ii. 6)

But if this is to explain how the life of the contented swine is categorically bad, then it must be that one won't count as an ideal appraiser unless one possesses a sense of dignity that reflects a belief in the value of activities that exercise one's capacities as a progressive being. But such a notion of idealization carries substantive evaluative commitments and is not content-neutral. Suitably idealized desire, understood this way, presupposes, rather than explains, the nature of a person's good. This is one sign that Mill's defense of higher pleasures might be best interpreted as expressing his commitment to a perfectionist conception of happiness, rather than one in which desire or preference plays an ultimate explanatory role.[39] But it also shows why ideal appraiser or advisor conceptions of well-being cannot accommodate our considered evaluative views about categorical goods and bads without relinquishing their distinctive reductive explanatory ambitions.

Finally, I would note that idealization seems unable to address the worry about the normative authority of desire. As long as idealization is a purely formal or content-neutral process, it cannot create normative authority where none existed before. If we lack an explanation about why we ought to care about the satisfaction of desire, as such, regardless of its historical pedigree or content, then we lack an explanation about why we should care about the satisfaction of fully and vividly informed desire, regardless of its historical pedigree or content. Laundering preferences may remove stains, but it does nothing to compensate for poor taste.

7. THE PER SE AUTHORITY OF DESIRE

Before turning to non-desiderative conceptions of practical reason and the good, it is worth considering a different rationale for a desiderative approach

[39] I go a little further in articulating this perfectionist reading of Mill in "Mill's Deliberative Utilitarianism" *Philosophy & Public Affairs* 21 (1992), 67–103.

to practical reason. In an interesting and resourceful article entitled "The Authority of Desire" Dennis Stampe defends the thesis that practical reason can begin in desire because desire enjoys per se rational authority.[40] Stampe rests his case for the authority of desire on an analogy between the way in which perception has authority in theoretical reasoning and the way in which desire has authority in practical reasoning.

Stampe characterizes the difference between beliefs and desires in terms of their different *directions of fit* with the world.[41] On a now familiar version of this view, we might see the difference between beliefs and desires as a special case of a more general difference between representations and pro-attitudes. On this view, representations, such as beliefs, are states of the agent whose content she adjusts to conform to information she receives about the state of the world. By contrast, pro-attitudes, such as desires, are states of the agent on the basis of which she acts to make the world conform to them. We can think of the difference in terms of the response to a perceived mismatch between the content of the intentional state and information about the way the world is. If the state is a belief, the agent tends to respond to such a mismatch by modifying the content of the intentional state to match the way the world is or appears. If the state is a desire, the agent tends to respond to such a mismatch by acting so as to modify the world to conform to the content of the state. On this sort of belief–desire psychology, agents act in order to satisfy their desires based on their beliefs about the world, in particular, their beliefs about the causal means to and necessary conditions of satisfying their desires.[42]

Despite this important difference in the functional profiles of beliefs and desires, Stampe thinks that they play analogous roles in theoretical reasoning and practical reasoning, respectively. Just as what one perceives provides defeasible reason for belief, so too, he claims, what one desires provides defeasible reason for action.[43] Stampe thinks that the parallel is strengthened by seeing desire as directed at the good, as belief is directed

[40] Dennis Stampe, "The Authority of Desire" *Philosophical Review* 96 (1987), 335–81.

[41] Ibid. 354–6.

[42] See e.g. Elizabeth Anscombe, *Intention* (Ithaca, NY: Cornell University Press, 1957), 56; I. L. Humberstone, "Direction of Fit" *Mind* 101 (1992), 59–83; David Velleman, "The Guise of the Good" *Noûs* 26 (1992), 3–26; and Smith, *The Moral Problem*, ch. 4.

[43] However, even on Stampe's proposal, there is a disanalogy between the role of perception in theoretical reason and the role of desire in practical reason. For, on his view, it is the perceiveds, rather than perceivings, that figure as the starting point for perceptual reasoning, whereas it is desirings, rather than the desireds, that figure as the starting points for practical reasoning ("The Authority of Desire," 335–7). I remain somewhat unclear about the bearing of this disanalogy on Stampe's argument.

toward the true. On this conception of desire, it is the perception of things as valuable. This, Stampe concludes, gives desire per se authority for action comparable to the per se authority that perception seems to have for belief. In the case of perception, perception appears to provide pro tanto but defeasible reason to believe. My perceiving something to be the case provides me with per se reason to believe, such that I have some reason to believe it even when I have no other reasons to believe accordingly or even other reasons to disbelieve. Similarly, Stampe claims, my desiring something confers per se authority on bringing it about, such that I have reason to bring it about even when I have no other reason to behave that way or even have other reasons not to behave that way.

Stampe's argument for the per se authority of desire depends on his good-dependent conception of desire. This raises questions about whether his view really assigns desire a fundamental explanatory role in its account of practical reason, inasmuch as desire is treated as the perception of value. However we resolve that issue, Stampe's argument is problematic. We can and should reject the per se authority of desire even if we accept the good-dependence of desire. Moreover, it's doubtful that desire, as such, is essentially good-dependent.

First, the per se authority of desire does not follow from the good-dependence of desire. Even if I do conceive of the objects of desire as good, my desires need not confer reason for action if they are based on false beliefs about the value of the objects of my desire. Stampe says that my desire for something that I otherwise believe or know to be valueless nonetheless gives me pro tanto reason for action just as my perceptual belief that the needle on the gas gauge in my car points to Full gives me reason to believe that my tank is full even if I believe or know my gauge to be broken (e.g. stuck on Full).[44] These are reasons, Stampe says, even if they are outweighed by other reasons or not even good reasons.[45] But though we should recognize pro tanto reasons that fail to be reasons all things considered, I don't know what a reason is that is not a good reason. In particular, I don't see why perception provides reason to believe or why desire provides reason to do when all the other evidence suggests that the perceptual belief is false or that the object of desire is valueless.

Moreover, I think that we should be skeptical of the assumption that desire is essentially good-dependent. No doubt many of our desires are in fact good-dependent in the sense that the desire was generated by or is sustained by the belief that the object of desire is valuable. As we will see shortly, the possibility of good-dependent desire in this sense is essential to agency. But we can admit this without concluding that desire,

[44] Stampe, "The Authority of Desire", 364–5.

[45] Ibid. 342, 364.

as such, is good-dependent. I am inclined to recognize various kinds of good-independent desires. First, I recognize the possibility of desires for things the agent regards as thoroughly bad, as might be the case with the self-loathing drug addict or the self-loathing pedophile. Second, I recognize the possibility of desires produced by sub-rational processes, such as hypnosis or suggestion, and these seem not to be produced or sustained by the thought that the objects of desire are good. Finally, I recognize the possibility of desires in animals and small children where these states are apparently not mediated by value concepts for the simple reason that the subjects themselves seem to lack value concepts.

These possibilities motivate skepticism about the assumption that desire, as such, is good-dependent. However, it would be nice to have an account of desire that explained what desire is such that it need not be good-dependent. But we have the beginnings of such an account in the familiar idea, which Stampe himself endorses, that desire is an intentional state with a specific functional profile given by its direction of fit to the world. Desires are states of the agent or subject in which she tends to adjust the world so as to make it conform to the content of the state. Genuine agents may well have such states as the result of beliefs about the way in which the world ought to be, but actors who are not agents, such as brutes and small children, and even genuine agents can have states that dispose them to change the world so as to conform to the content of these states independently of any belief about the value of the world so represented. Insofar as Stampe's defense of the per se authority of desire depends upon this good-dependent conception of desire, we should reject it.

8. NORMATIVE PERFECTIONISM

Despite their promise to reconcile resonance and fallibility, idealized desire conceptions of practical reason and the good fail to deliver a satisfying account of normativity. In particular, they score poorly along the dimension of normative accommodation, and they lack a clear rationale for the normative authority of desire. We might consider two apparently different ways forward.

We saw that Mill achieves normative accommodation, explaining what is objectionable about shallow and undemanding lives, by appeal to a conception of ideal desire in which ideal appraisers are guided by their sense of their own dignity as progressive beings. On this reading, Mill is appealing to good-dependent desires. He needn't assume, as Stampe does, that all desire is good-dependent, only that it can be good-dependent. This suggests that we might understand well-being in terms of objective goods.

One form of objectivism is a list of objective goods, such as knowledge, beauty, achievement, friendship, and equality.[46] Such a list may seem the only way to capture the variety of intrinsic goods. But if it is a mere list of goods, with no unifying strands, it begins to look like a disorganized *heap* of goods.[47] One objective conception of the good that goes beyond a mere list of goods is *perfectionism*. There is a venerable perfectionist tradition, common to Aristotle, John Stuart Mill, and T. H. Green, among others, that identifies a person's good with the perfection of her nature and, in particular, with the development of her deliberative competence and the exercise of her capacities for practical deliberation.[48]

Not only might we understand well-being in terms of objective goods. We might understand practical reason in terms of objective goods. What we have reason to do, on this view, is what is objectively good. This sort of view might explain the good in terms of the personal good, representing things as good insofar as they contribute to people's well-being, or it might recognize goods that are fundamentally impersonal. Such a view would embrace the Reason–Value Link, but it would treat value as the explanatorily more basic notion, and provide an objective conception of value. We might treat any such good-dependent conception of practical reason as a *teleological* conception. But this kind of teleology can be substantively ecumenical. In particular, it need not presuppose consequentialism, because central among the objective goods may be moral goods, and rational action can involve either *honoring* or *promoting* objective values.[49] Moore is one prominent example of someone who embraces this sort of good-dependent conception of practical reason, but there are other proponents as well.[50]

Any conception of well-being or practical reason that appeals to objective value is likely to fare well along the dimension of normative accommodation,

[46] Moore endorses an objective list in *Principia Ethica*, ch. 6, as does Ross in *The Right and the Good*, p. 140. Derek Parfit discusses such theories sympathetically in *Reasons and Persons* (Oxford: Clarendon Press, 1984), 493–502.

[47] This is like the criticism, made by Joseph, among others, that the intuitionist's objective list of right-making factors amounts to nothing more than an "unconnected heap" of obligations. See H. W. B. Joseph, *Some Problems in Ethics* (Oxford: Clarendon Press, 1931), 67. Just as a suitably structured or unified theory of the right avoids Joseph's heap objection, so too a suitably structured or unified theory of the good avoids this heap objection.

[48] A vigorous contemporary statement of perfectionism is Thomas Hurka, *Perfectionism* (Oxford: Clarendon Press, 1993).

[49] I borrow this useful distinction from Philip Pettit, "Consequentialism" reprinted in *Consequentialism*, ed. S. Darwall (Oxford: Blackwell, 2002).

[50] A good-dependent conception of practical reason is at work in Thomas Hurka, *Virtue, Vice, and Value* (Oxford: Clarendon Press, 2002) and Donald Regan, "The Value of Rational Nature" *Ethics* 112 (2002), 267–91 and in unpublished work by Derek Parfit and by Diane Jeske.

because it can appeal to whatever values are necessary to vindicate our intuitions about well-being and practical reason. However, not all teleological conceptions provide a rationale for the normative authority of objective values.

This seems especially true for many lists of objective goods. For example, why should beauty, knowledge, friendship, or equality engage my will? Of course, if it is a plausible list, most of us will already care about the items on the list. But to have normative authority, we must be able to explain why we should maintain our concern for items on the list if we already care about them and why we should care about items on the list if we do not yet. Of course, if the Reason–Value Link is correct, then we do have reason to be concerned about and promote anything that is good. And if the Reason–Well-being Link is correct, then we have reason to be concerned about something for someone's own sake just insofar as it is good for her. But if we make normative authority a condition of the good or well-being, then we ought to be able to explain for any candidate good how it enjoys normative authority. Standard lists of objective goods do not meet this demand.[51]

But perfectionist conceptions of the good may not be well positioned to address the issue of normative authority either. Perfectionists identify the good with perfecting one's nature. This might suggest that a perfectionist should base her conception of the good on claims about what is distinctive or essential about human nature. Some perfectionists understand the appeal to human nature as an appeal to a *biological essence*. But it is hard to find capacities that we have as a biological species that are essential and whose exercise provides reason for concern. For example, perfectionist ideals often prize creative achievements that exercise the agent's rational capacities in some way and condemn shallow and undemanding lives. But it is hard to see how this sort of perfectionist content could be justified by appeal to a biological essence. Genotypic and phenotypic diversity make it difficult to see how there could be a substantive species essence, especially one in which rational capacity figures prominently. One could appeal to the reproductive closure of the species, so that the species includes as members all and only individuals capable of breeding with other members of the species. But there are many members of the species human being that satisfy this

[51] I believe that the normative inadequacy of the simple appeal to objective values also animates Christine Korsgaard's criticisms of what she calls "substantive moral realism" in *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996), 28–48. While a normatively adequate account of objective values or moral requirements must explain why we should care about value or moral requirements, I don't see anything inherent in objective values or moral realism that prevents addressing this legitimate explanatory demand.

reproductive criterion that lack basic cognitive and affective capacities that we think of as normal and desirable. The biological perfectionist must claim that these individuals are abnormal. But we can imagine circumstances in which they need not be abnormal in a statistical sense. If they are abnormal, it appears that it must be in some normative sense. But this concession would defeat the project of deriving perfectionist norms from a biological essence.[52]

Once one recognizes the legitimacy of the question about normative authority, it can seem difficult to answer. For any putative standard of reason or value, we can always ask why we should care about conforming to that standard. This difficulty explains, I think, the appeal of a broadly Kantian approach that seeks a standard rooted in rational agency itself. For the demands of any such standard would be rooted in practical reason itself. To some minds, the Kantian appeal to agency and practical reason is fundamentally opposed to teleological approaches. However, I think that we can reconcile the Kantian insight with a form of perfectionism.

An important strand in the perfectionist tradition understands the appeal to human nature, not in biological terms, but in normative terms. I believe that this sort of *normative perfectionism* is evident in Aristotle, Mill, and Green. But I will focus on Green's version, as articulated in his *Prolegomena to Ethics*,[53] because the Kantian influence on his perfectionism is clearest. Green conceives of persons as agents who are responsible for their actions. Non-responsible agents, such as brutes and small children, act on their strongest desires; if they deliberate, it is only about the instrumental means to the satisfaction of their desires (§§ 86, 92, 96, 122, 125). By contrast, responsible agents must be able to distinguish between the intensity and authority of their desires, deliberate about the authority of their desires, and regulate their actions in accordance with their deliberations (§§ 92, 96, 103, 107, 220). This requires one to be able to distinguish oneself from particular appetites and emotions—to distance oneself from them—and to be able to frame the question what it would be best for one on the whole to do. So a person acts not simply on desires or passions but on the basis of ought judgments.

These deliberative capacities are essential for responsible willing and action, but they do not yet tell us what separates a good and bad will (§ 154). However, Green argues that it is the very capacities that make moral

[52] Philip Kitcher raises some related difficulties for Hurka's appeal to a biological essence in "Essence and Perfection" *Ethics* 110 (1999), 59–83.

[53] T. H. Green, *The Prolegomena to Ethics* [originally published 1883], ed. D. Brink (Oxford: Clarendon Press, 2003).

responsibility possible in the first place that determine the proper end of deliberation (§ 176). Responsible action involves self-consciousness and is expressive of the self. The self is not to be identified with any desire or any series or set of desires; moral personality consists in the ability to subject appetites and desires to a process of deliberative endorsement and to form new desires as the result of such deliberations. So the self essentially includes deliberative capacities, and if responsible action expresses the self, it must exercise these deliberative capacities. This explains why Green thinks that the proper aim of deliberation is a life of activities that embody rational or deliberative control of thought and action (§§ 175, 180, 199, 234, 238–9, 247, 283).

This sort of normative perfectionism promises to address questions about the normative authority of the good. For Green's defense of self-realization makes the content of the good consist in the exercise of the very same capacities that make one a rational agent, subject to reasons for action, in the first place. This promises to explain why a rational agent should care about the good conceived in terms of self-realization.

But why should we think that the exercise of practical deliberation must favor lives that embody or exercise rational nature? Green, like Kant, is interested in the question what one would care about insofar as one is rational. Consider an analogy. Insofar as one is a wine connoisseur, there are determinate things that one cares about. One cares about developing general wine competence (e.g. knowledge about wine varietals, conditions for growing and harvesting grapes, and methods of fermenting and aging wines) and about the consumption and appreciation of fine wines by themselves and as parts of meals. Similarly, insofar as one is a rational agent, one cares about developing one's deliberative competence and sensitivity to reason and one chooses environments, projects, and activities that allow scope for deliberative control of thought and action. In this way the exercise of practical reason can be the object of practical reason, much as the exercise of wine connoisseurship can be the object of the wine connoisseur. This addresses the issue of content, but not the issue of authority. But whereas assuming the perspective of the wine connoisseur appears rationally optional, assuming the point of view of practical reason cannot be comparably optional. Anything that practical reason, as such, would endorse necessarily enjoys normative significance.

This justification of self-realization also explains why Green treats the imperative of self-realization as a categorical imperative. Like Kant, Green seeks an account of the agent's duties that is grounded in her agency and does not depend upon contingent and variable inclinations. The goal of self-realization, Green thinks, meets this demand.

> At the same time, because it [self-realization] is the fulfilment of himself, of that which he has in him to be, it will excite an interest in him like no other interest, different in kind from any of his desires and aversions except such as are derived from it. It will be an interest as in an object conceived to be of unconditional value; one of which the value does not depend on any desire that the individual may at any time feel for it or for anything else, or on any pleasure that … he may experience. … [T]he desire for the object will be founded on a conception of its desirableness as a fulfilment of the capabilities of which a man is conscious in being conscious of himself. … [Self-realization] will express itself in [the] imposition … of rules requiring something to be done irrespectively of any inclination to do it, irrespectively of any desired end to which it is a means, *other then this end, which is desired because conceived as absolutely desirable.* (§ 193)

Because the demands of self-realization depend only on those very deliberative capacities that make one a responsible agent, they are categorical imperatives.

9. INSTRUMENTALISM, RESONANCE, AND PLURALISM REVISITED

Normative perfectionism promises to succeed along dimensions that desiderative conceptions of practical reason and the good fail. It addresses concerns about the normative authority of perfectionist goods better than desiderative conceptions address parallel questions about the normative authority of desire. Moreover, normative perfectionism is well positioned to accommodate and explain the evident fact that intellectually and emotionally rich lives are unconditionally good and intellectually and emotionally shallow lives are unconditionally bad for a person with the normal range of intellectual, emotional, and physical capacities. But if we are to take normative perfectionism seriously, it must have something plausible to say about the considerations that made us take desiderative conceptions seriously in the first place.

The instrumentalist makes several claims. One claim is that an agent has reason to take means or necessary conditions conducive to satisfying her desire, or at least her ultimate desires. Many people seem to assume that instrumental rationality, so conceived, is a part of practical reason. What separates the instrumentalist from others is that she assumes that this is not only a part but the whole of practical reason. The instrumentalist accepts this stronger claim, because she believes that we can reason only about means and not about desires, in particular, not about ultimate desires.

We can reject the stronger claim that practical reason is purely instrumental, because we can in fact reason about both the value and the authority of desires. Indeed, a great many desires are judgment-dependent in the sense that they are predicated on a belief in the value or appropriateness of the object of desire and, hence, are sensitive to reappraisals of the judgments of value and worth. If the perfectionist is right to locate the normative authority in the value of rational agency, then we can ask if a given set of desires is appropriate for us given the sort of beings we are. In particular, we can ask if a particular set of commitments is appropriate for agents who are capable of regulating their lives in accord with practical reason.

But we should reject even the weaker claim that instrumental rationality is a part of practical reason, provided we understand instrumental rationality as claiming that one always has a reason to adopt means or necessary conditions to the satisfaction of one's (ultimate) desires. For we have rejected the proposition that desire, as such, has any normative authority. If so, we must deny that instrumental rationality, so conceived, is even part of the correct account of practical reason. If this conclusion seems like throwing the baby out with the bath water, it is probably because we fail to distinguish instrumental rationality, so conceived, from a different conception of instrumental rationality that is genuinely indispensable. On this alternative conception, one has reason to adopt causal means to and necessary conditions of that which one *already has reason* to do. This conception of instrumental rationality is really a conception of *derivative* or *conditional* rationality.[54] It is in no way reductive, and makes no appeal to desire. Accepting instrumental rationality in this sense as part of the truth about practical reason concedes nothing to Humean instrumentalism.

The Humean instrumentalist also believes that reason can only be the slave of the passions. But practical reason, we just said, can judge some commitments appropriate and others inappropriate. But then one would expect desire to be capable of responding to reason. Judging a potential commitment appropriate tends to awaken desire, and judging an existing commitment appropriate tends to sustain desire. By contrast, judging a potential commitment inappropriate tends to produce aversion, and judging an existing commitment inappropriate tends to weaken desire.

These familiar observations are reinforced if we adopt a version of the sort of belief–desire psychology often associated with Humean moral psychology. On this view, as we have seen (§ 7 above), intentional action is viewed as the product of representational states, such as belief, and pro-attitudes, such as desire, which display different directions of fit with the world. On this sort of belief–desire psychology, agents act in order to satisfy their desires based on their beliefs about the world, in particular,

[54] Cf. Darwall, *Impartial Reason*, 79.

their beliefs about the causal means to and necessary conditions of satisfying their desires. But, on this sort of psychology, we can also understand how normative beliefs would tend to influence desire. For normative beliefs are beliefs about how the world should be. But if desires are precisely states that tend to make agents modify the world in accordance with their content, then we should expect normative beliefs normally to affect desires.[55] This is Green's view (§§ 130–6). He accepts belief–desire psychology, because of their different directions of fit, and argues that for this reason desire can be responsive to ought judgments. This shows how one can accept the Humean dictum that action depends on desire without accepting the Humean dictum that reason can only be the slave of the passions.

But if reason can be the master of the passions, then we can see how we can accept the resonance constraint without endorsing desiderative conceptions of practical reason or the good. Williams appeals to the idea that practical reason must be capable of resonating with agents to accept the internalist claim that practical reason must be relativized to elements of the agent's subjective motivational set. Because he assumes that something is capable of motivating someone only if it is conducive to satisfying antecedent desire, he concludes that reasons for action must be relativized to the agent's antecedent desires. We saw how idealized desire conceptions of well-being can be motivated by a similar argument. But if desire can be responsive to reason, then we can accept the demands that practical reason and well-being be resonant without concluding that practical reason or well-being be relativized to antecedent desire. If we accept belief–desire psychology, then desire is necessary for resonance. But desire can and will normally be consequential on recognition of reasons for action or value. This means that motivational capacity exerts no real constraint on the content of practical reason or well-being. Someone who recognizes imperatives of self-realization as imperatives of practical reason will, for that reason, tend to desire to conform to these imperatives. Such imperatives would be resonant and capable of motivating, even though they are not grounded in desire. The fact that desire can be responsive to

[55] To say that normative beliefs can and normally do influence desire is not to say that normative beliefs have such influence necessarily. Other things being equal, normative beliefs have conative influence. But other things need not be equal if there is some relevant form of psychological interference. In some cases of weakness of will, normative beliefs apparently motivate but provide insufficient motivation. In other cases of weakness of will, normative beliefs may not motivate at all. This second sort of weakness of will will be selective if the interference is intermittent; it will be systematic if the interference is systematic. Depression might produce selective weakness of will, but damage to the prefontal lobe of the cerebral cortex (as in the famous case of Phineas Gage) might produce systematic weakness of will.

reason means that the resonance constraint does not favor desire-dependent conceptions.

What it implies about internalism depends on how we understand that doctrine. If we understand internalism more generically as the view that normative facts must be capable of motivating the agent, then the proper moral to draw is that internalism follows from resonance but that it is a fairly ecumenical constraint and does not support a desiderative conception. Alternatively, if we understand internalism, as Williams sometimes does, as committed to the more sectarian claim that reason or value must be relativized to antecedent desire, then we should deny that resonance implies internalism and recognize that externalist conceptions of reason and value can meet the resonance constraint.

Finally, we should revisit the pluralist rationale for desiderative conceptions. Such conceptions seemed plausible, because they promised to sustain an attractive sort of pluralism about the content of reason and value. By contrast, objective conceptions of reason and value seemed hostile to pluralism. But this pluralistic rationale is misguided. First, objective conceptions can recognize a plurality of equally or incommensurably reasonable and good activities. This would certainly be true of conceptions of reason or value based on a list of objective goods. Activities and lives could combine different goods in different amounts, yielding the result that quite different activities and lives could be equally or comparably worthwhile. Moreover, the normative perfectionist can recognize that there is a diversity of activities and lives that exercise one's capacities for practical reason. The artisan who makes important decisions about the organization of her craft and the production and distribution of her product exercises deliberative control within her life just as much as the intellectual or artist. So pluralism is not the exclusive province of desiderative conceptions.

Morever, it matters how one justifies pluralism. Desiderative conceptions of practical reason and value are not just pluralist, but relativist. They are relativist, because they are content-neutral, placing no substantive constraints on the content of practical reason or well-being. But we saw that relativism faces problems of accommodation. Most of us are not prepared, on reflection, to judge that there are no substantive constraints on practical reason or the good. In particular, we said that shallow and undemanding lives are necessarily bad for those with a normal range of talents and capacities. It is a vice of desiderative conceptions that they derive pluralism from the more extreme and unsustainable commitment to content-neutrality. It is a virtue of objective conceptions that they can explain pluralism without the unsustainable commitment to content-neutrality. In particular, it is a virtue of normative perfectionism that it endorses pluralism while explaining what is wrong with shallow and undemanding lives, even when they are successful in their own terms.

10. THE SIGNIFICANCE OF CHOICE

In rejecting the content-neutrality of desiderative conceptions of practical reason and the good, we have rejected the normative authority of desire, as such. On the one hand, it seems right that the mere existence of a desire, regardless of its historical pedigree or content, has no normative significance. On the other hand, it certainly does seem in a great many cases that the fact that an agent wants something is a reason for her to care about and pursue it and often a reason for others to care about her caring about and pursuing it. How can we account for this?[56]

We ought to distinguish between the significance of choice and of desire. It is choice, rather than desire, as such, that has normative significance. Non-responsible actors have and act on desires. What makes someone a person or an agent is that she has the capacity to assess her options and act for reasons. She is not compelled to act on desire but can step back from existing desires, assess them, modify them, and form new desires. Kant appeals to this capacity to set ends as the source of normative significance. In *Religion within the Boundaries of Mere Reason* he writes, "Freedom of the power of choice has the characteristic, entirely peculiar to it, that it cannot be determined to action through any incentive *except so far as the human being has incorporated it into his maxim* (has made it into a universal rule for himself, according to which he wills to conduct himself)".[57] As Henry Allison has interpreted Kant's *incorporation thesis*, it implies that inclination or desire is not itself a reason for action but can become one through being incorporated into a maxim expressing a judgment about the principles on which one should act.[58] Green, who develops some Kantian claims within a perfectionist framework that treats moral personality as the source of value, distinguishes desire, as such, which has no normative significance, from the will, which does (§§ 139–42). An agent acts not simply on appetites or passions but on the basis of ought judgments or a

[56] Stephen Darwall raises this question in "'Because I Want It'" *Social Philosophy & Policy* 18 (2001), 129–53, though he provides a different answer than I will. Though our answers are different, I hope that they are not incompatible.

[57] Immanuel Kant, *Religion within the Boundaries of Mere Reason* [originally published 1793], ed. A. Wood and G. di Giovanni (Cambridge: Cambridge University Press, 1998), 6: 24.

[58] Henry Allison, *Kant's Theory of Freedom* (Cambridge: Cambridge University Press, 1990), 39–40.

conception of goods. Green thinks that when an agent endorses a course of action as a result of such judgments, this affects her desires; it can weaken or strengthen existing desires and create new desires. He identifies the will with post-deliberative desire or desire that is the product of deliberative endorsement. Only the will has genuine normative significance.

This perfectionist conception of the significance of choice or post-deliberative desire may sound remarkably like an informed desire conception of practical reason or the good. But notice some important differences. First, an informed desire conception defines normatively significant desire by appeal to a *counterfactual* condition. Is the desire one which *would* emerge from some suitable idealization of the agent's current desires? By contrast, the perfectionist conception appeals to an *historical* condition. Is the desire one which was produced or is sustained by a suitable kind of deliberation? Also, deliberation need not be ideal in order to have normative significance for the perfectionist; the normative significance of one's choices can be proportional to the amount of deliberation that produced them or sustains them. Moreover, whereas the informed desire conception appeals to a conception of idealization that is explicitly non-evaluative, the perfectionist conception appeals to an essentially evaluative conception of deliberation.

This perfectionist defense of the significance of choice will be of limited help if deliberative endorsement is a rare occurrence, making unusual demands on agents. But exercising one's will is not an exceptional feat accomplished only when one takes time consciously to survey and evaluate the alternatives and their grounds. One exercises one's will when one acts on standing principles and commitments that reflect those principles and when one concludes there is no special need or justification for renewed deliberation. One also exercises one's will when one acts on desires that are sustained by reflective endorsement, even if they did originate in reflective endorsement.

Choice is an exercise of the will and, as such, expresses agency. Because the perfectionist treats agency as the source of reasons for action and value, she regards choice as normatively significant. Indeed, if the will can be identified with desire that is the product of rational endorsement, then the perfectionist can explain why a significant class of desires has normative significance, even if she denies normative significance to desire as such.

11. WEIGHING CHOICE AND THE CONTENT OF CHOICE

But even if choice has significance, it is not the only thing that has significance. To treat choice as the only thing of significance would yield not just pluralism, but relativism. We would have serious problems of normative accommodation and would not have improved much on desiderative conceptions of reason and value. Any plausible conception of reason or value must recognize substantive constraints on the content of choice—constraints on which choices are reasonable, appropriate, or valuable.

For present purposes, I would like to remain agnostic about the precise source and nature of these constraints on the content of choice. In particular, I won't try to decide here between two different conceptions of the source of such constraints.

On a *monistic* view, the source of these constraints is the same as the source of the significance of choice, namely agency. Kant is usually read as this sort of monist. On one reading of Kant's *Groundwork*, he begins with the idea that moral requirements must be inescapable, which requires that they be represented as categorical, rather then hypothetical, imperatives (414, 416, 420, 425).[59] But this means that moral requirements must apply to people insofar as they are agents, that is, insofar as they have capacities for practical reason to set ends (408, 426). This is the source of both the Universality and Humanity formulations of the Categorical Imperative. It implies that moral requirements must have a sort of universality such that one may act only on maxims that one can will to be universal law (421). But it also sets the stage for recognizing the value of rational nature itself (428). For the one thing that one would value just insofar as one is rational is rational nature itself. This means that one should act only in ways that respect humanity or rational nature, whether in one's own person or that of others, as an end in itself and never merely as a means (429). In this way, rational nature is supposed to constrain and guide the content of the will.

Green, as we saw, is another monist who thinks that rational nature is not only a condition of the will and responsible action but also sets the proper object of the will (§ 176). He thinks that responsible willing requires consciousness of oneself as distinct from one's appetites and passions and as able to set ends. If responsible willing must aim to express the self, then it should aim to develop and exercise well those very capacities for setting ends. This requires undertaking projects that allow scope for the agent's deliberative control of his own fate. For reasons that defy easy reconstruction, Green also thinks that self-realization can only take place when an agent recognizes the reality of other agents, which leads him to claim that self-realization requires each agent to aim

[59] Immanuel Kant, *Groundwork for the Metaphysics of Morals* [originally published 1785], tr. M. Gregor (Cambridge: Cambridge University Press, 1996).

at good that is common to himself and others. In this way, he traces a perfectionist path from rational agency as condition of responsibility to something like Kant's Humanity formulation of the Categorical Imperative.[60]

However, some may doubt the adequacy of these monistic conceptions of the normative constraints on the content of choice. One might question whether one can really generate a constraint to treat all rational agents as ends in themselves or to promote the common good from the assumptions about agency required for responsible action. Alternatively, one might concede this but question whether the constraint to treat people as ends or to promote a common good exhausts the constraints and guidance about the content of choice that we want to recognize. One might think that an adequate account of the constraints on the content of choice must recognize values other than rational agency—objective values, sensitivity to which should guide autonomous choice.[61]

Whether monists or pluralists about constraints on the content of choice, we need to ask a question about how to weigh the significance of the fact of choice and the significance of the content of choice. In particular, one wants to know whether the fact of choice should have normative significance when the content of the choice lacks significance. Do a person's choices give her reason for action when they are substantively bad? Is it good for her for her choices to be successful even when her choices are substantively inappropriate? Let's consider briefly some different models.

Dualism of Choice and Content

In cases where there are substantive but comparatively minor problems with the content of choice, it is tempting to recognize the value of the choice itself. Most us make decent but non-optimal choices about many things, including career and friends. Surely, one has reason to act on such choices, and we might judge one's success in life at least in part relative to the content of such choices. If we generalize this intuition, we might recognize the choice itself and the content of choice as independent and potentially conflicting values. On this model, if one's choice is sufficiently substantively bad, this can outweigh, but not cancel, the value of the choice itself.

[60] I try to reconstruct and assess some aspects of Green's perfectionist defense of the common good in *Perfectionism and the Common Good: Themes in the Philosophy of T. H. Green* (Oxford: Clarendon Press, 2003).

[61] See e.g. Regan, "The Value of Rational Nature".

Choice Limited by Content

In cases where the choice is substantively deeply flawed relative to other available options, one might be tempted to deny any significance to the choice itself. Suppose that someone chooses to sell himself into slavery and has no very good justification for this choice. It was not forced on him by economic necessity; he just liked the idea of belonging to someone. The monist will have no problem explaining how this choice is substantively bad—it is an exercise of agency that abdicates agency. Pluralists may have other objections as well. One might be tempted not to accord any significance to this choice in determining what the person has reason to do or what would contribute to his well-being. Generalizing this response, one might say that if the choice is substantively problematic, then the choice itself has no significance. On this model, the substantive merits of the choice condition or limit the significance of the choice itself.

Choice Limited by Threshold Content

But this second model makes the significance of choice depend upon choice with ideal content. It seemed a virtue of the first model that it avoided this result. A compromise solution would be to modify the second model so that it accords significance to choice itself only when a threshold of substantive merit has been reached. Above the threshold, choice itself matters. Below the threshold, choice itself does not matter. But this third model leaves awkward questions often associated with thresholds. Where exactly do we locate the substantive threshold? And how can choice matter just above the threshold and not at all below the threshold?

Choice as Proportional to Content

An obvious response to worries about thresholds is to go scalar, claiming that the magnitude of the value of choice itself is proportional to the magnitude of the value of the substance of the choice. We can explain why choice itself is significant when the substance of choice is acceptable but not optimal. Moreover, we can explain why choice itself has little, if any, significance when the content of choice doesn't either.[62]

[62] Indeed, the scalar model presumably implies that the fact of choice has negative value when the merits of the choice do. Is this implication acceptable?

These are just four of the most obvious models for relating the significance of choice and content. Of these, the scalar model looks most promising. One might well prefer a model for weighing the significance of these two variables that had a deeper philosophical rationale, but at least this model has the virtue of initial plausibility. Until we identify a better or more theoretically satisfying model, we might defeasibly embrace the scalar model.

This page intentionally left blank

2

Fitting Attitudes and Welfare

Chris Heathwood

The purpose of this paper is to present a new argument against so-called fitting-attitude analyses of intrinsic value, according to which, roughly, for something to be intrinsically good is for there to be reasons to want it for its own sake. The argument is indirect. First, I submit that advocates of a fitting-attitude analysis of value should, for the sake of theoretical unity, also endorse a fitting-attitude analysis of a closely related but distinct concept: the concept of intrinsic value *for a person*, that is, the concept of *welfare*. Then I argue directly against fitting-attitude analyses of welfare. This argument, which is the focus of the paper, is based on the idea that whereas whether an event is good or bad for a person doesn't change over time, the attitudes there is reason to have towards such an event can change over time. Therefore, one cannot explain the former in terms of the latter, as fitting-attitude analyses of welfare attempt to do.

1. FITTING-ATTITUDE ANALYSES OF VALUE

1.1. Background

G. E. Moore famously argues (1903a, § 13) that the property of being intrinsically good is unanalyzable and that the predicate 'is intrinsically

Versions of this paper were presented at the Center for Values and Social Policy at the University of Colorado at Boulder in April 2006, at the Third Annual Metaethics Workshop at the University of Madison–Wisconsin in September 2006, and at the Pacific Division Meeting of the American Philosophical Association in San Francisco in April 2007. Thanks to those in attendance for helpful discussion, especially Gunnar Björnsson, Tom Carson, Steve Finlay, Pat Greenspan, Christian Lee, Don Loeb, Kris McDaniel, Bob Pasnau, Doug Portmore, Geoff Sayre-McCord, Mark Schroeder, Caj Strandberg, and Michael Tooley. Special thanks to my APA commentator, Bana Bashour. Ben Bradley, Fred Feldman, Jens Johansson, Mark LeBar, Brad Skow, and Stephan Torre each read a version of the paper and provided comments that helped me improve it; I am

good' is indefinable. If he is right, then friends of intrinsic value are in a bit of a bind: we believe in some property—indeed, we hold it to be a very important one, one significantly impacting how we are required to behave—but we've never seen it, we can't tell you what its nature is, and we can't define the word we use to express it. Wouldn't it be better if we could do this?

Some friends of intrinsic value have tried to do this, and in a way against which Moore's arguments lose some of their force.[1] Moore's open question argument is most forceful against naturalistic analyses of value, but these philosophers have proposed a certain kind of not-necessarily-naturalistic analysis: the *fitting-attitude analysis of value*. According to this view, to say that something is intrinsically good is to say, roughly, that we should desire it, or that we have reason to desire it, or to favor it, or to have some other pro-attitude towards it. The theory explains *value*, an axiological notion, in terms of some deontic notion, such as *obligation, fittingness*, or *reasons*. But it is "less reductive" than naturalistic analyses, since it doesn't attempt to reduce away the normativity of value. The analysans of a fitting-attitude theory contains a normative notion.[2]

Fitting-attitude theories are metaethical theories of *what it is* for something to be good, not normative theories of what things are good; so they should be compatible with any theory of the latter. When hedonists tell us that pleasure and pleasure alone is good, what the hedonists are saying, according to fitting-attitude theorists, is that pleasure and pleasure alone is fit to be desired. Fitting-attitude theories are, in a sense, *response-dependent* theories, but the relevant responses are not the ones we do or would have, but the ones we *should* have.

The earliest prominent advocate of this theory is usually thought to be Franz Brentano (1889). But Sidgwick endorses a version of the view in at least the third edition of *The Methods of Ethics* (1884: 108): "I

grateful to each of them. Finally, I'd like to thank those who contributed to a discussion of some of these topics on the weblog PEA Soup.

[1] Not all such philosophers are in fact friends of intrinsic value (if by 'intrinsic value' we mean value that supervenes on intrinsic nature). Some are more interested in theorizing instead about *final value*—the value that something has as an end, a value which it may turn out does not supervene on intrinsic nature (for discussion see Zimmerman 2004). Nothing in this paper turns on the debate over whether intrinsic value or final value should be the focus of axiology. For the sake of familiarity and historical continuity, I continue to use the term 'intrinsic value'. I also assume, though nothing depends on it, that states of affairs rather than concrete objects are the fundamental bearers of value.

[2] Fitting-attitude analyses are, however, compatible with a naturalistic reduction of the deontic notion they contain, a notion that, for the purposes of this paper, will be taken as primitive.

cannot, then, define the ultimately good or desirable otherwise than by saying that it is that of which we should desire the existence if our desires were in harmony with reason … ."[3] Many other philosophers have since defended, or at least seriously considered, more or less the same idea.[4] Here is how C. D. Broad (1930: 238) puts it: "I am not sure that 'X is good' could not be defined as meaning that X is such that it would be a fitting object of desire to any mind which had an adequate idea of its non-ethical characteristics." Though he doesn't mention Broad, T. M. Scanlon (1998: 97) has recently endorsed a view very much like Broad's. Scanlon's so-called buck-passing account of value has been receiving a lot of attention.

It is fitting to want such an analysis to be true. Fitting-attitude analyses do at least the following valuable things:

i. They reduce. I believe in value, and I believe in reasons to have certain attitudes. This view reduces the one to the other. So instead of having to believe in both of these things as fundamental, irreducible features of the world, I have to believe, fundamentally, in at most only one—reasons—and I get the other for free.

ii. They demystify. Some people are skeptical of intrinsic value, but fewer people are skeptical of the notion that we ought to have certain attitudes. We all, for instance, believe that there are certain things we ought to believe. So we are familiar with the idea that certain attitudes are required. If we see that facts about intrinsic value are nothing more than facts about what attitudes there are reasons to have, we may no longer find intrinsic value so mysterious.

iii. They explain why it is confused to wonder whether there is any reason to promote the good. If something is intrinsically good, then, given

[3] So Moore was evidently mistaken in taking Sidgwick to have agreed with him about the indefinability of goodness; Moore wrote, " 'Good,' then, is indefinable; and yet, so far as I know, there is only one ethical writer, Prof. Henry Sidgwick, who has clearly recognised and stated this fact" (1903a: § 14). Perhaps Moore was misled by the fact that Sidgwick held that "The term 'ought,' as used in moral judgments … is unanalysable" (1884: p. xvi). See also Sidgwick 1884: 27 or 1907: 25.

[4] See Brentano 1889: 18, John Maynard Keynes 1905, C. D. Broad 1930: 238, Richard Brandt 1946: 113 and 1959: 159, A. C. Ewing 1947: 152, J. O. Urmson 1968: 58–9, John McDowell 1985: 118, Roderick Chisholm 1986: 52, David Wiggins 1987 : 189, Allan Gibbard 1992: 980, Elizabeth Anderson 1993: 1–2, Richard Kraut 1994: 45, Noah Lemos 1994: 12, T. M. Scanlon 1998: 95, 97, Thomas Carson 2000: 158–9, Michael Zimmerman 2001: ch. 4, Derek Parfit 2001: 20, Wlodek Rabinowicz and Toni Rønnow-Rasmussen 2004, Jonas Olson 2004, Jussi Suikkanen 2004, and Philip Stratton-Lake and Brad Hooker 2005. Rabinowicz and Rønnow-Rasmussen 2004: 394–400, from which some of the above citations are drawn, contains a helpful historical summary. See also Carson 2000: 160–2, from which some other of the above citations are drawn. Keynes 1905 is unpublished; for discussion, see Baldwin 2006.

a fitting-attitude analysis, it follows automatically that there is reason to want it to occur.

1.2. A Formal Problem: The Wrong Kind of Reasons Problem

Fitting-attitude analyses face a certain sort of formal objection.[5] I call this objection 'formal' because it seems to me it doesn't strike at the heart of the basic idea of fitting-attitude theories; it instead guides us in how best to formulate the view, or in how best to understand the view we have already formulated.

One sort of problem case involves *desiring a bad thing for the bad thing's good effects.* According to fitting-attitude theories, to say that something is intrinsically good is to say that there is reason to want it to occur. But suppose I have a cut that needs to be cleaned with alcohol. Suppose that only if I am feeling the sting of alcohol in the cut is it being cleaned. I therefore want to be feeling this painful sensation. Given that avoiding infection requires feeling this sensation, it is sensible to want to be feeling it. But then the fitting-attitude theory implies the absurd claim that this pain is *intrinsically* good.

This case teaches us that fitting-attitude analyzers have *intrinsic desire*, or desire for something for its own sake, in mind when they say that to be intrinsically good is to be fit to be desired. A fitting-attitude theory restricted to intrinsic desire doesn't imply that my pain is intrinsically good (since I don't have reason to want it for its own sake, just for what will accompany it).

Another sort of case involves *desiring a bad thing for the desire's good effects*. Suppose a demon offers to end world poverty if only you will intrinsically desire something bad, like that Tiger Woods gets a terrible headache later today. Surely you ought to get yourself to desire for its own sake (assuming that's even possible) that Tiger Woods gets the headache later today. But still, Tiger's having the headache is not intrinsically good.

This case teaches us that that fitting-attitude analyzers have *object-given reasons*, rather than *attitude-given reasons*, in mind when they say that something is good just in case there is reason to desire it intrinsically.[6] As

[5] See D'Arms and Jacobson 2000, Rabinowicz and Rønnow-Rasmussen 2004, which introduces the expression 'Wrong Kind of Reasons problem', Jonas Olson 2004, Stratton-Lake 2005, Rabinowicz and Rønnow-Rasmussen 2006, and Danielsson and Olson 2007. Danielsson and Olson (2007) point out that Moore (1903b) may have put forth a version the Wrong Kind of Reasons problem as well.

[6] This distinction is drawn in Parfit 2001: 21–2, though not with these exact labels.

the case illustrates, sometimes reasons to have an attitude derive not from the *object* of the attitude, or from what that object would bring about, but from the *attitude itself*, or from what having it would bring about. In more typical cases, when we have reasons to have some attitude towards some object, the reasons are provided by the nature or effects of the object, as when, for example, we all have reasons to want world poverty to end. These reasons derive not (or not wholly) from any effects of having the attitude but instead from the badness of poverty, or from the non-evaluative features of poverty that make it bad.[7]

1.3. Overview of My Argument

My argument against fitting-attitude analyses of value is indirect. The first part of it says that anyone endorsing a fitting-attitude analysis of intrinsic value ought also, for the sake of theoretical unity, to endorse a fitting-attitude analysis of a closely related but distinct concept: the concept of intrinsic value *for a person*, that is, the concept of *welfare*. It would be surprising to learn that whereas intrinsic value *simpliciter* is analyzable in terms of reasons to have an attitude, intrinsic value for a person is not at all analyzable in this way, and is perhaps instead unanalyzable. This suggests (though admittedly does not entail) that fitting-attitude analyses of intrinsic value *simpliciter* were wrong all along.

The second part of my objection consists in arguing directly against fitting-attitude analyses of welfare. This is what the rest of the paper is about. Even if one rejects the first part of the argument, one can understand this paper to be arguing just against fitting-attitude analyses of welfare. This is not mere shadow boxing, for fitting-attitude analyses of welfare have been defended independently.[8]

[7] Rabinowicz and Rønnow-Rasmussen (2004: 404–8) maintain that restricting a fitting-attitude analysis to object-given reasons still doesn't avoid the Wrong Kind of Reasons problem. They complain that it is difficult to spell out the conditions under which a reason is object- rather than attitude-given (and that certain proposals for spelling it out don't help the fitting-attitude theorist). That may be correct, and I agree it would be best for a fitting-attitude theorist to be able to spell out the conditions. But since we have an antecedent, intuitive grasp of when reasons are object- rather than attitude-given, a grasp that does not depend upon intuitions about value (the analysandum of fitting-attitudes analyses), I think the appeal to the object-given/attitude-given distinction solves the problem well enough. Even if I'm wrong about this, however, granting that the problem is solved only helps the theory this paper is attacking.

[8] e.g. by Sidgwick (1907), Stephen Darwall (2002), and Rønnow-Rasmussen (forthcoming). Their theories are quoted below and briefly discussed.

2. FITTING-ATTITUDE ANALYSES OF WELFARE

2.1. Value *Simpliciter* vs. Welfare Value, or Impersonal vs. Personal Value

On the one hand there is welfare, or well-being, or personal value, or prudential value, or value for a person (or some other subject), or what makes for benefit and harm, or what makes a life in itself worth living. On the other hand, there is value *simpliciter*, or impersonal value, or "value for the world," or what makes the world a better or worse world. Although welfare is a kind of intrinsic value, it is also a *relational* kind of value. We express the idea when we say that something would be *good for* someone. Welfare value is intrinsic because we are saying the thing is good in itself for the person, and not merely good for what it leads to for the person. It is relational because it is a relation between the thing that's good and the person for whom it's good.

Both impersonal and personal value are *non-moral* kinds of value. When we say that something is good in one or the other of these ways, we do not mean the thing is morally good (as only an agent or an action can be). Personal value is rather the kind of value that makes a person's *life go better*. And impersonal value is the kind of value that makes an *outcome* better, or the *world* a better place, or, as we might say, makes *things go better*.

Despite their similarities, these concepts are independent from one another. We can believe that something is good for someone (that its presence makes his life go better) without believing that it is good (that its presence makes the world better, or makes things go better). Consider, for instance, the prospering of the wicked. We can believe Ted Bundy's enjoying something makes his life go better for him without thinking that this enjoyment makes things go better. If we believe injustice to be impersonally intrinsically bad, we might even think it makes things go worse when someone like Ted Bundy is having his life go better.

Likewise, we can think something helps make the world a good world even though it doesn't help make anyone's life good. If we think justice is impersonally intrinsically good, we may think that Bundy's getting the suffering he deserves is intrinsically good, even though this isn't intrinsically good for anyone. We may also think that equality, beauty, virtue, excellence, or noble action is intrinsically good without being intrinsically good for anyone.

The point of all this for our larger purpose here is that even if we have an acceptable analysis of value, we still have to deal with the concept of welfare. We have to be sure that our analysis of value "carries over" to our analysis of welfare, since it seems implausible, and would be theoretically unsatisfying, to be forced to say that these two kinds of value have radically different natures. But, as I will argue momentarily, fitting-attitude analyses do not in fact carry over to welfare.

2.2. Fitting-Attitude Analyses of Welfare

Sidgwick (1907: 112) held that a fitting-attitude analysis applies to welfare value as well as to value *simpliciter*. He interprets " 'ultimate good on the whole for me' to mean what I should practically desire if my desires were in harmony with reason, assuming my own existence alone to be considered" (1907: 112). More recent attempts to carry over fitting-attitude analyses to welfare are made by Stephen Darwall (2002) and Toni Rønnow-Rasmussen (forthcoming).[9] Darwall writes: "what it is for something to be good for someone *just is* for it to be something one should desire for him for his sake, that is, insofar as one cares for him" (2002: 8–9). Rønnow-Rasmussen writes: "An object has personal value for a person *a* if and only if there is reason to favor it for *a*'s sake (where 'favor' is a place-holder for different pro-responses that are called for by the value bearer)."[10] Since one obvious way to understand the expression 'for *a*'s sake' is as meaning the same as 'for *a*'s benefit' or 'for *a*'s welfare', the latter two theories appear circular. But both Darwall (pp. 1–2) and Rønnow-Rasmussen (pp. 14–17) suggest that desiring or favoring something for someone's sake is instead a *way of desiring*, and is therefore (if I understand them correctly) a purely psychological notion, not one whose analysis requires appeal to the notion of welfare.

Before presenting my argument against views such as these, I want to emphasize what seems to me to be an advantage of Sidgwick's version of the theory. It is the fact that, on Sidgwick's view, for something to be good for someone is for it to give *that person*, and not necessarily anyone else, the reasons to have the desire. Cases of the prospering of the wicked reveal the benefit this feature brings. As noted above, it might be a bad thing (or at least fail to be a good thing) for the wicked to prosper, even

[9] Page references for Rønnow-Rasmussen's paper relate to the manuscript available at <http://www.fil.lu.se/publicationfiles/pp88.pdf>.

[10] I should point out that Rønnow-Rasmussen does not mean by 'personal value' what I mean by it. As I use the term, personal value is the same thing as welfare, but as Rønnow-Rasmussen uses it, welfare is just one of two kinds of personal value (Rønnow-Rasmussen, forthcoming: 16). But since welfare is a species of personal value on his view, Rønnow-Rasmussen's theory commits him to the thesis my argument attacks: that if something has welfare-value for *a*, then there is reason to favor it for *a*'s sake.

though their prospering is good *for them*. I think that Ted Bundy's enjoying something is good for him, but it is reasonable to think that this enjoyment fails to be good *simpliciter*.[11] It is plausible that the fact that he would enjoy something doesn't give the rest of us any reason to bring it about—not even an outweighed reason. But Bundy himself still has reason to bring it about—he, after all, is the one who would benefit. So whereas Rønnow-Rasmussen's version of the view implies that if something would be good for Bundy, then *everyone* has reason to favor it, Sidgwick's version more plausibly implies only that Bundy himself has reason to favor it.[12]

Here, then, is the target theory. I include the necessary part analyzing negative welfare value since my argument is more naturally presented as against this part of the theory. I also intend this theory to reflect the solutions discussed earlier to the formal problems for fitting-attitude analyses generally.

FA1 x is intrinsically good for S iff x itself gives S reason to intrinsically desire x for S's sake;

x is intrinsically bad for S iff x itself gives S reason to be intrinsically averse to x for S's sake.

The expression 'x itself gives S reason' is meant to indicate that the relevant reasons are object- rather than attitude-given. Though I formulate the view in terms of reasons (as many fitting-attitude analyzers nowadays do), I will, in what follows, for stylistic variation, use other related notions, like that of an attitude being *fitting*, or *appropriate*, or *warranted*, or *rational*, or one

[11] This point applies even if some objective list theory of welfare is correct (I appeal to enjoyment here and throughout only because it is the least controversial example of a human good). Supposing, for instance, that *experiencing true beauty* is one of the great human intrinsic goods, then, intuitively, it will be good for Bundy to experience true beauty even if it fails to be a good thing that Bundy gets to do this. It is worth pointing out, however, that there is a problem here regarding the idea that there is a close connection between welfare and virtue. If only the virtuous can have good things happen to them, then it may be that anytime someone has something good happen to him, this is also a good thing (since it will automatically be a case of a good person getting something good for him). I thank Mark LeBar for discussion here.

[12] Rønnow-Rasmussen is aware of this objection, and discusses some strategies available to proponents of his view (Rønnow-Rasmussen, forthcoming: 24–5). But I think the Sidgwickian approach presented above is better than the ones Rønnow-Rasmussen considers. Darwall's care condition might seem to enable his theory to avoid this objection. But I think including this condition is in other ways problematic: it threatens to make the analysis circular, as I argue briefly in Heathwood 2003; and, as Rønnow-Rasmussen (forthcoming, 23) points out, it appears to commit the analysis to a subjective, attitude-based theory of practical reason. In any event, the argument I present below still applies to both Rønnow-Rasmussen's and Darwall's original formulations, as I indicate in the next footnote.

that *makes sense*, or one we *ought* to have. I also often leave the 'intrinsically' qualifier for the attitude implicit. I formulate these theses using desire and aversion since fitting-attitude analyzers commonly use these notions, but nothing hangs on this (and, in what follows, I will, for stylistic variation, make free use of other pro- and con-attitudes).

The expression 'for S's sake' is important. Without it, FA1 might imply that anything impersonally good (such as, e.g., Tiger Woods's deservedly enjoying a rousing ovation) is also good *for every subject*, since perhaps we all have reason to want impersonally good things to occur. But since we don't want, and don't have reason to want, these things "for our own sakes," FA1 avoids this implication.

Notice that FA1 involves two "sakes." If something would be good for someone, then she has reason to want it intrinsically for her own sake. But to want something intrinsically is to want it for its own sake. Thus she has reason to want it for its own sake for her own sake. She has reason to want it for its own sake—rather than for the sake it what it would lead to—and also has reason to want it for her own sake—rather than for my sake or for no one's sake at all.

3. WHY FITTING-ATTITUDE ANALYSES OF WELFARE FAIL

My argument against FA1 has to do with time. It is based on the idea that while an event's value for a person is unchanging, the attitudes he has reason to have towards such an event can change over time. We can illustrate this idea using Derek Parfit's ingenious case *My Past and Future Operations* (1984: 165):

> I am in some hospital, to have some kind of surgery. Since this is completely safe, and always successful, I have no fears about the effects. The surgery may be brief, or it may instead take a long time. Because I have to co-operate with the surgeon, I cannot have anaesthetics. I have had this surgery once before, and I can remember how painful it is. Under a new policy, because the operation is so painful, patients are now afterwards made to forget it. Some drug removes their memories of the last few hours.
>
> I have just woken up. I cannot remember going to sleep. I ask my nurse if it has been decided when my operation is to be, and how long it must take. She says that she knows the facts about both me and another patient, but that she cannot remember which facts apply to whom. She can tell me only that the following is true. I may be the patient who had his operation yesterday. In that case, my operation was the longest ever performed, lasting ten hours. I may instead be the patient who is to have a short operation later today. It is either true that I did suffer for ten hours, or true that I shall suffer for one hour.

> I ask the nurse to find out which is true. While she is away, it is clear to me which I prefer to be true. If I learn that the first is true, I shall be greatly relieved.

I make two additional suppositions. First, assume that, in the case, Parfit has his preference—his "bias towards the future"—in virtue of two facts: that he is strongly averse to, or strongly disfavors, his suffering for one hour tomorrow (as the nurse says he might); and that he has no aversion at all to the idea of his having suffered for ten hours yesterday. Thus, I am assuming (but only for now and only for simplicity) that Parfit's bias towards the future is *extreme*: he cares not at all about his own past suffering; he would not "buy" any reduction in past suffering, however large, in exchange for any increase in future suffering, however small. Second, suppose that, as a matter of fact, Parfit is the patient who had his operation yesterday, the operation that lasted ten hours.

Here is the argument against FA1:

(1) Parfit's suffering for ten hours yesterday was intrinsically bad for Parfit.

This is undeniable. We don't need to be hedonists to think suffering is bad for those who suffer—this is rather a datum that any theory of welfare must respect. Note that premise 1 is not saying that Parfit's suffering for ten hours yesterday is *all things considered* or *on balance* bad for Parfit. We can assume that this suffering is all things considered good, due to the good effects of the surgery, which outweigh the badness of the suffering and which would not have occurred had Parfit not suffered.

(2) If Parfit's suffering for ten hours yesterday was intrinsically bad for Parfit, then if FA1 is true, then Parfit's suffering for ten hours yesterday gives him reason to be intrinsically averse to that ordeal for his own sake.

This premise just applies FA1 to the case at hand. If in fact the event of Parfit's suffering for those ten hours really has negative value for him, then, by the second clause of FA1, we can conclude that it gives him reason to be averse to the fact that it happened.[13]

(3) But it is false that Parfit's suffering for ten hours yesterday gives him reason to be intrinsically averse to that ordeal for his own sake.

This is a crucial premise. But it should be intuitively compelling once one appreciates what it is saying. Parfit is being completely reasonable in preferring that his pain be in the past. In fact, even his no longer caring

[13] Rønnow-Rasmussen's analysis likewise implies that Parfit's suffering gives him reason to be averse to it, since, according to this analysis, Parfit's suffering gives everyone reason to be averse to it. Darwall's analysis also implies this, so long as we stipulate that Parfit cares for himself.

at all that it occurred is perfectly fitting—not at all inappropriate. Why should he care about it now? No reason—it's over and done with. When things become past, the reasons they provide can change.

But this isn't true for value; whether an event is intrinsically good or bad for a person doesn't change. It will always remain true that Parfit's ordeal was bad for him (just as it will always remain true that Parfit's ordeal actually occurred). His life is made worse as a result, and there's nothing anyone can do now to change that.

From these claims it follows that

(4) Therefore, FA1 is not true.

Since fitting-attitude analyses of welfare fail, the fitting-attitude approach to value generally looks less attractive. It is more reasonable to suppose that a unified account—one according to which the analysis of impersonal value, if any, carries over to the analysis of welfare—is correct than that personal value and impersonal value have radically different natures.

4. OBJECTIONS AND REPLIES

4.1. The Unfittingness of the Extreme Bias towards the Future

Having no bias at all towards the future does seem crazy, but the extreme bias you stipulate for Parfit is too extreme. Even though it's over and done with, it's still appropriate to be at least a little bit against, for your own sake, the fact that you underwent some terrible ordeal (even when you have no memory of it). After all, it was a terrible ordeal, and you really underwent it.

Suppose Parfit was deciding not between ten hours in the past and one hour in the future of equally intense suffering but between ten years of the most horrific torture in the past and one second of a barely noticeable pain later today. Maybe reason demands that he prefer the latter.

I happen not to be convinced that reason demands that he prefer the latter. So long as we're careful to keep in mind that the ten years of past agony has no bad side-effects—there are no memories of it (which would be bad to have), no post-traumatic stress, no injuries, no concomitant loss of goods[14]—then if Parfit insists, "I don't see why I should care at all about this ordeal that I evidently underwent but that is now over and done with," I don't think we could convince him that he ought to care about it. And I don't think his stubbornness would be unreasonable.

[14] To imagine this properly, we can imagine that, had Parfit not been tortured, he would have been in a coma for those ten years.

But this question is irrelevant anyway. This is because the extreme bias was stipulated only for the sake of simplicity of initial presentation. An analogous argument goes through without assuming that the extreme bias is appropriate.

To see this, first notice that FA1 is an oversimplification. Goodness and badness come in degrees—one good thing can be better than another good thing—but FA1 doesn't reflect this. A complete analysis of personal value would specify how *degree* of goodness or badness depends upon degree of fittingness of attitude, or upon fittingness of degree of attitude. The complete theory would analyze not the notion corresponding to 'x is intrinsically good for S' but the one corresponding to 'x is intrinsically good for S *to degree n*'. One natural view makes use of the notion of strength of reason, as follows:

FA2 x is intrinsically good for S *to degree n* iff x itself gives S reason *of strength n* to intrinsically desire x for S's sake;

x is intrinsically bad for S *to degree n* iff x itself gives S reason *of strength n* to be intrinsically averse to x for S's sake.

Since the better or worse something is, the more reason there is to be for or against it, fitting-attitude analyzers of welfare are committed to something relevantly like FA2.[15] But, since one hour of suffering is intrinsically better (in other words, less intrinsically bad) for the sufferer than ten hours of equally intense suffering, FA2 implies that Parfit's past pain gives him more reason to be averse to it than the future pain would give him to be averse to *it*.

But that's not right. The past pain is not cause for greater alarm than the future pain would be. Parfit's future pain, despite being less bad, provides more reason to be against it.

In response to this, a defender of FA2 might insist

Parfit's suffering for ten hours yesterday does *give him more reason to be averse to that suffering than the future suffering would give him. Just considering the pains themselves, Parfit has more reason to be averse to the past pain (it's a greater pain, after all). But, crucially, this does not commit me to the claim that Parfit should, all things considered, prefer that his operation be tomorrow. For taking into account all reasons, Parfit has more reason to be averse to the future pain, since there are (as this very case illustrates)* time-related reasons,

[15] So, for example, Darwall would be committed, if we were to mirror his formulation, to the view that what it is for something to be good for someone *to some degree* just is for it to be something one *to that degree* should desire for him for his sake, that is, insofar as one cares for him. This theory makes use of the idea of stringency of (prima facie) obligation.

and these time-related reasons tilt the balance so far in the other direction that, all things considered, Parfit has more reason to be averse to the future pain.[16]

I continue to think that it is just false that the past ordeal gives Parfit more reason to be averse to it than the future ordeal does. The past ordeal is over and done with, so it no longer merits much concern (even if, as we are now granting, it does merit a little). And it certainly merits less concern than the future ordeal would merit if it were looming.

But I don't need to rest my response upon this claim, because the appeal to these time-related reasons brings with it new problems for the fitting-attitude theorist. In making this appeal, the defender of FA2 is claiming that although Parfit's past pain *provides* more reason than would the future pain, Parfit nonetheless would *have* more reason to be averse to the future ordeal, due to these time-related reasons. But if Parfit would have this additional reason, it would have to come from somewhere. Something would have to provide this additional reason, something along the lines of the following fact (where t3 is the time of the future operation):

F *that Parfit suffers to degree 10 at t3, and t3 is in the future.*

On this picture, the reasons a pain provides are always proportionate in strength to the amount of pain in the pain, no matter the pain's temporal location. And then the further fact (like fact F) that some pain is in the future provides further reason.

Perhaps this picture is correct, but it cannot be combined with a fitting analysis like FA2. For on any fitting analysis of welfare, not only is it true that whenever there is welfare, there are reasons of a certain sort, it is also true that whenever there are reasons of that sort, there is welfare. If some fact such as F gives someone a reason of some strength to be intrinsically averse to it for his own sake, then, given FA2, it follows that that fact is intrinsically bad for the person to that degree, and makes his life that much worse.

But the idea that F would add disvalue to Parfit's life over and above the disvalue contributed by the suffering is implausible. To see this, consider a bad life that is at its midpoint and whose future is a perfect duplicate its past. This should allow us to say to its subject, "Although your future, like your past, won't be any good, at least it won't be worse than your past was." But, on the present suggestion, this won't be true. We would instead be required to say, "although your future is indiscernible from your past with respect to its non-evaluative features, it will be far worse

[16] I am grateful to Ben Bradley, Jens Johansson, and Doug Portmore for (independently) raising objections along these lines.

than your past was." Defenders of FA2 who adopt the above appeal to time-related reasons will be required to say this because, as we all agree, the subject of this life would have far more reason to be averse to his future than to his past. Given FA2, this will imply that the future is far worse. But clearly it isn't—this person's past is just as bad his future.

4.2. The Reasons Everyone Has to be Averse to Anyone's Suffering

Parfit's past suffering isn't just bad for him, it's a bad thing. Thus we all have reasons to wish it didn't happen. If one were to feel bad about what Parfit went through, that would be fitting (just as it is fitting for anyone now to be disturbed, say, over what one of Ted Bundy's victims went through). But surely Parfit can take up the same "impersonal point of view" towards his past self, and feel bad that there was this person who underwent this terrible ordeal. Such an attitude would be fitting. So premise 3 is false. Parfit has the same reasons we all have to be averse to his suffering yesterday.

I agree that, since Parfit's suffering was impersonally bad, he has the same reasons we all have to be averse to it. But this doesn't contradict premise 3. To contradict premise 3 in the way intended, it needs to be shown that Parfit has reason to be averse to his suffering yesterday *for his own sake*. But, for his own sake, why should he care at all about his past ordeal? It's over and done with.[17]

This reply can be made clearer if we pretend that Parfit is extremely wicked, so that any suffering he undergoes gives no one any reason to be averse to it. Given this supposition, his past suffering is no longer a bad thing (although it is still, of course, bad for him). Since "from the impersonal point of view" no one has reason to feel bad about Parfit's ordeal, he doesn't either, from the impersonal point of view. Thus the only reasons the past suffering could provide anyone to be intrinsically averse to it would be the reasons it provides Parfit himself to be against it "from the first person point of view." But, even if it provides him some such reason to be against it, it doesn't, as FA2 implies, provide more reason to be against it than the future ordeal would provide him.

[17] Here I am again assuming the fittingness of the extreme bias. But this is just for simplicity. Speaking just about what Parfit should want for his own sake, he has stronger reason to desire that he be the patient whose ordeal is over than that he be the patient whose ordeal is looming (even if, since an extreme bias is irrational, he has some reason to be averse for his own sake to both). In what follows, I will continue to assume, for simplicity, the fittingness of the extreme bias.

4.3. The Fittingness of Having No Temporal Bias

True, it is fitting to have a bias towards the future, but it is also fitting to have no temporal bias at all. Each is rationally permissible. Therefore, it would be ok for Parfit to prefer to be the patient whose operation is later today. But the arguments against FA1 and FA2 assume that having no such temporal bias is irrational.

I happen not to be convinced that reason permits Parfit to prefer to be the patient whose operation is later today. Were we in Parfit's shoes, then while the nurse is away, we all would reasonably anticipate, with dread, the possibility of being the patient whose operation looms. If Parfit had no temporal bias, then he would look backward, with a backward-looking analog to dread, to the possibility of being the patient whose operation is over. This doesn't just seem odd, it seems like a mistake. We'd say, "Look—don't you get it? If you're the patient whose operation was yesterday, then your suffering is *over and done with. It's a thing of the past.* Stop getting worked up about it. That doesn't make any sense."

But this question is irrelevant anyway. My point stands even if we grant the permissibility of having no temporal bias. For defenders of a fitting-attitude analysis of welfare like FA2 are committed to the claim that Parfit's past ordeal, since it is worse, gives him more reason to be averse to it than would be given by the lesser, future ordeal. So even if the defender of FA2 somehow nevertheless allows that it is rationally permissible for Parfit to prefer in the temporally biased way he does, she must claim that Parfit fails to prefer in the way that he has *most reason* to prefer. FA2 entails that the balance of reason tilts in favor—very strongly in favor, in fact—of Parfit preferring that his ordeal be in the future (since the possible future ordeal, according to FA2, provides significantly less reason to be averse to it than does the possible past ordeal, and since no other reasons are operating). So if Parfit's attitudes are to be in *full harmony with reason*, he needs to be much more strongly opposed to the thought of being the patient whose operation is over and done with.

But that is not true. And the implausibility of this claim is in no way compensated for by the concession that Parfit's bias towards the future, although way out of whack with what he has most reason to prefer, is nevertheless *permitted* by reason.

It would be nice if we had an argument (in addition to the appeals to intuition) for the claim that having a bias towards the future is fully rational. But I don't think this thesis about rationality can be explained in terms of, or subsumed under, any more general claim about rationality. It seems to me we've reached a brute fact about rationality. One can try to explain why the way Parfit prefers, and the way we all prefer, is perfectly rational by pointing out that Parfit *still has to undergo* the pain if his operation is later today, but the pain *is over and done with* if his operation was yesterday. But this adds nothing. It just repeats in different words what needed explaining.[18] Why is it preferable for a pain to be over and done with? I'm afraid the only answer may be: it just is.

Though the rationality of the bias towards the future may be inexplicable, it is important to note the implausibility of a tempting sort of debunking explanation of our intuitions in favor of its rationality. The debunker might claim that we all intuit that the bias towards the future is rational only because we all have the bias; our having the intuition is thus better explained by its being self-serving than by its being true. This argument is unpersuasive because there are other biases that are ubiquitous but that we nevertheless intuit to be positively *irrational*, such as the so-called bias towards the near—our tendency to care more about our nearer future than about our further future, as when we prefer to delay suffering and hasten enjoyment.

4.4. Time, Tense, and Temporal Indexing Strategies

Something funny is going on in your argument with tense. FA1 and FA2 seem to be stated in the present tense, but the welfare attributions in your argument are stated in the past tense. As a result, it is not clear that

(2) If Parfit's suffering for ten hours yesterday was *intrinsically bad for Parfit, then if FA1 is true, then Parfit's suffering for ten hours yesterday* gives *him reason to be averse to that ordeal for his own sake.*

is true. Since FA1 and FA2 are theories about what it means to say something is *bad for someone, they imply nothing concerning claims to the effect that something* was *bad for someone.*

Fitting-attitude analyses, I was assuming, were meant *tenselessly*. Surely fitting-attitude analyzers mean the theory to be general, so that it applies to all value and welfare judgments, irrespective of their tense.

Still, the theories do seem to be incomplete, since people have reasons to have attitudes at times. Just as FA1 was incomplete in failing to include degree indices, both FA1 and FA2 are incomplete in failing to including temporal indices. Perhaps the completed fitting-attitude analysis of welfare is

FA3 x is intrinsically good for S *at time t* to degree n iff x itself gives S reason *at t* of strength n to intrinsically desire x *at t* for S's sake;

[18] Cf. Parfit 1984:178.

> x is intrinsically bad for S *at time t* to degree n iff x itself gives S reason *at t* of strength n to be intrinsically averse to x *at t* for S's sake.[19]

FA3 is not *ad hoc*. The motivation for including the temporal indices is independent of the debate at hand. Perhaps this natural way to make the theory complete will also help it avoid my argument. This would show that my argument has force only against an oversimplification of a fitting-attitude analysis of welfare.

FA3 does not imply, as FA1 appears to, that Parfit now has reason to be averse to his past ordeal. The time at which Parfit's suffering for those ten hours yesterday has disvalue for him is *during those ten hours*. FA3 therefore implies only that Parfit *had* reason to be averse to his past ordeal, during those ten hours, which surely he did.[20] For similar reasons, FA3 does not imply, as FA2 appears to, that Parfit has more reason to be averse to his past possible long ordeal than his future possible shorter ordeal. This is because FA3 doesn't imply that Parfit has any reason to have any attitude about either possible ordeal.

But FA3 faces new problems. Suppose that Parfit will in fact undergo the future operation. We all agree that this future event gives Parfit reason now to be against it (intrinsically and for his own sake). It is reasonable for Parfit now to be dreading the fact that he will undergo it. But on FA3, whenever there are reasons at a time to have certain desires at that time, there is value at that time. So FA3 implies, absurdly, that Parfit's future suffering is bad for him *now*. But it's not bad for him now—it will be bad for him when it is occurring.

Let me be clear about what FA3 implies here. It implies that Parfit's future suffering is *intrinsically* bad for him now, and that is what's implausible. It is at least conceivable that Parfit's future suffering is *extrinsically* bad for him now, due to the anxiety it might be thought to give rise to now. I suppose this would be true if the following "backtracking" counterfactual were true: if Parfit weren't going to be suffering later today, then he wouldn't be feeling anxious right now. But it is not possible that Parfit's future suffering is *intrinsically* bad for him now. This is because the view that future suffering is intrinsically bad now (or that any future evils are intrinsically bad now) implies that one's present days are *made worse* by the existence of these future evils. It would thus imply, absurdly, that when someone asks, "How

[19] Views like this have been suggested to me by Campbell Brown and by Stephan Torre (personal correspondence).

[20] It may be more accurate to say that *at each individual moment* of Parfit's suffering, the suffering he experienced at that moment was bad for him at that moment; and, likewise, that at each individual moment of the ordeal, Parfit had, at that moment, reason to be averse to the suffering he was experiencing at that moment.

was your day?'', to answer accurately you need to consider not only what happened to you that day, but everything that will ever happen to you in all of your remaining days.

FA3 has another defect. It suggests that for any state of affairs that is good for a person, there is some particular time at which it is good for him. But some philosophers have independently endorsed certain normative theses about welfare that are hard to reconcile with this. For example, David Velleman (1991) claims that the narrative structure of a life can impact how good it is for the person (others before Velleman, including Brentano himself, have made similar claims[21]). So the state of affairs of your life having such-and-such structure could be good for you—it could make your life better than it would have been. But there doesn't seem to be any particular time at which the state of affairs of your life having this structure is good for you.

Another instance is Thomas Nagel, who discusses examples meant to illustrate that ''while [a] subject can be exactly located in a sequence of places and times, the same is not necessarily true of the goods and ills that befall him'' (1970: 77). I don't know whether Velleman's and Nagel's normative views are correct, but it would be better if our metaethical theory didn't rule them out right off the bat, making them conceptually confused, or false by definition.

An advocate of FA3 might reply to the Velleman case that for it to be true that there is no *particular* time at which the state of affairs of your life having its nice structure is good for you, it is enough that it be good for you *at every time*.[22] FA3 would then imply that you have reason, at every time, to be in favor of your life's structure. Though this claim about reasons is plausible, the associated claim that your life's structure is good for you at every time is not. It implies that each moment of your life is made better by the structure had by your whole life. It would thus imply, absurdly, that when someone asks, ''How was your day?'' to answer accurately you need to consider not only what happened to you that day, but also the overall structure of your whole life.

Perhaps, though, now that we have seen the problems with one temporal indexing strategy, we can use what we have learned to rig up an analysis that will spit out the results we want, results that harmonize with the idea that present and future goods and evils give us reasons (or

[21] e.g. Slote (1983) and Chisholm (1986). See also Lemos 1994 and Carson 2000. Carson 2000 contains a useful overview of the views of some of these philosophers on this topic.

[22] Campbell Brown and an anonymous referee have both suggested this reply. Perhaps a similar reply might be made to the Nagel cases as well.

reasons to a degree) that past goods and evils do not. Here is one such proposal:

FA4 x is intrinsically good for S to degree n iff x itself gives S reason of strength n to intrinsically desire x *before or during the time at which x occurs* for S's sake;

x is intrinsically bad for S to degree n iff x itself gives S reason of strength n to be intrinsically averse to x *before or during the time at which x occurs* for S's sake.[23]

According to FA4, if some event that either is occurring or will occur is bad for someone, then this event gives him reason to be intrinsically averse to it for his own sake. But if some valuable event for someone is over and done with, then FA4 does not imply that it now gives him any reasons to have any attitude. FA4 does imply that he *did* have such reasons in the past, before and during the time of the event. But FA4 is compatible with the idea that Parfit now has no reason at all to care about his past suffering.

But FA4 faces new problems. For one thing, it is *ad hoc*. It includes complicated epicycles in the form of disjunctive temporal qualifiers only to get the right result to a specific objection. Since *ad hoc* theories are less likely to be true, we should be dubious of FA4.

But more importantly, FA4 is not even extensionally adequate. To see why, first note an interesting distinction between certain kinds of alleged goods and evils, one that relates importantly to the bias towards the future. When it comes to our own pleasure and pain, the bias towards the future is ubiquitous and sensible. But there are other putative goods and evils about which we are not temporally biased. One example is behind-the-back ridicule. Some philosophers have argued that when you are ridiculed behind your back, this is bad for you, independently of whether you ever find out about it.[24] Though we never know about it, our lives are made

[23] Theories like this have been suggested to me by Michael Huemer and by Elizabeth Harman. It is interesting to compare FA4 to the following theory Sidgwick discusses (1907: 111–12): "a man's future good on the whole is what he would now desire and seek on the whole if all the consequences of all the different lines of conduct open to him were accurately foreseen and adequately realised in imagination at the present point of time." But there is no reason to think Sidgwick states this theory of welfare in terms of "a man's *future* good" in an attempt to avoid our time-related worries. I suspect he just finds it natural to state it this way, since it brings to mind the perspective of the deliberating agent.

[24] Nagel (1970: 76) may be the most prominent advocate of this view. Other examples of non-experienced evils according to Nagel include betrayal, deception, being despised, and having one's will ignored after one's death (1970: 76). See also Kagan (1998: 34–5). It is worth noting that the idea that ridicule, betrayal, hatred, and deception are bad independently of our awareness of them isn't just an intuition. It provides a simple and satisfying explanation for why their discovery is upsetting. Likewise, the idea

worse when such things happen to us. Maybe that's true, but, interestingly, when it comes to such evils, we do not prefer that their instances be in our past. Suppose we learn that it is either true that we did suffer ten ridiculings last week, or true that we shall suffer one ridiculing this week. If later we learn that the first is true, we shall *not* be greatly relieved. We simply (and reasonably) prefer fewer ridiculings, no matter their temporal location. Though this is not so with the "experienced evils" we have been discussing up until now, we have no bias towards the future when it comes to non-experienced evils.[25]

Non-experienced evils make trouble for FA4 in at least two ways. First, consider the possibility *pre-vital harm*. The notion of *posthumous harm* has been widely discussed and defended, but could there ever be pre-vital harm? Could there be an event that occurs before a person begins to exist that is nevertheless intrinsically bad for that person?[26] If there are non-experienced evils such as behind-the-back ridicule, I don't see how one can rule it out. If it is bad to be ridiculed independently of whether one could ever find out about it, then it should still be bad even if it occurs after its victim is dead.[27] And, likewise, it should still be bad even if it occurs before its victim is born.

So, it is a live option in normative ethics that events that occur before a person is born can be intrinsically bad for the person. But FA4 is incompatible with this. Consider some pre-vital harm, x, that occurs at t_0 and that is intrinsically bad for S, who begins existing at t_1. FA4 implies that x gives S reason to be averse to x *before or during* t_0. But this is impossible; nothing can give a reason to someone who doesn't exist. Since it can't be that x gives S a reason at t_0, FA4 is incompatible with x's being bad for S.

Note that this objection applies to FA3, the other temporally-indexed theory, too. The time at which x, the pre-vital harm, is intrinsically bad for S would seem to be t_0, the time at which x occurs (what other time would it be?). But then FA3 will likewise imply the contradictory thought that x gives S a reason to have an attitude at a time at which S doesn't exist.

Even if there is no such thing as pre-vital harm, non-experienced, non-temporally biased evils make trouble for FA4 in a more abstract way as well. The fact that some evils merit a bias towards the future while other evils do not makes FA4 seems like a bizarre view. There are these evils that

that ignoring the wishes of the dead harms the dead provides a simple and satisfying explanation for why we ought to honor the wishes of the dead.

[25] Brueckner and Fischer (1986: 216) note this point, too. Along similar lines, Parfit observes that "The bias towards the future does not apply to many kinds of event, such as those that give us pride or shame" (1984: 172).

[26] It is undeniable that an event that takes place before a person exists could be *instrumentally* bad for that person.

[27] Cf. Parfit 1984: 495.

provide their victims reason to be against them at every time the victim exists, whether this time is before or after the evil event. In "sending out" the reasons these evils send out, these evils don't discriminate between the past and the future—they send out their reasons in both directions, as it were. But FA4 does discriminate between the past and the future. FA4 in fact says that what *makes* these evils evil is merely the fact that the evils send their reasons into the present and the past. The fact that these evils also send reasons into the future is irrelevant to whether these evil events are evil. That is an odd thing for those attracted to a fitting-attitudes theory to say. Surely the spirit of the fitting-attitudes approach demands the idea that part of what makes my being ridiculed last week bad for me is that it now gives me reason to be against it.

4.5. Two Timeless Perspectives

i. Averaging Reasons over Time Normative theories of welfare of the preferentist sort face problems concerning preferences and time as well, for example, the so-called problem of changing desires.[28] Philip Bricker (1980) and Thomas Carson (2000: 86) have introduced ideas that may provide solutions to the problem of changing desires. A strategy shared by each is to construct a sort of "timeless standpoint" (Bricker 1980: 400), built out of all the desires (or rational desires) the subject has at particular times, to arbitrate between his changing desires. Perhaps an idea along these lines could be used by the fitting-attitude theorist. She could propose that what it is for something to be good for someone be explained, not in terms of the subject's reasons at any particular time, but in terms of the average of the strength of the reasons the person has throughout his whole life.[29] This proposal has promise for solving our time-related problems because it offers a fitting-attitude-theoretic way for value to be unchanging. The average of the strength of the reason someone has to want a certain thing over each moment of his life doesn't change.

Call this view 'FA5'. Instead of stating it formally, let's appreciate how it might deliver a more plausible result in a case like Parfit's *Past and Future Operations*. For purposes of illustration we can change the case so that Parfit must undergo *both* the ten-hour and the subsequent one-hour operation. FA5 yields a more acceptable result here than the earlier theories because, if we take as our data facts about what reasons Parfit has at what times, and we "input" these facts into FA5, the "output" FA5 generates is about right.

[28] See Brandt 1982: 179, Bykvist 1998, Carson 2000: 84–87, and Heathwood 2006: 541–2.

[29] I am grateful to Tom Carson for suggesting this application to me.

Even though, at the relatively brief time between the two operations, Parfit has more reason to be against the future suffering than he has to be against the past suffering, this temporary "reversal of reasons" will be swamped by the fact that, at every moment prior to the first operation, Parfit will have a much stronger reason (about ten times as strong) to be averse to the first ordeal, the one that is about ten times worse. So long as there are enough such moments, FA5 will deliver the result that the worse ordeal is about ten times as bad as the lesser ordeal.

But FA5 goes wrong precisely because there might not be enough such moments. Consider what happens if we make the interval between the two episodes of suffering a larger proportion of the whole life in question. If, say, the interval of time between the two operations equals the amount of time Parfit is alive before the first operation, the value of the first, worse episode of suffering will be brought down to only about five times as bad as the lesser episode—and this despite the fact that its intrinsic nature doesn't change between this case and the one above. Worse, if we continue changing the case in this way, making the interval of time between the two episodes almost as long as the whole life in question, the result will be that the second, much briefer ordeal is actually a worse harm than the first, much longer one (the time between the two intervals, during which Parfit has reason to be against the second but not the first, would swamp the time before the first, during which Parfit has reason also to be against the first).

The idea of the average reason someone has throughout his life to want some state of affairs to occur for his own sake is in at least one way more like the idea of the value of that state of affairs for him: both of these quantities are unchanging. But I think it is clear that they are not the same thing.

ii. Counterfactual Analyses The final proposals we'll consider are based around the idea that for something to be good for someone is not for it *actually* to provide her reason to want it but instead for it to be such that it *would* provide her reason to want it, if certain specified conditions held.[30]

Many counterfactual analyses of value specify conditions of *full information*. But this won't help here. Our judgment that Parfit has no reason to be averse to his past pain doesn't depend upon his lacking any information. Were Parfit fully informed, his past pain still wouldn't give him reason to be averse to it.

Another thought is that for something to be good for someone is for it to provide her reason to want to undergo it again, if she could. But this assumes that all goods are things that we undergo, and we have already

[30] Proposals along these lines have been suggested to me, in different ways, by Gunnar Björnsson, Fred Feldman, Pat Greenspan, and Peter Vranas.

discussed many putative examples—non-experienced goods, holistic goods, non-located goods—for which this does not hold.

The problem that time makes for fitting-attitude analyses of welfare stems from the fact that the subjects of welfare, to whose reasons welfare is reduced, are *located in time*. This might suggest that we consider making the relevant counterfactual conditions those in which the subject herself occupies some sort of timeless or atemporal perspective. If there is such a thing as the reasons one would be provided by some event were one to occupy a position outside of time, perhaps these reasons (unlike the actual reasons provided to us as we actually are in time) will be stable enough to provide a plausible grounding for value.

So consider

FA6 x is intrinsically good for S to degree n iff if S were to occupy an atemporal perspective, x itself would give S, while in this atemporal location, reason of strength n to intrinsically desire x for S's sake;

x is intrinsically bad for S to degree n iff if S were to occupy an atemporal perspective, x itself would give S, while in this atemporal location, reason of strength n to be intrinsically averse to x for S's sake.

According to FA6, when some event—an ordinary event that occurs in time, such as someone's being in pain at some time—is bad for some person, its being bad for this person consists in the following fact: if this person, who is actually located in time, were to be located outside of time, the event would, under these circumstances, give the person reason to be against it for his own sake.

FA6 is pretty wild, but that's not my main problem with it (though it may indeed be a problem[31]). It is rather just that, despite its extravagance, FA6 doesn't seem to help. Suppose Parfit were to occupy an atemporal location. From this perspective, he considers some episode of suffering his actual life contains (from this perspective, it is a merely counterfactual episode). Should he be averse to this episode for his own sake?

To try to answer this, I want to ask, Is it true of atemporal Parfit that, despite "currently" being in this atemporal location, he nevertheless *will undergo* the episode of suffering under consideration (as if he is looking down from eternity on the life he is about to begin)? If the answer is Yes, then I think he does have reason to be averse to it (and so FA6's verdict would be correct). But if this is how we understand FA6, then this theory really

[31] Since the concept of an atemporal location at which we could exist might be incoherent. It might be incoherent because it's impossible for time not to exist. Or it might be incoherent because, even it's possible for time not to exist, it would not have been possible for any of us to exist, had time not existed.

just amounts to the tensed theory, FA4, considered above, and inherits its defects. (I am ignoring the problems with the apparently incoherent idea that it can be true of atemporal beings that certain things *will* happen to them.)

So suppose instead that if Parfit were in this atemporal location, then it would neither be true that he will undergo nor true that he did undergo this episode of suffering (the most that is true is that he would have undergone the episode, had he been a temporal being). If this is how it is, then I think it is not at all clear what attitudes Parfit ought to have. We would essentially be asking atemporal Parfit this: suppose you, atemporal Parfit, were a temporal being and were to undergo such-and-such episode of suffering; how do you, as you are, feel about this merely possible episode of suffering, one that you would have underdone, had you been a temporal being?

I'm inclined to think that, just as it is reasonable not to care about one's past suffering, it is reasonable not to care about some episode of suffering that one knows is merely possible for one.[32] Moreover, even if we ought to care at least a little bit about some suffering we would have undergone in some counterfactual situation, surely we don't have as much reason to care about such suffering as we do about our actual future suffering (just as, as already discussed, even if we ought to care at least a little bit about our past suffering, we don't have as much reason to care about such suffering as we do about our actual future suffering).

There is an interesting complication to consider. Should we be inquiring into the reasons atemporal Parfit has to be averse to the suffering in question for the sake of atemporal Parfit, or for the sake of Parfit as he actually is? Either option, it seems, is unsatisfactory. The first option seems unsatisfactory, for why should atemporal Parfit feel bad *for his own sake* about some pain he himself never in fact undergoes? And the second option seems unsatisfactory, too. If actual Parfit still has to undergo the suffering in question, then it seems that, were he atemporal, he would have reason to want that suffering not to occur for his actual self's sake. And if actual Parfit already underwent the suffering in question, then it seems that, were he atemporal, he would not have reason not to want that suffering to occur for his actual self's sake. It's over and done with, after all. We are back again at the original problem.

I think that whatever plausibility FA6 might seem to have is gotten in an illegitimate, question-begging way. We consider some event that is bad for someone. We ask whether the person would have reason to be against it for his own sake if the person were to consider the event from an atemporal

[32] Note that I am not claiming the following: that we don't have reason to care about some future merely possible episode of suffering that nevertheless is, for us, *epistemically possible* (that is, is an episode of suffering that, for all we know, will actually happen).

point of view. I think, at least initially, we don't really know what to say about the reasons we would have in an atemporal location about the actual events of our lives; but we want to be accommodating and answer the question, so we infer that, since the event we're asked to consider is bad, we must have reason to be against it. But this inference makes use of a simplistic fitting-attitude view of welfare, the very theory under dispute. Once we appreciate, from considering the failure of past pain to provide reasons, that this simple inference is fallacious, we should refrain from using it in the atemporal case, too. And once we so refrain, we return to having no clear idea about what reasons we'd have concerning this bad event if we failed to be located in time. And then, finally, when we make efforts to overcome this puzzlement, I think it becomes clear enough that, were we to be atemporal, we'd have, at best, just a little reason to be against the bad event—certainly not as much reason as we have to be against our actual future suffering.

Certain things are good for us and certain things are bad for us. We also often have reason to want certain things to occur, or to have occurred, for our own sakes. Similarly, certain things are good and bad *simpliciter*, and we often have reason to want certain things to occur, or to have occurred. Surely these notions of value and of reasons have something to do with one another. But the connections are untidy, and, in any case, less tidy than any attempt to reduce one to the other requires.

REFERENCES

Anderson, Elizabeth (1993) *Value in Ethics and Economics* (Cambridge, MA: Harvard University Press).

Baldwin, Thomas (2006) 'Keynes and Ethics' in R. E. Backhouse and B. W. Bateman (eds.), *The Cambridge Companion to Keynes* (Cambridge: Cambridge University Press), 237–56.

Brandt, Richard (1946) 'Moral Valuation' *Ethics* 56: 106–21.

—— (1959) *Ethical Theory* (Englewood Cliffs, NJ: Prentice-Hall).

—— (1982) 'Two Concepts of Utility', in H. B. Miller and W. H. Williams (eds.), *The Limits of Utilitarianism* (Minneapolis: University of Minnesota Press), 169–85.

Brentano, Franz (1889) *The Origin of Our Knowledge of Right and Wrong*, tr. Roderick M. Chisholm (London: Routledge & Kegan Paul, 1969).

Bricker, Philip (1980) 'Prudence' *Journal of Philosophy* 77: 381–401.

Broad, C. D. (1930) *Five Types of Ethical Theory* (London: Routledge & Kegan Paul).

Brueckner, Anthony L., and Fischer, John Martin (1986) 'Why Is Death Bad?' *Philosophical Studies* 50: 213–21.

Bykvist, Krister (1998) *Changing Preferences: A Study in Preferentialism*, doctoral dissertation (Uppsala: Acta Universitatis Uppsaliensis).

Carson, Thomas L. (2000) *Value and the Good Life* (Notre Dame, In: University of Notre Dame Press).

Chisholm, Roderick M. (1986) *Brentano and Intrinsic Value* (Cambridge: Cambridge University Press).

Danielsson, Sven, and Olson, Jonas (2007) 'Brentano and the Buck-Passers' *Mind* 116: 511–22.

D'Arms, Justin, and Jacobson, Daniel (2000) 'Sentiment and Value' *Ethics* 110: 722–48.

Darwall, Stephen (2002) *Welfare and Rational Care* (Princeton: Princeton University Press).

Ewing, A. C. (1947) *The Definition of Good* (London: Macmillan).

Gibbard, Allan (1992) 'Reply to Blackburn, Carson, Hill, and Railton' *Philosophy and Phenomenological Research*, 42: 969–80.

Heathwood, Chris (2003) review of Stephen Darwall, *Welfare and Rational Care, Australasian Journal of Philosophy* 81: 615–17.

____ (2006) 'Desire Satisfactionism and Hedonism' *Philosophical Studies* 128: 539–63.

Kagan, Shelly (1998) *Normative Ethics* (Boulder, CO: Westview Press).

Keynes, John Maynard (1905) 'Miscellanea Ethica', unpublished.

Kraut, Richard (1994) 'Desire and the Human Good' *Proceedings and Addresses of the American Philosophical Association* 68: 39–54.

Lemos, Noah M. (1994) *Intrinsic Value: Concept and Warrant* (Cambridge: Cambridge University Press).

McDowell, John (1985) 'Values and Secondary Qualities', in T. Honderich (ed.), *Morality and Objectivity: A Tribute to J. L. Mackie* (London: Routledge & Kegan Paul), 110–29.

Moore, G. E. (1903*a*) *Principia Ethica* (Cambridge: Cambridge University Press).

____ (1903*b*) 'Review of Franz Brentano: *The Origin of Our Knowledge of Right and Wrong*' *International Journal of Ethics* 14: 115–23.

Nagel, Thomas (1970) 'Death' *Noûs* 4: 73–80.

Olson, Jonas (2004) 'Buck-Passing and the Wrong Kind of Reasons', *Philosophical Quarterly* 54: 295–300.

Parfit, Derek (1984) *Reasons and Persons* (Oxford: Oxford University Press).

____ (2001) 'Rationality and Reasons', in D. Egonsson, B. Petersson, J. Josefsson, and T. Rønnow-Rasmussen (eds.), *Exploring Practical Philosophy: From Action to Values* (Aldershot: Ashgate), 17–41.

Rabinowicz, Wlodek, and Rønnow-Rasmussen, Toni (2004) 'The Strike of the Demon: On Fitting Pro-Attitudes and Value' *Ethics* 114: 391–424.

____ ____ (2006) 'Buck-Passing and the Right Kind of Reasons' *Philosophical Quarterly* 56: 114–20.

Rønnow-Rasmussen, Toni (forthcoming) 'Analysing Personal Value' *Journal of Ethics*.

Scanlon, T. M. (1998) *What We Owe to Each Other* (Cambridge, MA: Harvard University Press).

Sidgwick, Henry (1884) *The Methods of Ethics*, 3rd edn (London: Macmillan).

—— (1907) *The Methods of Ethics*, 7th edn. (London: Macmillan).

Slote, Michael (1983) *Goods and Virtues* (Oxford: Oxford University Press).

Stratton-Lake, Philip (2005) 'How to Deal with Evil Demons: Comment on Rabinowicz & Rønnow-Rasmussen' *Ethics* 115: 788–98.

Stratton-Lake, Philip, and Hooker, Brad (2005) 'Scanlon versus Moore on Goodness', in T. Horgan and M. Timmons (eds.), *Metaethics after Moore* (Oxford: Oxford University Press).

Suikkanen, Jussi (2004) 'Reasons and Value: In Defence of the Buck-Passing Account' *Ethical Theory and Moral Practice* 7: 513–35.

Urmson, J. O. (1968) *The Emotive Theory of Ethics* (Oxford: Oxford University Press).

Velleman, David (1991) 'Well-Being and Time' *Pacific Philosophical Quarterly* 72: 48–77.

Wiggins, David (1987) 'A Sensible Subjectivism?' in D. Wiggins, *Needs, Values, Truth: Essays in the Philosophy of Value* (Oxford: Blackwell).

Zimmerman, Michael J. (2001) *The Nature of Intrinsic Value* (Lanham, MD: Rowman & Littlefield).

—— (2004) 'Intrinsic vs. Extrinsic Value', in E.N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy* (Fall 2004 edn), URL = <http://plato.-stanford.edu/archives/fall2004/entries/value-intrinsic-extrinsic/>.

This page intentionally left blank

3

The Argument from the Persistence of Moral Disagreement

Frank Jackson

1. Arthur and Martha disagree about what ought to be done. Arthur favours a course of action we will call Quantity. Martha favours an incompatible course of action we will call Quality. This disagreement occurs against a background of massive agreement on matters that can be expressed in non-moral (and non-normative) terms. They agree, for instance, on who will or would be made happy and to what extent, consequent on deciding between Quantity and Quality, and they agree on everything to do with the attitudes and thoughts of themselves and of others that can be given in non-moral and non-normative terms. There is literally nothing about how things are, specifiable in non-moral terms or *a priori* deducible from same, on which they differ. Indeed, we can go further and suppose that the explanation of their massive agreement is that they *know* all there is to know that can be captured in non-moral terms. How then is it that they differ on what ought to be done?

The answer turns out to be that Arthur gives a higher value to total happiness, by comparison with average happiness, than does Martha. They realize that this is the source of their moral disagreement. Arthur has attempted to convince Martha of the greater normative significance of totals over averages; Martha has attempted to convince Arthur of the reverse. Neither has got anywhere and they have agreed to disagree. Their moral disagreement has persisted despite their massive agreement on, and knowledge about, the non-moral, together with many attempts to reach agreement.

Thanks to the expressivists who forced me to write this paper, to Michael Smith for many discussions, and to comments from two readers.

Expressivists take the persistence of moral disagreement to be a powerful argument for their view.[1] Roughly, their thought runs as follows. The non-moral facts are all the facts there are. And as Arthur and Martha agree and know all about them, their disagreement can only be about something other than the facts. Ergo, moral claims are not claims about the facts—and that's the core claim in expressivism. Their disagreement lies in what they value, not in what they hold to be the case. Suppose, to use a familiar illustration, that I say 'Boo' to something that you say 'Hurray' to. We disagree, but not by virtue of our utterances attributing incompatible properties to whatever it is, or by virtue of the utterances making incompatible claims about how things are. 'Boo' and 'Hooray' aren't in the business of reporting or describing or attributing properties. Instead they express attitudes (or endorsements or 'disendorsements' or . . .). The same is true for ethical and normative sentences, according to expressivists. (I will mainly talk in terms of ethics but the discussion will apply equally to expressivism about normativity in general.)

The appeal of the argument from the persistence of moral disagreement is undeniable. Of course, it might be the case that Arthur and Martha are not disagreeing. It is possible to use the very same word in different senses, and it might be that this is what is happening in our example. Arthur and Martha might be differing in what they mean by 'ought'. In this case, although Arthur says 'Quantity ought to be done' and Martha says 'Quality (and not Quantity) ought to be done', they are not really disagreeing. However, there is a powerful intuition to the effect that this does not have to be the case. The intuition is really one side of the coin which has the open question argument on the other. The intuitive 'openness' of the question concerning what ought to be done after all the non-moral facts are in is essentially the same intuition as the intuition that genuine disagreement about what ought to be done, and about what is good, bad, just and so on, can survive massive agreement about the non-moral.

Some admit the power of the intuition but argue that, all the same, we must repudiate it as resting on one or another mistake or confusion. So one reply to the argument from the persistence of moral disagreement is to deny that disagreement can survive or persist in the way suggested by our example. At the end of the day, Arthur and Martha are talking past each other, for they must be giving different meanings to 'ought'. Another reply denies that the non-moral facts are all the facts there are, despite the supervenience of the moral on the non-moral. There is, that is, something factual left for

[1] See e. g. Gibbard 1990: ch. 1; Blackburn 1984: 168; and the discussion of agreement and disagreement in Stevenson 1944: ch. 1. The extent to which the doctrine in Gibbard 2003 counts as expressivism *in the sense of this paper* is a good question I set aside here.

Arthur and Martha to disagree about which survives their total agreement and knowledge about the non-moral. My aim here, however, is to object to the persistence of moral disagreement argument for expressivism without taking a stand on the viability or otherwise of these two replies. Even if we give expressivists a great deal, their argument still fails, or so I'll argue. I'll argue that it fails for a very simple reason: expressivists have to allow that disagreement can persist, but the only sense of disagreement on which this can be true, by their own lights, is one that cognitivists, in particular certain subjectivists, can equally allow.

I think the reason this objection to the argument has escaped attention is that there's been a tendency by expressivists to take the key notion of disagreement for granted. In fact what happens, as we'll see, is that there are two ways we might naturally cash out the notion of disagreement, and, on one, there is no disagreement in the case of Arthur and Martha by the lights of both expressivists and subjectivists, and, on the other, there is disagreement but it is a sense in which the existence of disagreement does not depend on taking an expressivist view of the situation—for example, a subjectivist could and should agree that there is disagreement in this second sense.

I start with some remarks on the distinction between expressivism and subjectivism.

2. The debate between expressivism and certain versions of subjectivism is not a debate over the existence and nature of pro and con attitudes, acts of valuing or endorsing or disvaluing or … . It is a debate over the relation between pro and con attitudes, valuing and endorsing etc., on the one hand, and something else, on the other. Traditionally, the something else has been ethical sentences and terms. On the traditional picture, everything one party says about the correct way to understand the relevant pro and con attitudes, acts of valuing and disvaluing, their connections to other attitudes and mental states in general, and the degree to which they are subject to rational debate, can be said by the other party. The debate is over whether or not, for example, the sentence 'Quality (Quantity) is the right course of action' reports or expresses the relevant pro attitude to, or endorsement of, Quality (Quantity). Subjectivists say that it reports or describes a certain kind of attitudinal mental state or endorsement. Thus subjectivism is a species of cognitivism in ethics in the sense that, according to it, ethical sentences serve to make claims about how things are, albeit how things are, in part or entirely, with those producing the sentences; whereas expressivists say that the sentence expresses the state or endorsement. Equivalently, subjectivists say that the sentence expresses one's *belief* that one has the relevant attitude to, or endorses, or values whatever it is, whereas expressivists say that it expresses the attitude, or the valuing, or the endorsement itself.

The fact that the difference resides in the relation to sentences and terms is explicit in Ayer's emotivist version of expressivism (1946: 108) but it is equally present in Gibbard's (1990) more developed version. Here's a representative passage from pp. 9–10, with emphases added. Gibbard is discussing expressivist treatments of rationality but the points apply *mutatis mutandis* to the ethical.

> According to any expressivist analysis, to *call* something rational is not, in the strict sense, to attribute a property to it. It is to do something else: to express a state of mind. It is, I am proposing, to express one's acceptance of norms that permit the thing in question. ...
>
> This may seem perverse. Surely a descriptive analysis would be better; that is to say, if a person *calls* something rational, it would be best to hear him as describing it, as ascribing a property to it. ...
>
> My broad response is that any such descriptivist analysis ... misses the chief point of *calling* something rational: the endorsement the *term* connotes.

However, sometimes the difference between expressivism and subjectivism is stated, not in terms of the relation between the pro and con attitudes (from now on I'll mostly include states of accepting and rejecting various norms, endorsing and 'disendorsing', and valuing and disvaluing, under the cover-all terms pro and con attitudes) and ethical or normative sentences and words, but in terms of the relation between the attitudes and certain thoughts or judgements, where we are counting identity as a relation. Here is a representative passage from Blackburn (1998: 50).

> the expressivist thinks that we can say interestingly what is involved for a subject *S* to think that *X* is good. It is for *X* to value it ...
>
> Expressivism denies that when we assert values, we talk about our own states of mind ... [as in subjectivism]. It says that we *voice* our states of mind, but denies that we thereby describe them.

Presumably, voicing is somewhat akin to what happens when you express approval with the word 'Hooray', but what's important for us in the above passage is what appears in the first part: the identification of thinking that something is good with valuing it. It gives us the other way of stating expressivism—as the view that thinking or judging that something is good, bad, right or wrong *is* having a certain attitude to it, where having a certain attitude includes valuing, disvaluing, and the whole range of attitudinal states that present-day expressivists insist marks out their view from the crudities of emotivism.

For some, myself included, this is an unfortunate way of stating expressivism. For us thinking, judging, and believing are all of a piece in the sense of being representational states, and, in consequence, to think that something is good, bad, right or wrong *is* 'to attribute a property to it', and Blackburn's characterization will be in serious tension with Gibbard's. Moreover, it will be obscure how such a view could count as expressivist. To think that *X* is good, if it is indeed a species of *thinking*, will be a thinking that things are thus and so. But what can the thus and so be in this case other than *X*'s being good? But for expressivists, *X*'s being good is not a way things might be; that's the cognitivists', the opposition's, position. Expressivists should view talk of thinking that *X* is good as like thinking that *X* is hooray, where although 'I think that *X* is good' is good English and 'I think that *X* is hooray' is bad English, the grammatical difference is not to the philosophical point. Neither is a genuine case of thinking.

One might hold that this is fussing about a verbal matter. We can, one might suppose, use 'thinks*' for thinking minus the requirement that it is a kind of representational state, and proceed to state expressivism as the doctrine that thinking* that something is good is valuing it; *mutatis mutandis* for thinking* that something is right, rational, irrational, and so on. The trouble is that being a representational state is central to thought; it is far from clear what happens to *thought* when you subtract being representational. It isn't like subtracting one leg from a four-legged stool to get a three-legged one. It accordingly seems best to state the difference between expressivism and subjectivism in terms of the difference in the relationship, according to the two doctrines, between certain attitudes and ethical and normative sentences.

It might be thought that this would leave out of account an important part of expressivism that has nothing to do with sentences and words, a part to do with the psychology of morals. In discussion I have sometimes encountered a position, dubbed a kind of expressivism, that denies that we have cognitive access to the relevant attitudes. The reason, runs this line of thought, that our ethical sentences don't report our attitudes is that we typically don't believe or think that we have the attitudes when we produce the sentences. There is no belief or similar cognitive state available. The key point, according to this approach to characterizing expressivism, is the lack of belief.

This would be a very unappealing version of expressivism. How is it supposed that we English speakers learnt when it is correct to produce a sentence like 'Quality (Quantity) is the right course of action'? In ignorance of our attitudes! And how are we supposed to regulate the production of a sentence like 'Quality (Quantity) is the right course of action'? Do we pay no attention to what we think our attitudes are? Here it is worth recalling that the sentences expressivists often cite when introducing their view, sentences like 'Hooray for happiness', 'Boo to inequitable distributions', and 'Three cheers for compassion', as being like ethical sentences in expressing without reporting attitudes, are ones that we *do* produce when we believe that we have the relevant attitudes. The sentences don't express the belief that we have the attitudes. If they did, they'd report the attitudes, and they don't. They express them. All the same, we do not produce the sentences in a cognitive vacuum. We typically produce them because of how we *take* our attitudes to be. Moreover, expressivists themselves presumably have no trouble in forming beliefs about the attitudes in question. How else could they tell us about the attitudes in their books and papers? And their audiences had better be in the same good cognitive shape. How else could they come to accept expressivism if they don't know about the key attitudes around which the theory revolves? Or think of the interesting accounts of ethical persuasion and debate in Gibbard (1990) and Blackburn (1998). These accounts would hardly make sense if the participants did not know, by and large, what their own attitudes were and what their discussants' attitudes were.[2]

The upshot is that by far the best way of distinguishing expressivism from a certain kind of subjectivism—the subjectivist counterpart to expressivism—has the difference lying in the relation between the attitudes and the sentences and words: according to expressivism, the sentences express the attitudes; according to subjectivism, they report them. The argument to come will be expressed in terms of this way of marking the difference between the two theories. I now turn to giving that argument.

3. If the argument from the persistence of moral disagreement is to succeed as an argument for expressivism over subjectivism, two conditions need to be met. First, disagreement in the sense in question must in fact persist on the assumption that expressivism is true—call this the persistence condition; and, second, this must be true given expressivism and false given subjectivism—call this the discrimination condition. We will see that these two conditions cannot be met together.

If I say that the Democrats will win and you say that the Republicans will win, we disagree. The obvious account of what makes it the case that we disagree is that we both make a claim about how things are, and our claims are inconsistent. Equivalently, we each express a belief about how things are and these beliefs cannot be true together. Call this kind of disagreement factual disagreement. It is what we usually have in mind when we talk about disagreement. It doesn't matter here whether the two claims are inconsistent in and of themselves, or inconsistent given what else is being taken for

[2] One might ask at this point how expressivism avoids making the connection between assertion and belief the right one for reporting rather than expressing. For worries for expressivism on this score, see Jackson and Pettit (1998).

granted (as in our example, where we are taking for granted various facts about elections and party membership).

In the factual sense of disagreement, the persistence condition fails. Far from persisting, there is and never was disagreement in the factual sense if expressivism is true. This is because the ethical sentences in question do not make claims about how things are that might or might not be inconsistent with each other. The expressivists cannot advertise their theory as explaining the persistence of *factual* disagreement. Their theory denies that it ever happens in the ethical case (and the same goes for factual agreement if it comes to that).[3] Of course, subjectivists say the same, except that the reason that there is no disagreement over the facts is that it is the same set of facts—namely, in our example, the different attitudes of Martha and Arthur—that make both sentences true. together.

4. However, as well as factual disagreement, certain differences in cognitive attitudes are naturally classed as disagreements. If I say 'I *believe* that the Democrats will win' and you say 'I *believe* that the Democrats will not win', we count as disagreeing, but of course we may both be right in the claims we are making about how things are. It may well be that I do believe that the Democrats will win, and you do believe that they won't. What about differences in conative attitudes? Differences in preferences *per se* do not count as disagreements. If I like red wine and you like white wine, we differ but do not disagree. But our difference here is not *substantive* in the sense that there is normally no bar to our both getting what we prefer: I spend my money on red wine; you spend your money on white wine. However, if there's only enough money for one bottle of wine at the dinner we are both attending, the situation is different. The situation is more like that in the belief case above. Just as you and I can't both be right in *what* we believe about the election result, you and I can't both get *what* we desire in the modified wine case.

There are nice questions as to when and why a *substantive difference* in cognitive or conative attitudes counts as disagreement. We don't need to enter into them. There is no denying the phenomenon and especially no denying the phenomenon by expressivists. It is a key part of expressivism that the disagreement that persists in our Arthur and Martha example and like cases is not a disagreement over the facts. The difference between Arthur and Martha, according to expressivism, is not that when Martha says, 'Quality is right' and Arthur says. 'Quantity is right', they differ over the

[3] Interestingly, Ayer (1946: 110–11) supposes that the factual sense is *the* sense. In his view, it is a consequence of expressivism, an acceptable one, that ethical disputes are not really *disputes*.

facts, the facts they report with those sentences. The difference lies in their attitudes, in the substantively different attitudes expressed but not reported by the two sentences. If that difference doesn't count as disagreement, expressivists will be denying the very phenomenon they use to argue for the superiority of their theory over any from of cognitivism (including subjectivism). Call this kind of disagreement attitudinal disagreement; it is the kind of disagreement you get when you have a substantive difference in attitudes that are of a kind, a kind we have illustrated without analysing, that amounts to disagreement.[4]

But now the discrimination condition fails. On the attitudinal sense of disagreement, expressivism and subjectivism *alike* save the phenomenon of disagreement in the face of massive agreement over the non-moral facts—that is, all the facts given our presumption that expressivists are right that the non-moral facts exhaust the facts. Although Arthur and Martha agree about all the facts, including the facts about their differences in attitudes (in what they endorse, in which norms they accept—say it how you prefer), if this kind of difference counts as disagreement—as it does, and as expressivists in particular must allow—subjectivists will rightly count them as disagreeing. In sum, the position is as follows: if disagreement means difference over the facts (factual disagreement, as we've called it), expressivists as well as subjectivists must deny the persistence of disagreement; if disagreement means difference in certain kinds of attitudes, both subjectivism and expressivism can explain the persistence of disagreement in face of massive agreement over the facts.

We should not be surprised that expressivism and subjectivism are on a par in respect to how they handle the persistence of disagreement. As we have seen, the difference between expressivism and subjectivism (the subjectivist counterpart of expressivism, that is) is over the relationship between certain attitudinal states of moral agents and ethical and normative sentences. But when people disagree, although the sentences they produce make public their disagreement, the disagreement *per se* is in their minds. If I say 'The Democrats will win' and you say 'The Republicans will win', we make public that we differ about the facts in a way that counts as disagreement, but we don't *create* the disagreement. If we hadn't spoken, our disagreement might have remained a secret, but it would have existed all the same. The phenomenon of disagreement is not essentially tied to words and sentences. This means that two theories differing only in how states of agents are related to words and sentences should not be expected to differ in their implications for the persistence of disagreement. There has always been something curious in the idea that expressivism was better placed to

[4] But see Schroeder (forthcoming) for some analysis.

handle the nature of moral disagreement than subjectivist theories, given that that ethical disagreement resides in that which prompts words (be the prompting a reporting or an expressing) and not in the words *per se*. Of course, sometimes the words we find ourselves producing are our best guide to our attitudes. Expressivists who hold that the relevant state of mind for positive moral evaluations is a kind of endorsing or adopting of a norm or of a plan of action may well say that often the act of endorsing or adoption is revealed to us when we affirm that something is a good thing to do—'I didn't know what I was committed to until I found the words coming out of my mouth'—all the same, the words don't make the attitude or the endorsement or the adoption; they reveal it.

There might, I suppose, be a performative version of expressivism according to which the role of the assertion (in words) that something is right is to *create* the endorsement, in somewhat the way that saying 'I baptize this ship the *Titanic*', in the right circumstances, *is* to baptize the ship the *Titanic*. On this version, the words would reveal the endorsement but also in part create it. But more would need to be said to differentiate expressivism on this construal from subjectivism. The sentence 'I baptize this ship the *Titanic*' reports what I do, as well as being an integral part of what I do. The performative version of expressivism would, that is, need to have *two* clauses: one saying that affirming that something is right makes it the case that one endorses it in the relevant sense; the other saying that affirming that something is right expresses without reporting one's endorsement. The subjectivist counterpart would equally have two clauses: one saying that affirming that something is right makes it the case that one endorses it in the relevant sense; the other saying that affirming that something is right reports one's endorsement (reports, that is, the endorsement that would not exist without the affirmation). Thus, it remains the case that if substantive difference in what's endorsed is sufficient for disagreement, as expressivists of the performative stripe must hold to satisfy the persistence condition, subjectivists have thereby their account of how disagreement persists and the discrimination condition fails. Giving expressivism a performative gloss doesn't save the argument from the persistence of disagreement.

5. I speculated earlier that the problem I've been raising has escaped the notice it deserves due to a failure by expressivists to ask what the disagreement in question comes to. They've taken the notion for granted. Once one asks the question, the destructive dilemma I've been posing for the argument from the persistence of moral disagreement is obvious. This suggests a way expressivists might respond to our objection. They might respond that the notion of disagreement in question should be thought of as

primitive, and it is notable that some expressivists set their discussion in the context of a scepticism about the possibility of reductive analyses.[5] Their scepticism is directed, in the first instance anyway, at attempts to analyse moral and normative concepts like rightness, goodness, and what ought to be done or believed, but a natural extension would be to scepticism about analyses of moral (and normative) disagreement. Be all this as it may, going *sui generis* on the notion of moral disagreement doesn't help the argument from the persistence of moral disagreement. Expressivists who appeal to the argument must maintain that in, for instance, the Arthur and Martha case there is moral disagreement. But their evidence for holding that there is can only be that Arthur and Martha differ in their attitudes, differ in what they endorse, etc. There is no other relevant evidence to hand by their lights. But subjectivists can appeal to exactly the same evidence. It is common ground that Arthur and Martha differ in their attitudes etc. Taking the disagreement that persists to be unanalysable doesn't affect the point that expressivists must maintain that the disagreement persists for reasons that are independent of which of expressivism or subjectivism is correct.

6. There is a final reason why the destructive dilemma I've been posing for the argument from the persistence of moral disagreement may have escaped attention: expressivists may have been committing a fallacy of equivocation. When they say that subjectivists cannot allow for the persistence of moral disagreement, they give 'disagreement' one meaning—that of factual disagreement; and when they say that expressivism can allow for it, they give 'disagreement' the other meaning—that of disagreement in attitude. Of course they won't say it this way. The fallacy of equivocation is a well-known fallacy. They'll say something like the following, or so I presume. Cognitivists in ethics, including subjectivists, have only one thing to mean by ethical disagreement; they can only mean factual disagreement. Expressivists in ethics have only one thing to mean by ethical disagreement; they can only mean disagreement in attitude. It follows that by the lights of the notion of disagreement that is appropriate for their respective theories, expressivists can allow for the persistence of moral disagreement whereas cognitivists, including subjectivists, cannot. This is why, runs the suggestion I am putting into the mouths of expressivists, the persistence of moral disagreement provides a strong argument for expressivism over any cognitivist theory, including the subjectivist counterpart to expressivism.

[5] See e.g. Gibbard (1990: 6n.), and particularly Blackburn (1998: 49), where he says 'Reductionism here, as elsewhere in philosophy, implies seeing one thing as if it were another'.

There are two serious problems for this possible response by expressivists to the charge of committing the fallacy of equivocation. The first is the presumption that when we are dealing with discourse that is to be construed in cognitive fashion, that is, as making a claim about how things are, the only thing to mean by disagreement is factual disagreement. Independently of whether or not expressivists are right that the sentences 'You ought to vote for the Democrats' and 'You ought to vote for the Republicans', in my mouth and your mouth, respectively, express rather than report our attitudes of approval, endorsements, intentions to act, adoption of norms or plans, or whatever, it is undeniable that sentences like 'I give my full support to (thoroughly approve of, endorse, intend to vote for, etc.) the Democrats' and 'I give my full support to (thoroughly approve of, endorse, intend to vote for, etc.) the Republicans' in my mouth and your mouth, respectively, do *report* our attitudes of approval etc. But it would be a mistake to infer that if you and I say, respectively, 'I give my full support to (thoroughly approve of, endorse, intend to vote for, etc.) the Democrats' and 'I give my full support to (thoroughly approve of, endorse, intend to vote for, etc.) the Republicans', we don't count as disagreeing because there is no factual disagreement in the sense that both sentences are, we may suppose, true. The point here is essentially the one we made earlier with the pair: 'I believe that the Democrats will win' and 'I believe that the Republicans will win'. The fact that, consistently with speaking the truth, I can say the first and you can say the second, does not mean that we are not disagreeing. Subjects can count as disagreeing by virtue of the fact-stating sentences they utter in cases where the sentences are true together.

The second serious problem relates to the question of the relevance of intuitions to deciding between metaethical theories. When Ayer (1946) advanced his version of expressivism, he was careful to state that he was describing, not prescribing.

> we are not, of course, denying that it is possible to invent a language in which all ethical symbols are definable in non-ethical terms … ; what we are denying is that the reduction … of ethical sentences to non-ethical statements is consistent with conventions of our actual language. That is, we reject utilitarianism and subjectivism, not as proposals to replace our existing ethical notions by new ones, but as analyses of our existing ethical notions. Our contention is simply that, in our language, sentences which contain normative ethical symbols are not equivalent to sentences which express psychological propositions, or indeed empirical propositions of any kind. (1946: 105)

The role of the intuition that moral disagreement persists has to be that of a pre-analytic datum to guide us in choosing between ethical theories, not something to be tailored to fit one's favoured theory after the event. The nature of the folk intuition of disagreement has to come first, and the question of which account of how we use ethical sentences makes best sense of that folk intuition comes second, telling us something about how we in fact use ethical and normative sentences and words.

It might be objected that this is to presume that there is some reasonably precise folk concept of disagreement to serve as the touchstone, whereas it may be that our folk concept is vague and indeterminate to one degree or another. However, expressivists cannot afford to be too critical of our folk concept. The argument from the persistence of moral disagreement needs it to be the case that there is some reasonably clear folk intuition, for otherwise it is no advertisement for their theory that it explains its persistence. We should seek to explain only that which has some degree of clarity and robustness. However, it does make good sense that our folk concept may be indeterminate between factual disagreement and disagreement in attitude. After all I am using terms of art to mark the distinction between the two concepts. But this degree of indeterminacy is no help to the argument from the persistence of moral disagreement. Run the argument for each precisification of the concept and, as we've seen, it fails on each.

REFERENCES

Ayer, A. J. (1946) *Language, Truth and Logic* (London: Victor Gollancz; 1st edn 1936).

Blackburn, Simon (1984) *Spreading the Word* (Oxford: Oxford University Press).

—— (1998) *Ruling Passions: A Theory of Practical Reason* (Oxford: Clarendon Press).

Gibbard, A (1990) *Wise Choices, Apt Feelings: A Theory of Normative Judgement* (Oxford: Oxford University Press).

—— (2003) *Thinking How to Live* (Cambridge, MA.: Harvard University Press).

Jackson, Frank, and Pettit, Philip (1998) 'A Problem for Expressivism' *Analysis* 58/4: 239–51.

Schroeder, Mark (forthcoming) 'Expression for Expressivists' *Philosophy and Phenomenological Research*.

Stevenson, C. L. (1944) *Ethics and Language* (New Haven: Yale University Press).

4

Moral Disagreement and Moral Expertise

Sarah McGrath

1. INTRODUCTION

The phenomenon of persistent ethical disagreement is often cited in connection with the question of whether there is any "absolute" morality, or whether, instead, morality is in some sense merely "a matter of personal opinion". Citing disagreement, many people who hold strong views about controversial issues such as the permissibility of abortion, eating meat, or the death penalty deny that these views are anything more than "personal beliefs". But while there might be inconsistencies lurking in this position, it is not obviously at fault for according the facts about disagreement some epistemic weight.

This paper addresses the question of whether and to what extent moral disagreement undermines moral knowledge. The most familiar arguments from disagreement in the literature purport to establish conclusions about the metaphysics of morality: that there are no moral facts, or that there are no moral properties, or that the moral facts are relative rather than absolute. Of course, the conclusions of some such metaphysical arguments might be perfectly consistent with the existence of considerable moral knowledge. For example, even if there is some successful argument from disagreement to the conclusion that moral facts are relative rather than absolute, this might very well be consistent with our having just as much moral knowledge as we

Earlier versions of this paper were presented at the 2005 Dubrovnik Metaphysics and Epistemology Conference and at the Second Annual Metaethics Workshop at the University of Wisconsin, Madison. I am grateful to the members of those audiences for their questions and feedback. Thanks also to Frank Arntzenius, Alex Byrne, Elisabeth Camp, John Collins, Adam Elga, Delia Graff Fara, Elizabeth Harman, Tom Kelly, Sarah-Jane Leslie, Ted Sider, and anonymous referees for *Oxford Studies in Metaethics* for helpful comments on earlier drafts.

ordinarily take ourselves to have. (Although of course, such an argument might alter our conception of what it is that we know.) On the other hand, a metaphysical argument from disagreement which successfully showed that there are no moral facts would presumably rule out the possibility of moral knowledge.

By contrast, epistemological arguments from disagreement purport to undermine moral knowledge by showing that, regardless of the metaphysics of the moral facts, we are not in a position to have anything like the amount of moral knowledge that we ordinarily take ourselves to have. For reasons that I explore below, there are various respects in which epistemological arguments from disagreement present a more formidable skeptical challenge than metaphysical ones. My main goal in this paper is to develop an epistemological argument that creates a difficulty for our controversial moral beliefs and to explore the extent to which it succeeds.

2. METAPHYSICAL ARGUMENTS

As a representative metaphysical argument, consider J. L. Mackie's well-known "argument from relativity" (1977: 36–8). According to Mackie, "radical differences between first order moral judgments" provide a compelling reason to doubt "the objectivity of values". While it is not entirely clear what Mackie means when he denies the objectivity of values, he does seem to mean, minimally, that all claims to the effect that something has a certain moral property are false. If Mackie is right about this, then we have very little moral knowledge—far less than we thought we had. Perhaps one could know that nothing is morally wrong, but one could not know of any particular action that it is morally wrong—for all claims to the effect that a particular action is right or wrong are false. (Just as, having learned that there aren't any witches, one could know that Marilyn Manson is not a witch. What one can't know is that anybody *is* a witch.)

Significantly, Mackie does not think that scientific disagreement supports an analogous conclusion about science. He argues that the skeptical inference is compelling in the moral but not in the scientific case because moral and scientific disagreements have different explanations. While scientific disagreement is best explained by the fact that scientists draw different conclusions from inadequate evidence, disagreement about moral codes is better explained by "people's adherence to and participation in different ways of life" (p. 36). In the moral case, "the causal connection seems to be mainly that way round: it is that people approve of monogamy because they participate in a monogamous way of life rather than that they participate in a monogamous way of life because they approve of monogamy".[1] The hypothesis that moral codes are mere reflections of ways of life better explains the pattern of moral variation than does the hypothesis that different people have different "seriously inadequate and badly distorted" perceptions of objective values (p. 37). Thus, there are no objective values.

One immediate concern about Mackie's argument is that it seems to prove too much: it is not true that, in general, where differences in belief covary with differences in ways of life, we ought to draw similar conclusions. For example, within the United States, beliefs about evolutionary theory seem to satisfy the relevant criteria. According to a Harris Poll conducted in the summer of 2005, only one-fifth of Americans believe that human beings evolved from other species; only half think that other plants or animals did; 64 percent believe that "human beings were created directly by God". The poll shows that variation in these beliefs reflects differences in the ways of life of the individuals who hold them, in the sense of reflecting the religious, political, and cultural features of the communities to which they belong: individuals who embrace creationism are more likely to be from the South, to be Republicans, to be religious, and to lack college educations. By contrast, Democrats, those from the Northeast and West, and those with college educations are more likely to believe in evolutionary theory. But while it does seem that people's beliefs about evolutionary theory reflect their ways of life in this sense, this does not support any surprising metaphysical conclusions about the facts at issue. In particular, it does not support the conclusion that there are no truths about the origins of the human species, or that all claims about human origins are false. Perhaps Mackie is correct in holding that the pattern of difference in moral beliefs corresponds to a pattern of difference in the cultural norms prevailing in the communities in which individuals were raised. But even if that is true, it does not show that there are no moral facts.

Of course, more could be said on behalf of Mackie's argument. In particular, one might argue that moral controversy and the controversy about human origins are disanalogous in ways that ultimately prove crucial. I will not explore arguments to that effect here, since I do not claim that the present difficulty is decisive. My purpose in raising this *prima facie*

[1] This quote leaves out some details: on Mackie's view, people's moral beliefs typically reflect "idealizations" of the ways that they actually live rather than simply reflecting those ways of living. So, for example, "the monogamy in which people participate may be less complete, less rigid, than that of which it leads them to approve" (p. 36). And some revisions in moral beliefs are explained not by changing idealizations but by the fact that people aim for consistency: thus, someone might change her belief about whether same-sex marriage is wrong because it conflicts with her other beliefs about what features of a relationship are relevant to whether people ought to marry.

difficulty for Mackie's argument is to highlight a quite general challenge for those who would have us draw conclusions about the metaphysics of morality from the existence of moral disagreement: such arguments naturally invite the charge that they prove too much. In order to successfully respond to this charge, proponents of such arguments must explain why we should not draw the same surprising metaphysical conclusions wherever we find apparently similar phenomena. Why, for example, doesn't widespread religious disagreement show that there is no fact of the matter about whether any gods exist, or that such facts are relative? Notoriously, it is difficult to explain why moral disagreement cries out for the metaethicist's favored metaphysical conclusion while similar disagreement in other domains does not. (Just as we would not want to conclude that all beliefs about human origins are false, so also we would not want to conclude that such beliefs are all relative, or by nature are knowable only by some special faculty of intuition.) Of course, this is not to say that the relevant explanation cannot be provided; only that the task of providing it cannot be avoided, and is far from trivial.

A second potential vulnerability for metaphysical arguments from disagreement is that in general such arguments have the form of inference to the best explanation arguments, according to which the best explanation of the kind of disagreement that we find in the moral domain is the preferred conclusion of the proponent of the argument: that there are no objective values, or, alternatively, that moral facts are relative facts, or that what look like moral claims are really just expressions of emotion, and so on. Because metaphysical arguments are inference to the best explanation arguments, one who offers such an argument must show that her favored conclusion better explains the data than any alternative hypothesis does. One competing hypothesis is the perfectly mundane one that the questions with respect to which we disagree are difficult ones, and at least some of us are getting them wrong; the others include the wide range of surprising candidate metaphysical hypotheses familiar from the metaethics literature. Again, there is no guarantee that such a case cannot be made on behalf of some preferred explanation. The point is just that it is not enough to point to a hypothesis that would adequately explain the relevant features of moral disagreement if it were true: one must show that the hypothesis better explains those features than would any competing hypothesis if *it* were true.

Thus, any metaphysical argument for the skeptical conclusion that we have little or no moral knowledge immediately inherits two potential vulnerabilities. First, to the extent that parallel reasoning applied to other domains would lead to conclusions that we are unwilling to accept, it is potentially vulnerable to the charge that it proves too much or overgeneralizes. Second, inasmuch as such an argument is an inference to the best explanation argument, it is vulnerable to the provision of formidable competing explanations of moral disagreement. In the next section, I consider a line of epistemological argument which possesses neither of these vulnerabilities.

3. AN EPISTEMOLOGICAL ARGUMENT

Consider the following passage from Henry Sidgwick's *The Methods of Ethics*:

> [I]f I find any of my judgments, intuitive or inferential, in direct conflict with a judgment of some other mind, there must be error somewhere: and if I have no more reason to suspect error in the other mind than in my own, reflective comparison between the two judgments necessarily reduces me temporarily to a state of neutrality. (p. 342)

Moreover, according to Sidgwick, "the absence of such disagreement must remain an indispensable negative condition of the certainty of our beliefs" (p. 342).

Let us call a belief ***CONTROVERSIAL*** just in case it satisfies the condition to which Sidgwick draws our attention. Thus your belief that p is CONTROVERSIAL if and only if it is denied by another person of whom it is true that: you have no more reason to think that he or she is in error than you are. Of course, a belief might be controversial without being CONTROVERSIAL. This is the case, for example, when some view that you hold is disputed, but you have reason to think that those who dispute it are more likely to be in error than you are.

As we have noted, Sidgwick holds that no belief that is CONTROVERSIAL can be *certain*. But a parallel claim about *knowledge* also seems attractive. That is, it seems plausible that

> If one's belief that p is CONTROVERSIAL, then one does not know that p.

Suppose that you and your friend Alice intend to take the train together but discover that you have different views about what time it is scheduled to depart: you think that the train departs at a quarter past the hour, while she thinks that it departs at half past. Perhaps you have some good reason to think that Alice is the one who has made a mistake. For example, perhaps you know that she arrived at her view by consulting a train schedule that is out of date, while you arrived at yours by consulting the current schedule. Or perhaps you know that Alice is prone to carelessness with respect to such matters, as she has a past history of having made similar mistakes.

But suppose instead that you have no such reason to think that it is Alice who has made the mistake: as far you know, it is just as likely that you are mistaken as that she is. In that case, it seems that your belief about what time the train leaves does not amount to knowledge.[2]

Of course, it's clear enough that your belief does not amount to knowledge if you are in fact the one in error, i.e., if your belief about what time the train leaves is false. But even if your belief is true, and Alice is the one who has misread the schedule, it seems that your belief does not amount to knowledge provided that you have no good reason to think that she is the one who has made the mistake. Even if your belief would amount to knowledge in the absence of Alice's holding a contrary belief, the fact that she believes as she does can preclude your knowing in the circumstances. For plausibly, this would be a case in which misleading evidence undermines knowledge.[3]

This suggests the following epistemological argument for a certain kind of moral skepticism:

P1 Our controversial moral beliefs are CONTROVERSIAL.

P2 CONTROVERSIAL beliefs do not amount to knowledge.

C Therefore, our controversial moral beliefs do not amount to knowledge.

The first premise and the conclusion of the argument refer to "our controversial moral beliefs". By this, I mean our beliefs about the correct answers to the kinds of questions that tend to be hotly contested in the applied ethics literature as well as in the broader culture: questions about

[2] Cases broadly similar to this one have recently been discussed in the epistemology literature devoted to the question of how we should respond to "peer disagreement". This literature has not directly addressed the question of how disagreement affects knowledge, which is our primary concern here. Significantly, however, a number of contributors to this literature (notably Feldman 2006, Christensen 2007, and Elga 2007) either endorse or express considerable sympathy for the view that peer disagreement should lead the peers to suspend judgment about the disputed question. Presumably, if one ought to suspend judgment as to whether p, then one does not know that P. Kelly (forthcoming) explicitly argues against the view that one is rationally required to suspend judgment in the face of peer disagreement but holds that one should nonetheless become less confident of one's original opinion, and that, all else being equal, as the number of peers on both sides of the issue increases, the push towards agnosticism grows stronger.

[3] Does the fact that your true belief is denied by the relevant kind of person suffice to undermine its status as knowledge, or must one also be *aware* that it is denied by such a person? This is a special case of a substantive and disputed question in epistemology, the question of whether (or in what circumstances) the existence of misleading evidence undermines knowledge when it is not possessed by the would be knower. For discussion, see Harman 1973, Lycan 1977, and Ginet 1980. In what follows, I sidestep this issue by focusing on cases in which one is aware of the disagreement.

the circumstances (if any) in which it is morally permissible to administer the death penalty, or to have an abortion, or to eat meat, or about how much money we are morally obligated to donate to those in dire need, and so on. It is clear that our beliefs about the answers to such questions are controversial ones. It is of course much less clear that they are also CONTROVERSIAL, i.e., that P1 is true. A good part of what follows is devoted to scrutinizing this claim. I begin, however, with a few preliminary remarks about the argument.

First, one who endorses the argument might remain studiously agnostic about the metaphysics of morality, and in particular, about whether there are any moral facts. That is, one who endorses the argument need not take a stand on whether such facts exist, or even on what, if anything, the relevant kind of disagreement suggests about their existence. The contention of one who endorses the argument is rather that the kind of disagreement that we find with respect to controversial moral questions precludes our knowing the correct answers to these questions, *regardless* of whether such questions have correct answers.

Second, the conclusion of the argument is that our beliefs about controversial moral matters do not amount to knowledge. The conclusion is not that it is unreasonable to hold those beliefs in the face of disagreement, or that we are rationally required to suspend judgment with respect to controversial moral matters. However, if the argument is successful, then the skeptic would seem to have made significant headway towards establishing these apparently stronger claims. For it has been argued, with considerable plausibility, that if one is not in a position to know whether p, then the reasonable course is to suspend judgment about whether p until further evidence becomes available; that is, one should not believe when one is in no position to know[4]. Thus, if the above argument is sound, then this would at the very least seem to put considerable pressure on the idea that it is rational for us to maintain our controversial moral views.

Third, in the previous section, we noted that metaphysical arguments from disagreement generally take the form of inference to the best explanation arguments, and that this fact presents a potential line of resistance to such arguments. Notice that the epistemological argument presented here is *not* best reconstructed as an inference to the best explanation argument. The suggestion is not that the best explanation of the disagreement is that no one knows; rather, the suggestion is that the circumstances of the disagreement are inconsistent with one's knowing. Thus,

[4] This conclusion will be especially attractive to those who take knowledge to be the aim of belief; for defense of this claim, see esp. Williamson 2000.

the argument does not share at least one of the two potential vulnerabilities characteristic of metaphysical arguments.

It is less clear, however, that the argument avoids the second potential vulnerability of metaphysical arguments: that of susceptibility to the charge of overgeneralizing, or proving too much. This issue is the focus of the next section.

4. DOES THE ARGUMENT OVERGENERALIZE?

In Section 2, we noted that metaphysical arguments run the risk of overgeneralizing. On the face of it, Mackie's argument that there are no objective values would, if it succeeded in showing claims about value to be false, show the same for claims about human origins. Since we can confidently assume that it would be a mistake to draw that conclusion about human origins, it seems that we can conclude that Mackie's argument falls short of showing that there are no objective values. Does the epistemological argument from disagreement similarly overgeneralize?

One might suspect that the epistemological argument does prove too much. Thus, in responding to a similar line of argument to the one under consideration here, Russ Shafer-Landau poses the following dilemma:

> Either intractable disagreement among consistent intelligent parties forces them to suspend judgment about their contested views, or it doesn't. If it does, then we must suspend judgment about *all* of our philosophical views, as well as our belief that there is an external world, that I am an embodied being, that the earth is older than a second, etc. All of these have been challenged by brilliant, consistent, informed skeptics over the millennia.
>
> Alternatively, if we are warranted in any of our beliefs, despite the presence of such skepticism, then justified belief is possible, even in the face of persistent disagreement. And so we could retain our moral beliefs, especially those we have carefully thought through, despite an inability to convince all of our intelligent opponents. (2004: 108–9)

Here, the suggestion is that our controversial moral beliefs are in the same epistemic boat as our beliefs that *there is an external world* and that *the earth has existed for more than one second*. If this were the case, then we could safely conclude that the argument from disagreement does prove too much, since (I assume) we do know that there is an external world and that the earth has existed for more than one second.

However, the idea that beliefs of this kind and our controversial moral beliefs are equally jeopardized by disagreement seems dubious. After all, my belief that *the earth is older than one second* faces much less opposition than my belief that *the death penalty is morally impermissible*. Even if it is true that brilliant skeptics have disputed the former[5], they are vastly outnumbered by reasonable people who disagree. By contrast, with respect to, say, the moral permissibility or impermissibilty of the death penalty, the division of opinion is not that of lone geniuses vastly outnumbered by the opposition.

As Shafer-Landau interprets the skeptical challenge, it is the *absence of unanimity* among the relevant class of people which suffices to generate the skeptical conclusion. Even the existence of a single formidable dissenter who cannot be won over would suffice to undermine whatever justification one's belief originally enjoyed. This interpretation allows him to plausibly suggest that such a requirement, if consistently applied, would yield a sweeping and global skepticism. However, this is not the most charitable interpretation of the skeptical challenge. On a more charitable construal of that challenge, it is the fact that there is a substantial division of opinion with respect to controversial moral questions that undermines the possibility of knowing the answers to those questions.

In short, the beliefs that *the earth is older than one second* and that *there is an external world* are not CONTROVERSIAL. Even if these beliefs have on occasion been denied by some, including some of formidable intelligence (etc.), it does not follow that one has no more reason to suspect error in such minds than in one's own. Plausibly, one does have such reasons, reasons provided by facts about the distribution of opinion among the relevant class of people. If you and Alice have conflicting beliefs about what time the train is scheduled to depart, then it might be that both of your beliefs are CONTROVERSIAL. However, if you and Alice subsequently discover that ten other people have independently arrived at your belief while none shares hers, your belief is no longer rendered CONTROVERSIAL by the fact that Alice denies it. For now you do have reason to think that she is the one who has made the mistake. On the other hand, her belief—supposing she maintains it—is CONTROVERSIAL: she lacks any parallel reason.

Of course, it is no objection to the skeptical challenge under consideration that it fails to single out our controversial *moral* beliefs. Parallel arguments might be constructed to show that one lacks knowledge with respect to a significant number of topics—for example, philosophy, public policy, and

[5] One might quibble with Shafer-Landau's choice of examples. Skeptics about the external world or the past are best understood, I think, not as disputing first order propositions about the world such as *there is an external world* or *the world is older than one second*, but rather epistemic propositions such as *we know that there is an external world* or *we know that the world is older than one second*. It is clear that some have held views inconsistent with the latter propositions; it is less clear that anyone has held views inconsistent with the former. Still, if it could be shown that our controversial moral beliefs are no worse off than the relevant epistemic beliefs, I would take this to constitute an adequate vindication of the 'companions in guilt' strategy for resisting disagreement-inspired moral skepticism.

religion.[6] But this does not show that the argument *over*-generalizes. For it is far from clear that the answers to much disputed questions in such domains are known; in any case, that some of us have such knowledge is not a *datum* to which one might appeal in attempting to discredit the argument.

5. IN SEARCH OF MORAL EXPERTISE

The previous section defended the argument against the charge of overgeneralization. This section addresses the question of why one should think that one's beliefs about disputed moral questions are CONTROVERSIAL in the first place. We have emphasized that, even if a belief is controversial, it might not be CONTROVERSIAL. That is, even if the truth of a given belief is contested, a person who holds that belief might have good reason to think that anyone who thinks otherwise is more likely to be wrong than she herself is. Indeed, some beliefs might be *extremely* controversial without being CONTROVERSIAL. Consider again our earlier example of evolutionary theory. The proposition that *human beings evolved from other species* is vigorously denied by many, but it would be a mistake to conclude that it is therefore not known by any of those who believe it. Indeed, the fact that it is denied by many does not even preclude its being known by some who are relatively unfamiliar with the scientific evidence in its favor. Crucially, the proposition in question is not controversial among those who are known to possess the relevant expertise. Certain scientific questions might be highly controversial among the population as a whole, but when a consensus or near consensus exists among those with the relevant expertise, one need remain in a state of agnosticism only for as long as it takes to discover the content of that consensus. Thus, despite the large number of people who deny that human beings have evolved from other species, awareness of the expert consensus on the opposite side of the issue provides good reason to think that those who deny it are in error.

It might be thought that there is a parallel defense of one's controversial moral beliefs. That is, it might seem plausible that, although many dispute these beliefs, they are not CONTROVERSIAL, because they are not controversial among "the moral experts". This raises two questions. First, are there genuine moral experts? And second, if there are, how can they be recognized—either by themselves or by others—as such? Let us set aside the first question, and concede for the sake of argument that individuals with genuine moral expertise exist. How might they be identified?

[6] Cf. van Inwagen 1996.

The task of identifying those with genuine expertise will be a much less straightforward matter in some domains than in others. For the most part, the epistemology literature devoted to the topic of disagreement has focused on the idealized case, in which facts about relative expertise and "epistemic peerhood" are treated as given; the question that has dominated that literature concerns how we should respond to disagreement with our epistemic peers or equals.[7] But in actual, real-life cases, others do not typically wear their relative levels of competence on their sleeves. Of course, on occasion they do: most of us have good reason to think that the person whose shirt reads "Expert Plumbers" is someone to whose judgment we should defer with respect to whatever plumbing questions might arise. But in other cases, facts about relative levels of expertise and competence are far from transparent.

In general, identifying those with genuine expertise in some domain will be most straightforward when we have some kind of *independent check*, one not itself subject to significant controversy, by which we can tell who is (and who is not) getting things right. In certain domains, it is relatively easy for us to acquire evidence which bears straightforwardly on questions about relative expertise. Consider, for example, weather forecasting. Two weather forecasters might offer what seem to be equally compelling cases for their conflicting predictions about what tomorrow's weather will be like. But once tomorrow's weather rolls in, we will have an answer to the question of which of today's two conflicting predictions was more accurate. Thus, in the weather forecasting case, inductive track record evidence about who is more reliable is relatively easy to acquire. Moreover, crucially, such evidence can be readily assessed and assimilated by the layperson: one need not be an expert weather forecaster in order to reliably identify those who possess genuine expertise with respect to weather forecasting.

But significantly, we possess no similar independent check for moral expertise. If moral expertise stands to morality as weather forecasting expertise stands to weather, then a moral expert would be someone who consistently arrives at the correct answers to non-trivial moral questions (or at least, someone whose reliability with respect to such questions significantly exceeds that possessed by the average person, when the average person does not form his moral opinions by deferring to a moral expert). Given such a straightforward understanding of moral expertise, there is nothing particularly problematic about the idea that some individuals possess such expertise. The difficulty lies in arriving at compelling grounds for attributing such expertise, either to oneself or to others. A natural suggestion is that the possession of certain academic credentials, or professional concern with

[7] See esp. the work referred to in n. 2 above.

ethics, is good evidence that one possesses reliable moral judgment. I am acquainted with the ethics literature in a way that my plumber is not; moreover, I have taught ethics classes to college students and attended conferences devoted to the subject. He can claim no similar experiences. But in the absence of an independent check on my relative ability to therefore get the answers right, such facts would seem to constitute a relatively meager basis on which to conclude that I am his superior with respect to the reliability of my moral judgment. Again, contrast a case in which we know that one of two weather forecasters is more reliable than the other on the basis of his superior past track record. It would be a mistake, I think, to suppose that in these circumstances I have anything like the kind of evidence for the superiority of my moral judgment that is available in the weather-forecasting case.

Simply put, there is no obvious way to locate oneself in the space of moral expertise relative to others. It is true that professional philosophers who work in applied ethics have thought about the arguments longer than the average person has. Here, as elsewhere in philosophy, this has not resulted in a convergence of opinion. Yet even if these professionals were to converge on the view that, say, killing is no worse than letting die, on the grounds that no adequate metaphysical basis for imputing moral significance to this distinction could be found, it is not clear that ordinary people of the opposite conviction need treat this as conclusive. For it is less clear in the moral case than in various other cases that reliable judgment with respect to the relevant domain is the typical upshot of formal training. Here again the lack of an independent check seems crucial. If a moral expert is someone who tends to get the hard questions right, then good moral training is presumably whatever confers the relevant capacity. That studying structural engineering at MIT is good training for solving the kinds of problem that confront structural engineers can be more or less readily checked by, for example, examining the stability of bridges built by MIT-trained engineers. But in the moral case, since it is unclear how to check who is getting things right, it is unclear how to check whether MIT is a good place for moral training. Thus, while one might think that good moral training would consist in taking a series of ethics courses devoted to the critical examination of arguments on both sides of divisive issues, an equally plausible answer might be that good moral training consists in being raised by virtuous people who devote relatively little time to scrutinizing arguments for and against their views. Similarly, one might think that the best training for appreciating the permissibility or impermissibility of causing animal suffering would involve, among other things, witnessing such suffering. But we could just as easily imagine that the judgment of those best acquainted with the slaughterhouse tends to become artificially deadened to the thought that animals matter.

If the population is substantially divided about, say, the moral permissibility of abortion in certain circumstances, then, assuming that there is some non-relative fact of the matter, a large number of us are wrong. Unfortunately, we possess no analogue to an eye exam, by which we might determine whose moral vision is askew and whose is in good working order. Thus, the truth about where one stands in the space of moral expertise might prove elusive, even for intelligent, thoughtful people.

The upshot of these considerations is that it is quite unclear how one might argue, in a way that is not transparently question-begging or circular, that one's controversial moral beliefs are uncontroversial among the moral experts. But if, for all one knows, there is no consensus among the moral experts in favor of one's controversial moral beliefs, then one cannot appeal to the existence of such a consensus in order to show that those beliefs are not CONTROVERSIAL.

6. MORALITY AND THE CASE OF UNIQUE GREEN

In some ways, moral disagreement seems to parallel the diversity of opinion as to which shade of green is *unique green*. Unique green is that shade of green that is neither bluish nor yellowish. When asked to select the shade which is unique green, different subjects with normal color vision will select different shades.[8] As in the case of our controversial moral views, opinion about which shade is unique green not only fails to be unanimous, but is substantially divided. Perhaps if there were relatively widespread agreement as to which shade is unique green, then the dissenting judgments of a few who possessed otherwise normal color vision could be dismissed. But the fact that the actual division of opinion is substantial suggests that human beings are not reliable detectors of the relevant property. That relevantly similar creatures—creatures with the same type of visual system—arrive at different verdicts when similarly situated seems to show that that kind of creature is simply not well equipped to detect the presence or absence of the property in question. That human beings are not, as a species, reliable detectors of unique green seems to tell against crediting any individual with

[8] See Hardin 1988 for detailed information about the phenomenon. Hardin himself draws a conclusion analogous to Mackie's, viz. that nothing is colored. Cohen 2003 draws a relativist conclusion similar to Harman's: that objects have colors only relative to perceivers and circumstances of viewing. Byrne and Hilbert (2003) conclude that we should suspend judgment about which particular things are unique green.

knowledge that a certain shade is unique green, particularly if the individual knows of this general lack of reliability and has no good reason to think that he is exceptional in this respect.

Note that although questions about which shade of green is unique green are hard questions for human beings, such questions do not present themselves to us as difficult ones. In fact, most subjects are quite confident of their initial judgments; each person's view strikes her as obviously correct. This seems parallel to the moral case: in the moral case too, many find that their own views about controversial moral questions strike them as obviously correct. Cases in which we are quite confident in our original judgments, only to discover there is a substantial division of opinion among those with relevantly similar cognitive capacities, highlight the fact that there is more than one way to discover the relative difficulty of a given intellectual task. While one might learn that a given problem is hard by attempting to think it through and finding oneself struggling or unable to come up with an answer, one might, alternatively, learn that a given problem is hard by discovering that beings with the same cognitive capacities have arrived at wildly different answers. One might learn that a particular philosophical problem—say, about what makes it the case that I am the same person over time—is difficult by attempting to answer it, and finding oneself at a loss. But alternatively, one could learn that it is hard by discovering that there is a great deal of controversy about it. In the case of unique green, learning about the disagreement is the crucial way of finding out that the question is hard: each individual finds a shade that seems straightforwardly neither bluish nor yellowish to her; it is only upon discovering the extensive variation in judgment among those with similar cognitive capacities that the intellectual task is revealed to be difficult. The fact that with respect to various controversial moral questions, many of those on *both* sides of the issue experience their own view as obviously true suggests that here, as in the case of unique green, the better route to appreciating the relative difficulty of the problem is the more indirect of the two.

Judgments to the effect that a given shade is unique green are controversial; are such judgments also CONTROVERSIAL? Suppose that, in fact, some among us *are* reliable detectors of unique green: the initial judgments of members of this subpopulation are quite accurate, and non-accidentally so. Relative to the general population then, the members of this group are "experts" in the straightforward sense employed above. Consider their position, once they come to learn that others—whose color vision otherwise resembles their own—make contrary judgments. Do the experts have more reason to think that it is the others who are in error than that they are? It seems that they do not. For although their judgments are, *ex hypothesi*, more accurate, they have no reason to think so. Thus, even true, reliably formed beliefs about which shade of green is unique green are CONTROVERSIAL. On the assumption that CONTROVERSIAL beliefs are not knowledge, neither are these beliefs.[9]

But while the case against attributing knowledge to even reliable detectors of unique green is quite strong, one might accept this conclusion while denying that the kind of disagreement that surrounds our controversial moral beliefs plays a similarly undermining role. The challenge, then, would be to point to some compelling difference between the moral and color case. Of course, there are some potentially relevant disanalogies. Here I will mention two, and argue that neither is sufficient for a successful defense of moral knowledge.

First, in the case of unique green, the subjects arrive at their judgments completely *independently* of one another. You select a particular shade as unique green, and those who select some other shade as unique green are neither influenced by one another, nor by some common influence. This stands in sharp contrast to the moral case, in which individuals do not arrive at their views about controversial moral issues in isolation. One might claim, then, that disagreement in the moral case creates significantly less skeptical pressure than it does in the unique green case, inasmuch as those on the other side of a particular moral controversy did not independently converge on their view.[10]

Now it is certainly true that, in some cases, learning that those who think otherwise did not arrive at their view independently can substantially reduce the skeptical pressure on one's own view. Consider an extreme case: the population is more or less evenly divided with respect to some question, with large numbers of people on both sides. Suppose you learn that all or almost all of those on the other side of the issue believe as they do because they unquestioningly defer to the judgment of a single charismatic individual whom they regard as a guru. If most of those on your side of the issue arrived at their view independently of one another, then it seems that you might reasonably conclude that your belief is not CONTROVERSIAL despite being controversial. After all, this would be a case in which there is substantial convergence on your view among individuals who made up their minds independently of one another. (Even if the guru himself arrived at the opposite opinion on his own, he is greatly outnumbered by those on the other side.)

[9] Notice that this conclusion can be accepted even by those drawn to reliabilist accounts of knowledge. On sophisticated versions of such views (e.g. Goldman 1986), true, reliably produced belief does not amount to knowledge when the subject is in possession of undermining evidence. (See Goldman's discussion of this point, 1986: 109–13.)

[10] On the importance of independence, see esp. Goldman 2001 and Kelly (forthcoming).

The difficulty with this response is just that actual moral controversies do not seem to exhibit the kind of asymmetry that might make the disagreement less threatening for either of the two sides. Granted, the many people who contest some controversial moral opinion of yours did not converge on their contrary view independently: the correct explanation of why that view is held by a substantial number of people will undoubtedly attribute a great deal to mutual influence, influence of common sources, and the like. But it is not the case that those who share *your* view have independently converged upon it. Thus, although it is true that people arrive at their judgments about unique green in relative isolation as compared to the moral case, it is far from clear that this disanalogy helps to defuse the skeptical challenge facing our moral beliefs.

A second potentially relevant disanalogy between moral controversy and the unique green case is the following. Judgments to the effect that a particular shade is unique green are *non-inferential* judgments. But for many of us, that is not how things are with our controversial moral beliefs. Many of us can provide reasons or arguments in favor of our controversial moral beliefs and against those of our opponents. Suppose, for example, that I disagree with Alice about whether abortion is morally permissible. She says that the fetus has a right to life, and that it follows that abortion is impermissible. But I have read Judith Jarvis Thomson's "In Defense of Abortion", and so I can supply an argument that even if the fetus does have a right to life it does not follow that abortion is impermissible. I take myself to have rebutted her argument and thus to have more reason to think that she is in error than that I am; I thus conclude that the fact that Alice disagrees with me about abortion does not render my belief CONTROVERSIAL.

More generally, one might think that non-inferential beliefs face skeptical pressure from disagreement that is not faced by beliefs that are based on arguments or discursive considerations.[11] Those who find this line of thought plausible will think that the key disanalogy between moral disagreement and the diversity of judgments about unique green lies here: many of us take ourselves to have compelling arguments for our controversial moral convictions while judgments about unique green are non-inferential, brute judgments.

Of course, the mere having of reasons cannot be sufficient to defuse the skeptical challenge. Presumably, if one can only offer bad reasons for one's

[11] For example, Walter Sinnott-Armstrong argues that moral intutionism leads to moral skepticism inasmuch as the moral intuitionist maintains that moral beliefs are non-inferentially justified. According to Sinnott-Armstrong, "disagreement creates a need for inferential justification" (2002: 312).

view, that is not sufficient to break the symmetry between oneself and one's opponent. And neither is the offering of genuine reasons on behalf of one's controversial beliefs, when those genuine reasons can be matched by similar reasons on the other side. After all, in the case of unique green, each of us can at least cite as a reason that this shade appears unique green to *me*. As each person's reason seems equally compelling, the symmetry remains, along with the skeptical pressure.

But suppose that one correctly recognizes that the argument on which the other person bases her belief is fallacious. One is then in a position to conclude that one's own belief is not rendered CONTROVERSIAL by the fact that *this* person holds a contrary view. If one could do this more generally, then one could establish that one's belief was not CONTROVERSIAL. But often those who engage in moral debate are dialectically skilled proponents of the rival views. In such cases, there will be non-fallacious arguments on both sides. The disagreement will then effectively reduce to one about the relative plausibility of the fundamental premises from which the arguments proceed. However, once the disagreement has been reduced to the question of whose premises are more compelling, the gap between the case of moral disagreement and the case of unique green seems to close. Of course, the premises of my argument seem more compelling to me than the premises of Alice's argument; but by the same token, the premises of Alice's argument seem more compelling to her than the premises of my argument.

Alice might show me pictures that motivate a premise of her argument for the conclusion that abortion is impermissible. But the pictures might not move me to agree that that the premise is true. Can I break the symmetry, then, by assuring myself that the reasons that I have are more compelling than hers? This seems no better than simply privileging my judgment about a given shade of green over Alice's contrary judgment.

Once again, it seems that I need some principled line of reasoning by which to privilege my judgment over that of those with whom I disagree. The final section of this paper examines a recent account of such reasoning due to Adam Elga.

7. ELGA'S PROPOSAL

In his recent paper "Reflection and Disagreement", Elga defends a view known in the epistemology literature as the "the equal weight view". According to the equal weight view, one is required to give equal weight to the judgment of an *epistemic peer* as to one's own judgment. You consider someone your epistemic peer with respect to a given question just in case: in advance of either of you reasoning about the issue, you would have predicted that the person in question was just as likely as you to arrive at the correct answer. For example, if I would have predicted that you and I would be equally likely to arrive at the correct solution to some mathematical problem in advance of our actually performing the calculation, then I consider you my epistemic peer with respect to that problem. According to the equal weight view, if you and I arrive at different answers, I am required to suspend judgment.

On the face of it, the equal weight view seems to have far-reaching skeptical consequences, requiring us to suspend judgment with respect to countless controversial questions. Elga, however, argues that this is not the case. His general strategy is to show that one's circle of epistemic peers includes only those with whom one is in substantial agreement on issues closely related to the one under dispute. Thus:

> Consider Ann and Beth, two friends who stand at opposite ends of the political spectrum. Consider the claim that abortion is morally permissible. Does Ann consider Beth a peer with respect to this claim? That is: setting aside her own reasoning about the abortion claim (and Beth's contrary view about it), does Ann think Beth would be just as likely as her to get things right?
>
> The answer is "no". For (let us suppose) Ann and Beth have discussed claims closely linked to the abortion claim. They have discussed, for example, whether human beings have souls, whether it is permissible to withhold treatment from certain terminally ill infants, and whether rights figure prominently in a correct ethical theory. By Ann's lights, Beth has reached wrong conclusions about most of these closely related questions. As a result, even setting aside her own reasoning about the abortion claim, Ann thinks it unlikely that Beth would be right in case the two of them disagree about abortion ... The upshot is that Ann does not consider Beth an epistemic peer with respect to the abortion claim. (pp. 492–3).

It is clear enough how this general line of thought might be adapted so as to apply to the argument with which we are concerned. Again, a belief of yours is CONTROVERSIAL if and only if it is denied by another person of whom it is true that: you have no more reason to think that he or she is in error than that you are. But in what circumstances do you have no more reason to think that the other person is in error than that you are? Perhaps: exactly when, in advance of either of you reasoning about the case at hand, you would have predicted that the other person was just as likely as you to arrive at the correct answer. That is, we might take a CONTROVERSIAL belief to be one that is disputed by someone who is an epistemic peer in Elga's sense. In that case, one might appeal to the line of reasoning Elga provides and hold that one has significantly fewer CONTROVERSIAL beliefs than one might have thought, since those with whom one frequently disagrees over controversial moral questions are outside one's circle of peers. Following Elga's lead, one might say: even though Ann knows that Beth disagrees with her about abortion, this has no tendency to make Ann's view about abortion CONTROVERSIAL.

Elga anticipates a natural objection that runs as follows: Ann cannot legitimately take her own views on the surrounding issues for granted and use them as a basis for concluding that Beth is more likely to get things wrong with respect to abortion. Rather, Ann should think of the entire cluster of related issues as a single compound issue, and take into account Beth's disagreement about this single compound issue. Once she does this, Ann will no longer be in a position to penalize Beth for having, by Ann's lights, false views about the surrounding issues. Hence, the case for skepticism is restored.

In response, Elga offers the following:

> Consider the cluster of issues linked to abortion. Contrary to what the objection supposes, Ann does *not* consider Beth a peer about that cluster … That is because there is no fact of the matter about Ann's opinion of Beth, once so many of Ann's considerations have been set aside … To set aside Ann's reasoning about all of these issues is to set aside a large and central chunk of her ethical and political outlook. Once so much has been set aside, there is no determinate fact about what opinion of Beth remains. (pp. 495–6).

He motivates this claim with an example: plausibly, there is no determinate answer to the question of what your opinion of Jennifer Lopez is, setting aside your views that humans have bodies and that the Earth exists (p. 25).

However, while it seems right to say that there is no fact of the matter about your opinion of Lopez setting aside your beliefs about human embodiment and the existence of the Earth, the same maneuver seems less plausible when applied to the case of Ann and Beth. Recall that Elga characterizes Ann and Beth as "two friends at opposite ends of the political spectrum". We might then think of Ann as a conservative Republican who takes abortion to be morally abhorrent in most circumstances, and Beth as a liberal Democrat who thinks that it is morally permissible in most circumstances. No doubt, we would expect Ann and Beth to disagree about a wide range of moral issues. Significantly, however, this is perfectly consistent with a very substantial amount of moral agreement between the two. Indeed, we would expect Ann and Beth to agree about the answers to any number of moral questions. We would expect them to agree, for example, that slavery is morally abhorrent, that it is wrong to cause others pain for the sake of one's own amusement, that lying is *prima facie* wrong, and about countless other issues. Moreover, notice that many of the issues on which they are likely to agree are highly non-trivial, at least when judged by world-historical standards. (Consider, for example, their shared belief that slavery is morally abhorrent.) With respect to moral sensibility, we would expect Ann and Beth to resemble one another far more than either resembles, for example, a committed Nazi, or an ancient Hittite lord for whom the thought that slavery is abhorrent was simply not on the moral map. In short, Ann and Beth's disagreements about abortion and related matters, although substantial, almost surely take place against a relatively wide background of shared moral beliefs. It seems wrong, then, to say that there is no fact of the matter about Ann's opinion of Beth's moral judgment, setting aside abortion and the many related issues about which they disagree. Indeed, once these disagreements are bracketed, the relatively wide background of agreement seems to tell in favor of Ann's taking it that Beth is more or less equally likely to get the hard questions right.

Once again, a comparison with the case of unique green seems apt. Subjects with normal color vision make contrary judgments as to which shade is unique green. Thus, those contrary judgments take place against a relatively substantial background of shared color judgments. Plausibly, this fact tends to increase the skeptical pressure: a shared background of agreement strengthens the case for counting the conflicting judgments as CONTROVERSIAL. The same would seem to be true in the moral case. Elga's proposal suggests that I can simply rank myself above those who disagree with me about controversial moral issues on the grounds that our disagreement is substantial. But this seems like a dubious procedure for locating myself relative to others in the space of moral expertise—all the more so when those who disagree do so against a wide background of *agreement*. After all, according to me, they usually get it right.

REFERENCES

Byrne, A., and Hilbert, D. (2003) "Color Realism and Color Science" *Behavioral and Brain Sciences* 26: 3–21.

Christensen, D. (2007) "Epistemology of Disagreement: the Good News" *Philosophical Review* 116: 187–217.

Cohen, J. (2003). "Perceptual Variation, Realism, and Relativization, or: How I Learned To Stop Worrying And Love Variations In Color Vision (Open peer commentary on Byrne and Hilbert, *Color Realism and Color Science*)" *Behavioral and Brain Sciences* 26: 25–6.

Elga, A. (2007) "Reflection and Disagreement" *Noûs*, LXI(3): 478–502.

Feldman, R. (2003) *Epistemology* (Upper Saddle River, NJ: Prentice-Hall).

—— (2006) "Epistemological Puzzles about Disagreement" in *Epistemology Futures*, ed. S. Hetherington (Oxford: Oxford University Press).

Ginet, C. (1980) "Knowing Less by Knowing More" *Midwest Studies in Philosophy.* 5: 151–62.

Goldman, A. (1986). *Epistemology and Cognition*. (Cambridge, MA: Harvard University Press).

——— (2001) "Experts: Which Ones Should You Trust?" *Philosophy and Phenomenological Research* 63.

Hardin, C. L. (1988) *Color for Philosophers* (Indianapolis: Hackett).

Harman, G. (1973) *Thought* (Princeton: Princeton University Press).

——— and Thomson, J. J. (1996) *Moral Relativism and Moral Objectivity* (Cambridge, MA: Blackwell).

Harris Poll (2005) "Nearly Two-thirds of U.S. Adults Believe Human Beings Were Created by God." Harris Interactive, Inc. http://www.harrisinteractive.com/harris_poll/index.asp?PID=581

Kelly, T. (2005). "The Epistemic Significance of Disagreement" in *Oxford Studies in Epistemology*, ed. J. Hawthorne and T. Gendler Szabo (Oxford: Oxford University Press), 167–96

——— (forthcoming) "Peer Disagreement and Higher Order Evidence" in *Disagreement*, ed. R. Feldman and T. Warfield (Oxford: Oxford University Press).

Lycan, W. (1977) "Evidence one does not possess" *Australasian Journal of Philosophy* 55: 114–26.

Mackie, J. L. (1977) *Ethics: Inventing Right and Wrong* (New York: Penguin).

Rosen, G. (2001) "Nominalism, Naturalism, Epistemic Relativism" *Philosophical Perspectives* 15: 69–91.

Shafer-Landau, R. (2004) *Whatever Happened to Good and Evil?* (Oxford: Oxford University Press).

Sher, George. (2001) "But I Could Be Wrong" *Social Philosophy and Policy* 18: 64–78.

Sidgwick, H. (1907/1981) *The Methods of Ethics* (Indianapolis: Hackett).

Sinnott-Armstrong, W. (2002) "Moral Relativity and Intuitionism" *Philosophical Issues* 12: 305–28.

van Inwagen, P. (1996). "It is Wrong, Always, Everywhere, and for Anyone, to Believe Anything, Upon Insufficient Evidence" in *Faith, Freedom and Rationality*, es. J. Jordan and D. Howard-Snyder (Lanham, MD: Rowman & Littlefield).

Williamson, T. (2000). *Knowledge and its Limits* (Oxford: Oxford University Press).

This page intentionally left blank

5

Moral Dependence

Nick Zangwill

1. BECAUSE AND NECESSITY

What is the relation between moral and natural properties? And how do we conceive of this relation? By '*moral*' properties I will mean properties such as being evil, just, or virtuous or having duties or rights; and by '*natural*' properties I will mean properties such as psychological, sociological, and physical properties.[1]

Suppose we judge that Queen Isabella of Spain was evil in 1492, or at least that many of her actions in 1492 were evil. Then we do not think that she had various natural properties in 1492—such as being a torturer, a bigot, and desiring other's pain—and *by an astounding coincidence* she or her actions also had the moral property of evil. Rather, we think that she or her actions were evil *in virtue of* those natural properties; we think that her moral properties *depend* on her natural properties; we think that she had her moral properties *because* of her natural properties. In general, when we make a moral judgment we judge not just that something has a moral property, but that it has a moral property *because* it has some natural property. This is a fundamental principle of our moral thought.

I am very grateful for many interesting questions raised by audiences at the University of Melbourne, the Centre for Applied Philosophy and Public Ethics in Canberra, and at the Third Annual Metaethics Workshop at the University of Wisconsin, where there were many penetrating questions. Two referees for the Press also made helpful comments.

[1] This stipulation has the consequence that God's psychological properties are natural properties, but I don't think this matters much. Adding a spatial requirement for natural properties would allow that God's mental states are not natural. (Moore distinguished, 'naturalist', 'non-naturalist', and 'metaphysical' ethics in Moore 1903.) I do not focus on so-called 'thick' concepts for many reasons, which I shall not detail here.

We may think that a natural property is that in virtue of which an act has a moral property whether or not we think that there is a *further* explanation of why it is so. For example, we may think that an act may be wrong in virtue of being a lie whether or not we think that there is a further explanation of why lying is wrong. That is, we may or may not think that some other more basic moral–natural dependence underpins the particular one. But we cannot judge that an act is barely wrong; we must judge that it is wrong because of some natural property of the act, such as being a lie.[2]

But, as John Mackie famously asked, "What in the world is signified by this 'because'?" (Mackie 1977: 41). What exactly is this relation of moral–natural dependence?

A plausible thought is that we need a *necessity* to make sense of the "because": moral and natural properties are necessarily connected. What is called moral–natural "supervenience" is the usual way of articulating precisely what the necessary connection is.

In this paper I shall argue, to start with, that this is incorrect or at least oversimplistic, even though necessary ties are implied by moral–natural dependence. I then consider the epistemic status of moral–natural dependence and moral–natural supervenience, and the relation between them. I end by drawing conclusions about the kind of metaethical theory we should seek.

For the most part, I shall restrict my attention to moral normative properties, but I am inclined to believe that the same applies to other normative properties such as epistemic properties.

2. MORAL CONDITIONALS, RESPONSIBILITY, AND RELEVANCE

Shadowing the dependence relations signified by this "because", there are conditionals with natural antecedents and moral consequents. For example, a conditional judgment might be that that *if* someone likes torturing *then* he is evil, or if someone lies then he does wrong. I call these 'moral dependence conditionals'. Should we understand moral dependence conditionals as holding with necessity?

One thing that might tempt us to think that these moral conditionals hold with necessity is that statements of strong moral supervenience

[2] Elsewhere I probe this requirement, which I call the 'Because Constraint' (Zangwill 2005). There are some special cases, such as when we judge on authority, where although we must think that there is some responsible natural property, we need not know exactly which it is. But such cases are secondary.

involve such necessary conditionals in their consequents (Zangwill 1995). Strong moral supervenience is a general framework principle of our moral thought that says that if something has a moral property M, then it has a natural property N, such that if anything at any time is N, then it must also be M. (Or, in only six words: moral instantiations have natural sufficient conditions.) A quite plausible thought is that in moral judgments we assert various specific conditional consequents of the overall strong supervenience conditional. The overall claim is grasped by every moral thinker. But the overall claim embeds an existential quantification over natural properties, and thus in effect it existentially quantifies over necessary natural-to-moral conditionals. *Which* necessary natural-to-moral conditionals obtain, though, is a substantive question and not something every moral thinker knows. We are thus led to the view that although the antecedents of moral dependence conditionals obtain contingently, the moral dependence conditionals themselves hold necessarily. It is contingent that Isabella had certain mental states, performed certain actions, and whatever else, in virtue of which she was evil in 1492; but, necessarily, given that she did, she was evil.[3] On this way of thinking, moral cognition is primarily of necessary relations, and only given additional information about which natural properties are actually instantiated can we conclude anything about the morality of actuality. We may have an 'egocentric' (!) moral interest in actuality; but our primary moral understanding is of necessary links between moral and natural properties.

This line of thought is attractive, but it must be resisted or at least considerably modified.

Let us call the natural properties of a thing that *make* it good or bad, the "responsible" natural properties. By contrast, let us call "relevant" those natural properties that generate moral–natural counterfactual conditionals. We can distinguish natural properties that are *responsible* for a moral property from those that are merely *relevant* to it. For example, some act might be bad that *is* an act of intentionally causing pain to a child but is *not* an act of intentionally causing him pain in order to medically benefit him. Intuitively, intentionally causing pain *makes* it bad in a way that not being an act of intentionally causing pain in order to medically benefit him does not, even though it is true *both* that if it had not been an act of intentionally causing pain then (other things being equal) it would not have been bad, *and also* that if it had *not not* intended to be medically beneficial then (other things being equal) it would not have been bad.

[3] Similarly, a sample of a substance might not have existed; but given that it does, whatever composes it does so necessarily (Kripke 1980; Putnam 1975).

Making a moral judgment about a thing does not mean a commitment to having located all of a thing's natural properties that are counterfactually relevant to its instantiating the moral property—that is, which allow its instantiation. (For one thing, there are an infinite number of negative properties that are counterfactually relevant.) That would be asking too much. What we are committed to is the existence of a subset of a thing's natural properties that *make* it good (or bad or whatever). We may be ignorant of *many* natural features of what we are evaluating. But in so far as we make a moral judgment about it, we assume ourselves to have located natural properties that are *responsible* for its possessing the moral property. To judge that x has a moral property is to judge that x has some natural property that is responsible for x having the moral property. This notion of responsibility is hard to analyze. But it is fundamental. It is a notion we need all over philosophy. And it seems to differ from necessity.

I found two authors who have had similar thoughts. In a very interesting paper published in 1970, Michael Stocker wrote:

> I shall say that a *precondition* of a moral duty is a condition such that if it does not obtain one would not have that moral duty. ... There are some moral duties that we have (or might have) but would not have were that condition not met, yet which we do not have even in part *because* that condition is met. (Stocker 1970: 610, his emphasis; see also pp. 606 and 607)

Stocker's thought lay neglected for many decades. However, Jonathan Dancy has recently followed Stocker's lead (Dancy 2004, ch. 3).[4]

3. NEGATIVE RESPONSIBILITY?

We can often illustrate the responsibility/relevance distinction by picking positive and negative natural properties of a thing such that its negative properties are morally relevant while its positive properties are morally responsible. In my example, the positive natural property of intentionally causing pain is responsible for the badness, while the negative natural

[4] Dancy thinks that there are certain features of things that he calls 'enablers' that do not themselves make something good but which are necessary if something else is to make it good (Dancy 2004: ch. 3; see also Stocker's use of "enable" at Stocker 1970: 607). Dancy makes some interesting points about non-responsible natural properties that nevertheless play a role. He casts the issue in terms of 'reasons', which I think is unfortunate, but his points survive translation into property terms. Frank Sibley uses the words "result" and "responsible" for the relation between aesthetic and non-aesthetic properties (Sibley 1965: 138–40). In explaining aesthetic/nonaesthetic dependence he steers completely clear of both modal and conditional formulations. Sibley's papers on aesthetics might be read with profit by moral philosophers.

property of not causing pain for medical benefit is merely counterfactually relevant to it. There is a question about whether we should generalize and say that negative natural properties are never responsible for positive moral properties. I find this quite an attractive thesis.

My view is that, quite generally, negative properties do not determine positive properties in the way that positive properties do, even though both generate counterfactuals. Not being plutonium does not *make* something water in the way that being H_2O does, even though both generate counterfactuals. (This is a controversial general metaphysical thesis; see Zangwill 2003.) In morality, there is no denying that negative natural properties are *counterfactually relevant* to moral properties. Nevertheless, I am *inclined* to think that these negative natural properties are not *responsible* for the instantiation of (positive) moral properties in the way that some of its positive natural properties are.

Not every positive natural property of the thing that has the moral property is responsible for that moral property. Many subatomic properties, for instance, are positive natural properties that are not responsible for moral properties. They are not relevant either. Many positive natural properties of a thing are neither responsible nor relevant for its moral properties. Clearly, many negative natural properties of a thing (such as those in my example of causing pain to a child) are counterfactually relevant to its moral properties but not responsible for them. And others are neither responsible nor relevant. The question is whether *any* negative natural properties are responsible for moral properties.

In discussion, Lloyd Humberstone wondered whether my general claim that all moral properties have natural makers could cope with *permissions*. Surely the instantiation of the moral property of being permissible has no natural maker. One way to respond to this worry would be to take permissibility to be a negative moral property and to restrict the doctrine of moral–natural dependence to positive moral properties, like being required or being wrong. But that seems rather arbitrary. Instead, consider the judgment that stamp-collecting is permissible because it does not do any harm. Here it seems that the negative moral property has a *negative* natural maker. So the suggestion might be that positive moral property instantiations require positive natural makers and negative moral property instantiations require negative natural makers. Thus negative natural properties *are* responsible, but only for negative moral properties.[5]

[5] It might be argued that if the deontic moral operators are interdefinable, then calling some positive and others negative lacks meaning, and the whole issue lapses. One response would be that, although there are necessary biconditional relations between positive and negative deontic properties, still, the positive deontic properties are metaphysically prior.

But what if this were questioned? Perhaps some positive moral properties have negative natural makers and some positive natural properties have negative moral makers. One example is this. Some people might do wrong because they do *not* drive with due care and attention. In this case, the positive moral property seems to have a negative natural maker. Similarly, some people might be permitted to drive on the 'wrong' side of the road because they are policemen. This is a negative moral property with a positive natural maker.[6] One possibility is that we can deal with such cases by saying that the maker is a hybrid conjunction of a positive property and a negative property. Perhaps the wrong-maker is driving *and* failing to take care.

I shall leave this topic here (apart from making a suggestion in a footnote in the next section). The important point for this paper is the distinction between making and allowing—although the issue of positive and negative moral determination, it seems to me, warrants further exploration.

4. THE CONDITIONAL CRITERION, PRAGMATISM, AND CONTEXTUALISM

But how are we to distinguish responsible from relevant natural properties? How can we know which are right-makers and which are right-allowers?

Thus, although it is true that necessarily A ought to do X if and only if A is not permitted not to do X, nevertheless, A is not permitted not to do X *because* A is obligated to do X, and *not* vice versa. The obligation is metaphysical prior. (See further, section 5 on divine commandment theory.) Similarly, although it is true that necessarily A is permitted to do X if and only if A is not obligated not to do X, nevertheless A is not obligated not to do X *because* A is permitted to do X, and *not* vice versa. Here *permission* is metaphysically prior. (On this view, obligations and permissions are both positive moral properties.) A different response would be to deny that there are necessary biconditionals. The number 7 has the misfortune to lack rational agency and moral responsibility. Therefore it lacks obligations and permissions. It is neither obligated to travel on the bus nor obligated not to travel on the bus. But that does not mean that it is *permitted* to travel on the bus. The number 7 is not the bearer of moral properties of *any* sort—obligations or permissions. It is not true that if something is not obligated not to do something then that thing is permitted to do it, and it is not true that if something is not permitted not to do something then that thing is obligated to do it. (The dialectic is like that over the so-called disquotation theory of truth, according to which all there is to say is necessarily ['p' is true if and only if p]. One response is to assert the necessary equivalence but nevertheless give the right-hand side metaphysical priority. Another response is to deny that the right-hand side entails the left-hand side, so the necessary biconditional is not true at all.)

[6] Dancy thinks that some of what he calls 'enablers' are positive. In principle I could accept this. His example is of something we should do because we promised to do it. It might be a positive property of the promise that it was done freely. But being done freely doesn't make the act right. It is merely an enabler or a precondition. The contrary view would be that it is only the conjunction of being a promise and being freely done that is the right-maker.

It is one thing to make a principled distinction and another to have a means of telling when it applies. Furthermore, if we make some abstract philosophical distinction but could not apply it in practice, that would cast doubt on the abstract distinction.

I propose that one important difference between relevance and responsibility is that when the instantiation of a natural property N is responsible for the instantiation of a moral property M, that (typically) generates the *factual* conditional if Nx then Mx, whereas merely relevant natural properties do not. By contrast, a *counterfactual* of the form if not–Nx then not–Mx, is generated *both* by relevant and responsible natural properties. (A 'factual conditional' has a true antecedent whereas a 'counterfactual conditional' has a false antecedent.) Counterfactual conditionals do not allow us to discriminate relevance from responsibility, but factual conditionals do. Moral dependence or responsibility is a stronger notion than that of the moral factual conditional. But dependence or responsibility is a relation that explains the holding of factual conditionals. And every moral judgment involves a commitment to such a factual conditional.[7]

Appealing to factual conditionals enables us to reply to an important objection, which is that the distinction between natural properties that are responsible (makers) and those which are merely relevant (allowers) is arbitrary and depends on our interests. Jamie Dreier put the following case to me. Suppose a doctor knowingly fails to do something that is medically beneficial, out of laziness or callousness. In this case that natural property—failing to medically benefit—*is* a wrong-maker, unlike the case I had in mind earlier one, where the wrong-maker is being an intentional pain-causing, and failing to medically benefit was merely counterfactually relevant. So, the argument is that the making/allowing distinction is merely a pragmatic one, as it is when we pick out something as a cause as opposed to a 'mere background condition'. Depending on our interests, we might, in some situations, say that the spark caused the fire, but the presence of oxygen was a mere background condition, and in other situations, perhaps in deep space, we might say that the presence of oxygen caused the fire. But in fact both have equal claim to being causes. Similarly, it might be said, for wrong-makers and wrong-enablers. What we pick on as significant or salient depends on our interests, and there is nothing metaphysically privileged about one element rather than another element. Call this view 'pragmatism'.[8]

[7] Statements of moral dependence factual conditionals will typically deploy names and not universal quantification; for example: If Isabella tortures, then she does wrong.

[8] This is also Caj Strandberg's suggestion in his contribution to this volume.

In order to reply to pragmatism, let us separate three cases. The first two cases are what I had in mind in Section 2. And the last is the sort of case that pragmatists use to argue that the maker/allower distinction is not fundamental, and merely reflects our interests.

Case 1 (a) X intentionally caused pain to a child (=A): (b) X did not intentionally cause pain for the child's medical benefit (=not-B); (c) X is not a doctor (=not-C). That is: A&¬B&¬C

Case 2 a doctor intentionally causes pain to a child for the child's medical benefit. That is: A&B&C. (Perhaps the doctor needs to check that the child's nervous system is working.)

Case 3 a lazy or callous doctor fails to cause pain to a child for the child's medical benefit, which he knows would benefit the child and that is easily within his power to do. That is: ¬A&¬B&C.

It is the interpretation of case 3 that is important. The conditional criterion can be deployed. In case 1, which is the actual case, it is true that if A then X did wrong; and it is not true that if ¬B then X did wrong. Hence the criterial factual conditionals say that A is a wrong-maker and ¬B is a wrong-allower. In case 2, a doctor causes pain in order to make sure that something is working properly. In that case, the conditional if X caused pain for medical benefit then X did right (if B then X did right) holds. So B is a right-maker. (I am not sure what to say about the intentional causing of pain in that case; I am not sure whether it is a right- or wrong-maker, but I am not sure whether much hangs on the question.) Case 3 is the important one—that of the callous or lazy doctor. In *that* case, yes, it is a conditional fact that if X did not cause pain for medical benefit then X did wrong (if ¬B then X did wrong). The fact that X did *not* intend to medically benefit *is* a wrong-maker in that case. However, the fact that the conditional fact holds in case 3 does not mean that the same conditional does not fail in case 1. So ¬B is a wrong-maker in one case but not the other.

This is a *contextual* account, which is quite different from a pragmatic account. In case 1, in the actual case, the criterial conditional [if ¬B then X did wrong] holds. It is true, of that case, that *if* X were a lazy or callous doctor, X would be wrong not to benefit the child by causing pain. But that does not mean that in the actual case, in which X is *not* a doctor, failing to medically benefit someone makes an act wrong. The conditional fails *there*. So failing to medically benefit can be an allower in a non-doctor case and a wrong-maker in a doctor case. And this has nothing to do with our interests.

The view here is similar to Alvin Goldman's treatment of his famous barn example, where Henry is surrounded by fake barns but happens to be looking at a rare real barn (Goldman 1976). In this case we are invited to think that Henry does not know that he is looking at a barn. In this example, whether or not we know depends on whether there are actually many fake barns nearby. Or, in a structurally similar but more everyday example from Fred Dretske, whether we know that the kind of bird we are looking at is a Gadwall duck can depend on whether or not there has been an unusual freak migration of Siberian Grebes into the area (Dretske 1982). Knowledge is dependent on the actual context. That is, the actual context matters—not our interests, which is a question of how it is with the person making the knowledge attribution. That latter is a psychological matter that has nothing to do with the actual situation being evaluated. A context-dependent account is quite different from a pragmatic account, and the contrast allows us to validate the making/allowing distinction.[9]

5. CONTINGENT DEPENDENCE

Let us return to the question of the modal status of the dependence-conditionals that we are committed to in making moral judgments. Are these conditionals necessary? The tempting thought—which I am recommending that we should resist—is that the conditionals are necessary and they are instances of the necessary conditionals that figure in the consequent of the strong supervenience principle.

Instead, I propose that when we make a moral judgment, we must have in mind some specific dependency, but we need not have to have in mind some specific sufficiency. While we have in mind deliberately causing pain as that which makes the act bad, we do not have in mind all the possible defeaters without which it would not have been bad. We do not have in mind all the negative natural properties that must be conjoined with the positive natural property of being a deliberate pain-causing in order to yield a conjunctive property that suffices for wrongness.

This means that moral dependence conditionals are *contingent*. This might seem strange. How can dependence be contingent? There is no denying that this seems odd. Dependency relations are intuitively stronger than necessary relations, and in many cases explain them. Surely—we might think—dependence implies necessity but necessity does not imply

[9] The factual conditional approach may help with the issue over hybrid conjunctive properties that we considered at the end of the last section. It could be argued that the hybrid property generates both factual and counterfactual conditionals whereas the negative conjunct of the hybrid conjunction alone does not generate factual conditionals. That might vindicate the idea of positive moral power and negative moral impotence. The wrong-maker seems to be the conjunctive fact: driving *and* failing to take care.

dependence. Consider Kit Fine's example of Socrates and the singleton set {Socrates}—the set with only Socrates as a member: there is a two-way necessity relation between the existence of these two things but only a one-way dependency relation between them (see Fine 1995). That is: necessarily Socrates exists if and only if the set exists; but the set depends on Socrates, not vice versa. Or consider the properties of being good and being what God would approve of. These are necessarily coinstantiated. But many think that even so there is only a one-way dependence relation. Divine commandment theorists run the dependence one way, while their opponents—autonomists about morality—run the dependence in the other direction. The divine commandment theory says that things are good because God approves of them while the autonomist thinks the opposite. But both sides agree on the necessary coinstantiation. (Necessity is really quite a *weak* relation!) By contrast with these two cases, however, in the cases of moral dependence, it seems that dependence relations are *contingent*. There is no doubt that the idea of contingent dependence seems somewhat oxymoronic.

We can be helped to feel more comfortable with contingent moral dependence if we compare moral dependence with causal dependence. We might hold that a spark caused an explosion, even though it was not sufficient for the explosion. Various (positive) background conditions were part of the overall state of affairs that *did* suffice for the explosion.[10] Had there been no oxygen, then the explosion would not have occurred. Nevertheless, to think that the spark caused the explosion is (typically) to hold the conditional that if there were a spark then there would be an explosion. But that conditional holds neither with nomological nor with causal necessity, even though it is generated by a dependency relation.

The contrary view would be that in ordinary moral judgments we isolate a dependency that *is* sufficient. For example, we might think that something is wrong because it is torturing for fun. Maybe there are conceivable saving-the-world scenarios where there are other factors that are counterfactually relevant in that had they obtained, which they did not, it would have been alright to torture for fun. But it might nevertheless be maintained that torturing for fun is still necessarily wrong (that is, it suffices for wrongness). It is just that this wrongness is what is called 'pro tanto' wrongness, which can be outweighed by other values, such as the value of saving-the-world. But the necessity still holds. Where this kind of outweighing scenario obtains—if it does—dependence coincides with sufficiency.

However, it is unlikely that that all cases of moral dependence coincide with sufficiency. Although there may be some cases, like the torturing for

[10] Or in cases of probabilistic causality, that the spark plus background conditions suffice for the probability of the explosion. I assume this qualification in what follows.

fun case, where the wrong-maker also suffices for wrongness, in many other cases the situation is not like this. In many cases, there are *negative* natural properties without which the wrong-maker would not have sufficed, even *pro tanto*, for wrongness but which are intuitively not wrong-makers. In the earlier example, we judge that an act was wrong because it was an act of intentionally causing pain. Being an act of intentionally causing pain is the wrong-maker. But there are many defeaters that did not obtain, such as the fact that the act was *not* an act of causing pain in order to bring medical benefit, in which case it would not have been wrong. If it had *not not* been an act of causing pain for medical benefit, then it would not have been wrong. And there are countless other possible defeaters. So what sufficed for wrongness in this case is *being* an act of intentionally causing pain, *and not* being an act of intentionally causing pain in order to bring medical benefit *and not* being an intentionally causing pain that is just punishment *and not* being ... and so on.

Hence, dependence and sufficiency come apart in our moral thinking. Oxymoronic as it may seem, we have to come to terms with contingent dependence in morality.

Two corollaries deserve mention. First, the notion of contingent dependence and of a making/allowing distinction is, or should be, important in epistemology—another normative domain. For example, it allows—what some may find unthinkable—that, for some important notions of 'is', the existence of Gettier cases does not show that knowledge is not justified true belief. Knowing might *depend* on justified true belief, and being justified true belief might *make* something knowledge, even though justified true belief is not *sufficient* for knowledge. Gettier cases refute an identity claim; but there are many other robust relations that they may not threaten, such as dependence or constitution. The only epistemological accounts in the Gettier literature that I know of that have the potential to recognize the making/allowing point are 'defeasibility' accounts (for example, Lehrer and Paxon 1969). But such accounts typically made the mistake of seeing not-being-defeated as a fourth 'condition' for knowledge—that is, as a knowledge-making factor. Making a making/allowing distinction might make a large difference to epistemology.

Secondly, there are consequences for 'projectivist' views of moral thought. In this paper I have assumed realism about moral properties. However, a projectivist view, which sees moralizing as a matter of having attitudes as opposed to beliefs about moral properties, might nevertheless allow that we speak and think as if there were moral properties. Such a 'quasi-realist' theory might attempt to understand our a priori commitment to moral–natural supervenience as arising from a constraint of consistency among our attitudes (Blackburn 1985). The aim would be to show why we speak and think as if moral dependence holds as an a priori constraint, and perhaps also why it is right for us to do so (Zangwill 1993). This can only be achieved, I think, if it is insisted that the systemization of a set of attitudes is an essential condition of their being *moral* attitudes. What makes an attitude a moral attitude is not, on this view, an intrinsic property of an attitude any more than being a soldier is an intrinsic property of a certain person. On this view, moral/natural supervenience arises from the practical necessity of operating with a systematic set of attitudes, either as individuals or as a group. Morality could not serve its purposes unless we imposed consistency. The trouble is that this systemization strategy will only deliver the idea of supervenience, not that of dependence. Presumably the projectivist will want to make the making/allowing distinction. But how is that to be done? Of course, our attitudinal reactions are *to* the natural features of things; for example, we find certain natural properties pleasurable. But that will not suffice to explain the more robust idea of dependence. Projectivists need to explain and justify an a priori principle of moral dependence as distinct from moral supervenience.

6. A PRIORI DEPENDENCE AND A PRIORI SUPERVENIENCE?

At first sight, it seems hard to square the contingency of moral dependence with the principle of strong supervenience. What is the relation between these relations? Why should we think that dependence and supervenience go together? And why should we think that an a priori commitment to dependence goes along with an a priori commitment to supervenience? Maybe they have nothing to do with each other. But if so, what then has become of the idea that strong supervenience is an overarching *a priori* framework principle governing our moral thought? Moral dependence seems to be obviously a priori, as we saw earlier. But there is a philosophical tradition of claiming that strong supervenience is a priori. How then are these ideas related?

My view is that although the principles are distinct, they are closely related, and both are a priori. In making moral judgments we are committed to two ideas: when we judge something to have an *M* property M, we are committed *both* to there being some *N* property N# that *makes* the thing M, *and* we are also committed to there being some *N* property N* that is *sufficient* for M. In at least most cases, N# and N* are not identical. If so, the conditional if N# then M does not hold necessarily. For there are many other *N* features of the thing that is M, and only the *N* property N# *together* with these other features is sufficient for the instantiation of M. It is possible to have N# instantiated plus or minus various other *N* properties, such that M might not be instantiated without them. Nonetheless, as things actually were, N# *made* the thing M, even though N# was *not* sufficient for M in the way that N* is. But N# is a conjunct of N*, which is sufficient for M.[11] That is, if we judge that something x has an *M* property M, then we must judge (that is, it is a priori) that:

(1) x has some *N* property N# such that Mx because N#x, and

(2) x has some conjunctive *N* property N*, such that N# is a conjunct of N*, and necessarily anything that is N* is M.

By the a priori strong supervenience principle, we *do* indeed know that there is *some N* property, N*, which *is* strictly *sufficient* for M. But what we need is the claim that if we think that *some N* property N# is responsible for M then we also think that N# is a conjunct of a complex *N* property N* such that N* suffices for M.[12] For example, where we think that an act actually is bad because it is a deliberate pain-causing we do also think that *something* sufficed for it—deliberate pain-causing *plus* other conditions. There must be *some* sufficiency nearby. So I think that it would be an over-reaction to conclude that dependence and supervenience had nothing to do with each other.[13]

[11] The N* natural property is the conjunction of all the natural properties that the thing possesses that together *suffice for* the moral property. This conjunction will typically include negative natural properties. However, it is unlikely to include absolutely all of its positive natural properties. The exact position of atoms does not matter much morally. N* natural properties need not be what are called 'total' natural properties, which are the conjunction of *all* of a thing's natural properties.

[12] Am I endorsing the idea that moral dependence conditionals are *ceteris paribus* or *pro tanto* moral truths of the sort that David Ross liked? I don't think so. The moral judgment says that a certain thing is M in virtue of being N, and there is some condition C such that N and C suffice for M. In this, there is no commitment to an array of *pro tanto* principles that might combine together to yield an overall, or all things considered, judgment.

[13] To express this commitment, we could harmlessly semantically ascend and talk in terms of truth, which is the preferred genre of many of those who discuss essentialist issues. In these terms, we could say that moral conditional judgments of the form [If something is N then it is M] may not themselves hold true with metaphysical necessity. But, where C is a conjunction of N and all of a thing's other natural properties, then the conditional judgment [If something is N and C then it is M] *does* hold true with metaphysical necessity. It does not follow that *moral* knowledge is a sub-class of *modal* knowledge; for our moral knowledge is of the *truth* of the conditional, not of its *necessary* truth. The knowledge of the moral conditional is added to an a priori framework principle that says that if such a moral conditional judgment is true then it or its cousin (with an enhanced antecedent) is necessary if true. This supervenience framework principle is known a priori; it is constitutive of what it is to make moral judgments. Supervenience principles can in this way be (re)cast in terms of truth of moral and natural judgments (or 'sentences' or 'statements' or 'propositions'). However, I believe

The lurking discomfort with the idea of contingent dependence, in both morality and causality, can be eased somewhat by accepting that a dependency entails that there is some necessary relation nearby. In both the cases of morality and causality, if A depends on B, then B is *part* of a condition C such that necessarily if C then A. If A depends on B—morally or causally—then B has a cousin C, of which B is a part, that is sufficient for A.[14] So dependence and necessity are indirectly connected.

Despite the important similarities between morality and causality, it seems that causation and morality differ in some respects. These are similar in so far as we should allow that something is a cause even though it is not sufficient for the effect, so long as a conditional holds. We can allow for contingent causal dependence, so long as the cause is part of a sufficient condition. On the other hand, in the case of causation, there is a sense in which all of the conditions that are part of the overall sufficient condition have equal status, and this is unlike the moral case, if there is an making/allowing distinction. The distinction between background causal conditions and foreground causes is only pragmatically significant (varying with our interests). But this is unlike the distinction between makers and allowers in morality, which is a metaphysical rather than a pragmatic distinction.

7. A PRIORI MORAL SUPERVENIENCE

In the last section, I assumed the a priori strong supervenience framework principle and I worried about its relation to the principle of a priori moral dependence. But have we now lost the rationale for believing a priori strong supervenience? If we have a priori dependence, why do we also need a priori strong supervenience? Perhaps strong supervenience is true, but why should it be a priori? Maybe it is enough that moral dependence is a priori. Why do we need the idea that we have an a priori commitment to necessities linking moral and natural properties? Maybe we should give up the a priority of strong supervenience. Moreover, it might even be suggested that we should *abandon* the modal doctrine altogether. Perhaps it is not only not a priori but also not true at all! Why must every dependence relation be accompanied by sufficiencies? Although we know a priori that

in the principle of 'semantic gravity': whatever goes up must come down! That is, one can 'semantically ascend' and talk in terms of truth; but only because of non-semantic facts about objects and properties.

[14] This is Mackie's INUS account (Mackie 1965). Even in a probabilistic framework, B&C is sufficient for the probability of A's occurrence.

every instantiation of a moral property is due to some right-making natural property, why must there be some natural property that is sufficient for the instantiation of the moral property? So we have two disturbing questions: why hold that strong moral supervenience is a priori? And: why hold that it is true at all?

It is true that the claim of strong moral supervenience is not *as obviously* a priori as moral dependence. But there are other very closely related modal claims, also labeled 'supervenience' claims, which *are* obviously a priori. For example, it is obviously a priori that if something *changes* morally then it must change naturally, and that *two things* that are morally different must be naturally different. These are what Jaegwon Kim has called 'weak' supervenience claims, since they only make claims about what must hold, as he put it, 'within a world' (Kim 1984). But it is also obviously a priori that something *could* not be morally different from how it actually is unless it were naturally different. This obviously a priori modal claim takes us 'across worlds'. But this obviously a priori cross-world modal claim is only inches away from the not obviously a priori strong moral supervenience principle to the effect that if something has a moral property then it has some natural property that suffices for that moral property.

Let us attempt to traverse these inches. The obviously a priori cross-world claim does not imply that things with moral properties must have natural properties, unlike the strong a priori moral supervenience principle (although it does imply that there could not be many things that differ morally that all had no natural properties). But this *is* implied by the a priori dependence principle, which is obviously a priori. So let us try *conjoining* the obviously a priori dependence principle with the obviously a priori cross-world claim. Perhaps together they will yield the not obviously a priori strong moral supervenience principle. How so?[15]

The a priori dependence principle means that something with a moral property must have a natural property. And the modal principle says that moral differences (across worlds) imply natural differences. Contraposing, that implies that that complete natural similarities (across worlds) imply moral similarities. But how does this show that for all moral properties, there are natural properties that suffice for those moral properties? Well, we know from the obviously a priori dependence thesis that something with moral properties must have natural properties. Moreover, if something has natural properties then it is safe to assume that it has some *total* natural property, where a total natural property is the conjunction of all of a thing's

[15] The argument of this section does not depend on discriminating between positive and negative natural properties.

natural properties. It is not too controversial to add that. But that total natural property is sufficient (more than sufficient) for the moral property. So, given the obviously a priori cross-world modal claim, it means that things with moral properties must have natural properties that are sufficient for them. This natural property will not be the same natural property as the one that the moral property depends on. Still, something with a moral property must have some natural property that suffices for it, even though the natural property that it depends on does not. Thus the strong supervenience principle is after all a priori, even if it is not as obviously a priori as some other claims.

8. SUI GENEROSITY

Saul Kripke famously defended principles to the effect that if identity statements between proper names are true then they are necessarily true and that if identity statements between natural kind terms and terms for molecular structures are true then they are necessarily true. He *also* claimed that these conditional principles were known "a priori by philosophical analysis" (Kripke 1980: 109; see also Sidelle 1989). Nathan Salmon, Keith Donnellan, and Hilary Putnam were surely right to object that this epistemological claim is false in the case of natural kinds: although the conditional modal principle is true, it is a posteriori, not a priori (see Salmon 1982, appendix II, citing unpublished work by Donnellan; and Putnam 1992). It seems like something we once did not know and that we have discovered as part of the growth of scientific understanding of what natural kinds, such as water and gold, actually are. In particular, we were only in a position to know the conditional modal principle for natural kinds given the advent of Daltonian chemistry. The ancient Greeks did not know it. Hence the principle is known empirically and is not known a priori.

We must distinguish two claims:

(A) If something is water *and* has some molecular micro-structure *then* necessarily anything with that molecular structure is water.

(B) If something is water *then* it has some molecular micro-structure such that necessarily anything with that molecular structure is water.

Perhaps most ancient Greeks would have accepted (A) if someone had gone back in time and put it to them and explained it. But surely most ancient Greeks would have rejected (B), since they thought that water was a basic substance. Thales may not have been an average ancient Greek, but he would certainly have rejected (B), since he thought that water was *the* basic substance. Given that Thales and many fellow ancient Greeks were perfectly good water-thinkers, but thought of water as a basic substance, or in Thales' case, *the* basic substance, (B) is very unlikely to be an a priori conceptual constraint.[16] If so, then Kripke seems to be either wrong or misleading about the a priori and conceptual status of his claim about natural kinds, although he may still be right about the a priori and conceptual status of the claim about proper names. In the moral case, however, the modal framework principle—moral supervenience—*is* known a priori and conceptually, along the lines that Kripke had in mind for both natural kinds and proper names. Hence, in this epistemic respect, moral supervenience is unlike the supervenience of natural kinds on molecular structure.[17]

The same is true of moral dependence: moral dependence is an a priori conceptual constraint on thinking in moral terms. In this respect, moral properties contrast with both natural kinds and sensory properties. It is not knowable a priori that water depends on anything. Thales may have been wrong but he was not confused to think that nothing makes water water. Similarly with colour. It may be that something makes blue things blue. There are presumably physical properties of the surfaces of things, or of light reflectance, or of standard observers, that make blue things blue. But in order to think in terms of blueness, not only need we not know what those blue-making properties are, we need not think, explicitly or tacitly, that there are any such properties. Blueness might, for all we need to know, be a primitive property of things. It was a discovery that it is not. But moral dependence is unlike those two cases, for we know a priori that moral properties are not basic and that they depend on other properties.

I believe that the implications of this are significant. It means that we should distance ourselves from two currently popular realist metaethical schools. A common American approach models moral kinds on natural kinds, and moral/natural dependence on natural kind/molecular structure dependence. And a common British approach models moral kinds on sensory kinds and moral/natural dependence on the dependence of sensory properties on whatever sensory properties are thought to depend on (which varies with different theories). But both analogies are flawed because moral dependence has quite different epistemic characteristics from natural kind

[16] Here I *may* be distancing myself from what has recently been called "two-dimensional semantics".

[17] Perhaps there are *some* terms for natural kinds that do imply a certain micro-structural composition; "DNA" might be an example. However, while there may be some natural kind terms that are bound by a priori principles of composition, they could not all be like that. Furthermore, whether such natural kind terms have actual instantiations is an empirical question.

and sensory property dependence. This suggests that we need a moral theory that does not model itself on these alleged analogies and that what we need is a more *sui generis* theory. Moral kinds may be a distinctive kind of kinds; and moral dependence may be a distinctive kind of dependence. Or moral kinds and moral dependence may be of a broader normative kind. Either way, moral kinds and moral dependence are very different from non-normative kinds and non-normative dependence.

REFERENCES

Blackburn, Simon (1985) "Supervenience Revisited" reprinted in *Essays on Quasi-Realism* (Oxford: Oxford University Press).

Dancy, Jonathan (2004) *Ethics Without Principles* (Oxford: Oxford University Press).

Dretske, Fred (1982) "The Pragmatic Dimension of Knowledge" reprinted in *Perception, Knowledge and Belief: Selected Essays* (Cambridge: Cambridge University Press, 2000).

Fine, Kit (1995) "Ontological Dependence" *Proceedings of the Aristotelian Society* 95, 269–90.

Foot, Philippa (1958) "Moral Arguments" reprinted in her *Virtue and Vices* (Oxford: Oxford University Press, 1978).

____ (1959) "Moral Beliefs" reprinted in her *Virtue and Vices* (Oxford: Oxford University Press, 1978).

Goldman, Alvin 1976 "Discrimination and Perceptual Knowledge" *Journal of Philosophy* 73/20: 771–91.

Kim, Jaegwon (1984) "Concepts of Supervenience" reprinted in his *Supervenience and Mind* (Cambridge: Cambridge University Press, 1993).

Kripke, Saul (1980) *Naming and Necessity* (Cambridge, MA: Harvard University Press).

Lehrer, Keith, and Paxon, Thomas, Jr. (1969) "knowledge: Undefeated Justified True Belief" *Journal of Philosophy* 66, 225–37.

Mackie, John (1965) "Causation and Conditionals" *American Philosophical Quarterly*.

____ (1977) *Ethics* (Harmonsworth: Penguin).

Moore, G. E. (1903) *Principia Ethica* (Cambridge: Cambridge University Press).

Putnam, Hilary (1975) "The Meaning of Meaning" in *Philosophical Papers, ii. Mind, Language and Reality* (Cambridge: Cambridge University Press).

____ (1992) "Reply to Alan Sidelle" *Philosophical Topics*.

Salmon, Nathan (1982) *Reference and Essence* (Oxford: Blackwell).

Sibley, Frank (1965) "Aesthetic-Nonaesthetic" reprinted in his *Approach to Aesthetics* (Oxford: Oxford University Press, 2001).

Sidelle, Alan (1989) *Necessity, Essence, and Individuation: A Defense of Conventionalism* (Ithaca, NY: Cornell University Press).

Stocker, Michael (1970) "Moral Duties, Institutions and Natural Facts" *Monist* 54: 602–24.

Nick Zangwill (1993) "Quasi-realism, Justification and Explanation" *Synthese* 97: 287–96.

——— (1995) "Moral Supervenience" *Midwest Studies in Philosophy* 20: 240–62.

——— (2003) "Negative Properties, Conditionals and Determination" *Topoi* 22: 127–34.

——— (2005) "Moral Epistemology and the Because Constraint" in *Contemporary Debates in Moral Theory*, ed. Jamie Dreier (Oxford: Blackwell).

This page intentionally left blank

6

Particularism and Supervenience

Caj Strandberg

1. INTRODUCTION

One of our most fundamental notions of morality is that in so far as objects have moral properties, they have non-moral properties that *make* them have moral properties. Similarly, objects have moral properties in virtue of or because of having non-moral properties, and moral properties depend on non-moral properties. In ethics it has generally been assumed that this relation can be accounted for by the supervenience of moral properties on non-moral properties. However, this assumption is put into doubt by an influential view in contemporary ethics: particularism. Thus, one of particularism's most important implications is thought to be that supervenience is incapable of accounting for the notion that non-moral properties make objects have moral properties. At least, this is what Jonathan Dancy, the leading proponent of particularism, argues in his recent book *Ethics Without Principles*, and elsewhere.

In the present paper, I defend supervenience against this challenge. That is, I argue that particularism does not threaten the ability of supervenience to account for the notion that non-moral properties make objects have moral properties. While doing so, I hope to contribute to our understanding of what is involved in this notion. In the next section, I consider a general argument put forward by Dancy against supervenience and criticize his

I am grateful to participants of various conferences and seminars in Lund, Gothenburg, Lisbon, Madison, and Chapel Hill for valuable discussions on earlier versions of this paper. Special thanks for helpful comments are due to Gunnar Björnsson, Ragnar Francén, Elín Hafsteinsdóttir, Sven Nyholm, Ingmar Persson, Wlodek Rabinowicz, Geoff Sayre-McCord, Jussi Suikkanen, and two anonymous referees for Oxford University Press. A substantial part of my work with the paper was supported by a research grant from the Bank of Sweden's Tercentenary Foundation.

alternative, resultance. In Section 3, I develop a version of supervenience that I call Specific Moral Supervenience, SMS, and which I think avoids Dancy's argument. There are basically two conceptions of particularism: what is known as 'holism' and the contention that there are no true moral principles. In Section 4, I argue that the view that SMS provides a basis for an account of the notion that non-moral properties make objects have moral properties is compatible with the pertinent version of holism. However, in Section 5 we see that SMS is incompatible with the view that there are no true moral principles. Particularists find support for this view in the distinction between non-moral properties that make objects have moral properties and so-called enablers. On Dancy's conception of this distinction, it follows that SMS does not refer to non-moral properties that make objects have moral properties and that there are no true moral principles of the relevant kind. In Sections 6 and 7, I defend SMS against these two consequences. In doing so, I distinguish two uses of 'make' and provide a pragmatic account of the distinction between non-moral properties that make objects have moral properties and enablers.

2. GENERAL MORAL SUPERVENIENCE AND RESULTANCE

Dancy formulates the version of supervenience that he focuses on roughly in the following way:

General Moral Supervenience (GMS) It is necessary that if an object has a moral property, then any other object which shares all the non-moral properties with the first object has the moral property too.[1]

According to this principle, Dancy contends, the 'supervenience base ... consists in *all the non-moral features*' of an object.[2] He then argues that even though the principle holds, it fails to account for our notion of the way in which non-moral properties make objects have moral properties. The reason is that we do not believe that it is *all* of an object's non-moral properties that make it have a moral property; on the contrary, we assume that it

[1] Dancy (2004: 86). Dancy does not explicitly claim that the principle he is concerned with holds with necessity, but what he writes indicates that he understands it in that way. As formulated here, GMS is a version of weak supervenience; however, this does not affect Dancy's argument. For a useful discussion of different versions of supervenience, see McLaughlin (1995: 16–59).

[2] Ibid. See also Dancy (1981: 381), and Dancy (1993: 78). Cf. Grimes (1991: 88–9), and McKeever and Ridge (2006: 8).

might be the case that only *some* of an object's non-moral properties have this function. Dancy therefore concludes that supervenience fails to account for the notion at issue. As an alternative to supervenience, he introduces the concept of resultance which he claims does not have this shortcoming. His argument for this contention, when applied to wrongness, is that '[t]he "resultance base" for the wrongness of a particular action consists in those features that make it wrong, the wrong-making features', not all of the action's non-moral features.[3]

There is reason to believe that GMS fails in the indicated manner. However, I do not think we should stay satisfied with resultance. When Dancy characterizes this concept, he does so in terms of 'make', 'in virtue of', 'because' and 'depend'.[4] Indeed, Dancy admits that resultance resists further explication. This is problematic for various reasons. Most obviously, it means that resultance is uninformative since it does not provide us with any account of the relation between non-moral and moral properties that improves our understanding of what this relation involves. In fact, Dancy characterizes resultance by using the very terms we were hoping that the concept would illuminate. Moreover, it might be argued that this makes resultance vulnerable to a version of J. L. Mackie's argument from queerness. Mackie famously argues that unless it is possible to explain 'what *in the world* is signified by this "because"', this relation is metaphysically queer.[5] He believes that no such account can be provided and concludes therefore that this relation never is instantiated, in which case there are no moral properties.

3. SPECIFIC MORAL SUPERVENIENCE

In the last section, we saw that there are reasons to believe that neither GMS nor resultance succeeds to account for the notion that non-moral properties make objects have moral properties. However, I think there is a version of supervenience which is more promising in this respect. Consider:[6]

[3] Ibid.

[4] See e.g. Dancy (1981: 367, 380–2), Dancy (1993: 73–7), and Dancy (2004: 85–9).

[5] Mackie (1977: 41). It might be replied that Mackie's argument fails since if it were correct, it would apply to all dependence relations between properties in which case it would follow, implausibly, that there are no properties which depend on other properties. For related objections, see Brink (1989: 173–4), and Shafer-Landau (2003: 88). Unfortunately, I cannot discuss Mackie's argument here.

[6] As easily can be seen, (i) in SMS is a version of one of Jaegwon Kim's formulations of strong supervenience; see e.g. Kim (1993 (1984): 57–67). When I refer to a set of

Specific Moral Supervenience (SMS): (**i**) It is necessary that, for any object x, and for any moral property M, if x has M, then there is a set of non-moral properties G such that (**A**) x has G, and (**B**) it is necessary that, for any object y, if y has G, then y has M.

(**i**) says, roughly put, that, necessarily, if an object has a moral property, it has some set of non-moral properties which is such that, necessarily, whatever object has that set of non-moral properties has the moral property. Although (**i**) constitutes the basic part of SMS, this principle should be understood to include at least two further claims.

Let us first observe that SMS should state a set of non-moral properties which does not contain any superfluous elements. A set of non-moral properties G should in other words contain those, and only those, elements that bring about that a given object has the moral property in question. In particular, it should contain just as many non-moral properties that are sufficient for an object to have the moral property, but not more. In order for SMS to meet this demand, we may add the requirement that there is no part of G which can be included in the formula instead of G and yet preserve its truth. Thus, SMS should be understood to include the following claim:

(**ii**) There is no proper subclass of G, G*, such that if G* is substituted for G, (**i**) is true.

Let us next observe that SMS should state an asymmetric relation between non-moral and moral properties. That one kind of properties makes objects have another kind of properties is an asymmetric relation; hence, non-moral properties make objects have moral properties but not the other way around. As SMS is formulated so far, however, it expresses neither a symmetric nor an asymmetric relation. In order to be able to account for this notion, SMS should therefore include some kind of asymmetry claim. It is a difficult issue how SMS should be developed to meet this demand and I cannot give it sufficient attention here. However, one simple suggestion is to add the requirement that the reverse relation between the properties in question does not hold.[7] On this proposal, SMS should be understood to include the following claim:

(**iii**) It is not the case that the reverse of the relation between moral properties and non-moral properties holds, where the relation is of a kind stated in (**i**).

non-moral properties, such as G in SMS, I have in mind a non-empty set which consists of a single non-moral property or a conjunction of such properties.

[7] Cf. Kim (1993 (1990): 144–5).

This requirement can presumably be spelled out in different ways. However, it seems reasonable to understand it simply as the negation of the relation between non-moral properties and moral properties stated in (**i**). Thus understood, (**iii**) claims that the following is not the case: It is necessary that, for any object x, and for any set of non-moral properties F, if x has F, then there is a moral property M such that x has M, and it is necessary that, for any object y, if y has M, then y has F.[8]

It might be objected that the latter claim is too weak to secure the required asymmetry. If (**iii**) is understood in this way, SMS implies that there is a general asymmetric relation of a certain kind between moral properties and non-moral properties. However, one may want to argue that this does not guarantee that the relation between a particular set of non-moral properties G and a particular moral property M is asymmetric and, as a consequence, that it is insufficient to account for the notion that the former makes objects have the latter. According to (**B**), it is necessary that, for any object y, if y has G, y has M. In order to secure the required asymmetry, it may therefore be tempting to suggest that the reverse relation should be ruled out. On this proposal, SMS should be understood to include a claim to the effect that the following is not the case: It is necessary that, for any object y, if y has M, y has G. I think, however, that such a requirement would be too strong. On this requirement, a moral property M cannot be necessarily coextensive with a single set of non-moral properties G. According to a common view of property identity, it follows that a moral property cannot be identical to a non-moral property.[9] However, it seems consistent with a correct use

[8] If (**iii**) is understood in this way, SMS implies that there is a certain general asymmetry between moral and non-moral properties. According to (**i**) in SMS, it holds, necessarily, for *all* moral properties that if an object has a moral property, it has some set of non-moral properties which is such that, necessarily, whatever object has that set of non-moral properties has the moral property in question. However, according to (**iii**) in SMS, it is *not* the case that it holds, necessarily, for *all* sets of non-moral properties, that if an object has a set of non-moral properties, it has some moral property which is such that, necessarily, whatever object has that moral property has the set of non-moral properties in question. It should be noted that this claim leaves open the possibility that the last-mentioned relation holds for *some* sets of non-moral properties. As a consequence, understood in this way SMS is compatible with a moral property being necessarily coextensive with a single set of non-moral properties and, consequently, with the possibility that a moral property is identical to a non-moral property.

[9] Kim argues that strong supervenience entails that the supervenient property is necessarily coextensive with a property which consists of the disjunction of the various sets of properties it supervenes on. See e.g. Kim (1993 (1990): 150–5). (For a similar argument, see Jackson (1998: 122–3).) As noted earlier, (**i**) in SMS is a version of one of Kim's formulations of strong supervenience. According to Kim's argument, a moral property M is consequently necessarily coextensive with a non-moral property which consists of a disjunction of the various sets of non-moral properties it supervenes on

of 'make' to claim that a certain non-moral property makes objects have a certain moral property even if one believes that they are identical. To see this, suppose that a utilitarian claims that what makes actions right is that they maximize happiness and suppose further that she believes that rightness consists in maximizing happiness. As far as I understand, we would not object that her use of 'make' is erroneous and this fact indicates that a correct use of the term is compatible with the view that identity is the case.[10] Now, identity is evidently a symmetric relation. It is therefore legitimate to ask how it comes that it is legitimate to use the term in this way in spite of one believing in the identity between a moral property and a non-moral property. This is, however, a difficult issue that I cannot deal with satisfactorily here. However, one answer which suggests itself is that there is a general asymmetric relation between non-moral and moral properties of the kind indicated above.[11]

It seems reasonable to assume that many metaethicists embrace, explicitly or implicitly, principles like SMS and that it is compatible with various metaethical views. There are for example different ways of understanding the occurrences of 'necessary'. Moreover, SMS does not entail that there is one particular set of non-moral properties which objects must have whenever they have a moral property; it leaves in other words room for moral properties being multiply realizable. As already indicated, the preferred understanding of SMS is also compatible with moral properties being identical to non-moral properties.

(G_1, G_2, etc.). It might be claimed that this means that a moral property can be identical with a non-moral property *even* if there is no necessary implication from a moral property M to a *single* set of non-moral properties G. This is an important argument which I cannot do justice to here. But it should be noticed that it is highly controversial whether properties are closed under disjunction and that it is generally assumed that necessary coextension is not sufficient for property identity.

[10] Cf. Post (1999: 320–5).

[11] Cf. Kim (1993 (1984): 67). Another possible answer is that there is some kind of non-metaphysical asymmetry between non-moral and moral properties. In particular, there might be an explanatory asymmetry in virtue of the fact that non-moral properties are explanatorily prior to moral properties. Cf. Post (1999: 325–30). Consider the following statements: (1) 'What makes actions right is that they maximize happiness' and (2) 'What makes actions maximize happiness is that they are right'. (1) seems more plausible than (2). One reason might be that we understand (1) and (2) as explanations and that we take (1) to be a more plausible explanation than (2). One reason for the latter view may in turn be the idea that explanations in terms of the non-moral property are part of a comprehensive ethical theory, such as utilitarianism, whereas explanations in terms of the moral property are not. As a result, we believe that the former are explanatorily more powerful—explain more phenomena—and informative than the latter. As a result (1) seems more plausible than (2). I will briefly return to this idea in Section 7.

As these remarks suggest, it should be stressed that SMS as it stands does not provide a complete account of the notion that non-moral properties make objects have moral properties. Most obviously, in order to claim that it does, it would be necessary to say more about its different parts than I can do here; especially, the asymmetry requirement should be discussed more fully. Moreover, it would be vital to specify its various key elements; especially, the two occurrences of 'necessarily' should be identified.[12] In Section 7, I also suggest that SMS should be supplemented with certain pragmatic considerations in order to capture a use of 'make' which is essential when we claim that non-moral properties make objects have moral properties. There might also be other ways in which SMS needs to be amended. In addition, objections against it should be discussed and responded to.[13] However, if what I have said so far is fairly correct, SMS constitutes at least a basis for the required kind of account.

The point I would like to stress here is that this version of supervenience is not vulnerable to Dancy's argument against supervenience I mentioned in the last section. The kind of set of non-moral properties referred to in SMS—G—does not need to contain all of an object's non-moral properties but may contain only some of these. Hence, there is no reason to believe that SMS in this particular respect is unable to account for the notion that non-moral properties make objects have moral properties. Moreover, unlike resultance, SMS does not utilize 'make' and related terms. There is therefore reason to believe that, if SMS is correct, it contributes to an informative account of this relation which improves our understanding of it. Accordingly, SMS seems capable of avoiding the main argument against resultance. Since SMS as it stands does not provide a complete account of the relation at issue, what I have said above is insufficient as a response to the pertinent version of Mackie's argument from queerness. However, there is reason to believe that SMS, unlike resultance, at least supplies a base for a response to this worry.

Let us now turn to the question of what consequence particularism has for supervenience. It might be maintained, as Dancy evidently does, that particularism means that supervenience quite generally is unable to account for the notion that non-moral properties make objects have moral

[12] Most metaethicists would presumably agree that the first occurrence of 'necessary' should be understood as analytic necessity, but they disagree as to how the second occurrence should be understood. Elsewhere I argue for a particular reading of SMS; see Strandberg (2004). Among other things, I suggest that the first occurrence of 'necessary' should be understood as analytic necessity and the second occurrence as a certain kind of synthetic necessity.

[13] Principles like SMS have been put into question on various grounds; see e.g. Blackburn (1993 (1985)). I respond to such arguments in Strandberg (2004).

properties.[14] Thus, it might be argued that particularism implies that SMS fails to account for this notion. In the remainder of the paper, I will consider whether this is the case.

4. SPECIFIC MORAL SUPERVENIENCE AND HOLISM

According to one conception of particularism, it consists in the view Dancy calls 'holism':

> Particularism, I want to say, is an expression of a general holism in the theory of reasons; it is the application of holism to the moral case. Holism in the theory of reasons holds that a feature that is a reason in favour in one case may be no reason at all in another, and in a third may even be a reason against.[15]

As Dancy formulates holism here, it is a claim about the features that constitute normative reasons for performing actions.[16] However, Dancy assumes generally that what holds for reasons also holds for the mainly metaphysical make-relation. Hence, he believes that what holds for the features that constitute reasons to perform actions holds for the non-moral properties that make objects have moral properties. Dancy accordingly advocates holism understood as a claim about the non-moral properties that make objects have moral properties as well.[17]

Understood in the latter way, holism is the view that the relevance of non-moral properties is *context-dependent*. It can be formulated in the following way: a non-moral property which, when instantiated in one object, contributes to the object having a certain moral property, might, when instantiated in another object, contribute to that object *not* having the moral property, and might, when instantiated in yet another object, contribute in *neither* of these ways.[18] Suppose A is such a property. The reason why A's relevance varies in this way is that some of the object's other non-moral properties determine whether A is relevant and, if it is, which relevance it has. The relevance of A is thus context-dependent, where the context is made up by other non-moral properties of the object.

[14] See e.g. Dancy (1981: 373–5), Dancy (1993: 77–9), and Dancy (2004: 85–9).

[15] Dancy (1999*a*: 144). Cf. Dancy (2004: 7).

[16] Strictly speaking, what provide reasons for action are presumably facts or truths rather than properties. When referring to reasons in what follows, I have in mind normative reasons.

[17] See e.g. Dancy (2004: 78–80).

[18] For similar formulations of holism, see e.g. Kirchin (2003: 54–5), Alm (2004: 312), and Robinson (2006: 333–7).

Advocates of this view find support for it in various thought experiments of which the following is an example.[19] Suppose an action causes pleasure and that we think that it is right because it has that non-moral property. This fact may have us believe that causing pleasure always contributes to actions being right. To see that this is not at all evident, imagine that an action of punishing someone causes pleasure to those who are witnessing the action. In that case, we might be inclined to say, causing pleasure does not contribute to the action being right. Perhaps we would even say that it contributes to the action *not* being right. If this is correct, the relevance of causing pleasure varies depending on the context made up by other non-moral properties of the action, for example, that of being a punishment.

Now, I think that the view that SMS provides a basis for an account of the notion that non-moral properties make objects have moral properties is compatible with the relevance of non-moral properties being context-dependent. The basic reason is this. A set of non-moral properties of the kind referred to in SMS, and which makes objects have a moral property according to this claim, might consist of a number of non-moral properties. Within such a set, the relevance of a non-moral property might be context-dependent, where the context is made up by other non-moral properties in the set.

Consider the following simple account of how this can be the case. Suppose an object has the following set of non-moral properties: A & −B & C. Assume that this set of non-moral properties is of the kind referred to in SMS and that it consequently makes an object have a moral property M according to the view suggested here.[20] In this set, A might contribute to the object having M. This can be understood in the following way: in this set, A is such that if the object had not had A, it would not have had M. That is, if the object had had the other non-moral properties in the set (i.e. −B & C), but not A, it would not have had M.[21] However, suppose another object has the following set of non-moral properties: A & B & C. Assume that this

[19] For this example, see McNaughton (1988: 193), and Dancy (1993: 61). It is doubtful whether not being a punishment is relevant in the required way. An advocate of particularism might reply, however, that what is relevant is instead, say, not causing sadistic pleasure. For other examples, see e.g. Dancy (1993: 60–2), Sinnott-Armstrong (1999: 3–4), Crisp (2000: 36–7), Cullity (2002: 173–4), Kirchin (2003: 57–8), and McKeever and Ridge (2006: 27).

[20] A set of this kind may of course contain a vast number of non-moral properties; here I provide merely a simple example. It might be argued that properties are not closed under negation and, thus, that features such as '−B' are not real properties. This means that what makes an object have a moral property need not strictly speaking be a set of non-moral properties, but rather a set that contains such negative features.

[21] Put differently: if the object had not had A, and so had had the set −A & −B & C, it would not have had M. If one is sceptical towards such formulations, the following one might be preferred. Let *a* be the object in the example. We can then they say that, for any object, if it has −B & C, and is exactly similar to *a* as regards all other non-moral

set does not make the object have M. In this set, A might contribute to the object not having M. This can be understood in the following way: in this set, A is such that if the object had not had A, it would have had M. That is, if the object had had the other non-moral properties in the set (i.e. B & C), but not A, it would have had M. Or, in this set, A might contribute in neither of these ways.[22] Thus, whether A is relevant, and, if it is, which relevance it has, is determined by other non-moral properties in respective set of non-moral properties, i.e. A's relevant context. In the examples, the active part of the context is −B and B, respectively.

This account of the relevance of non-moral properties being context-dependent can be applied to the example above. Let the moral property, M, be 'rightness', A 'causing pleasure' and B e.g. 'being a punishment'. In the first set of non-moral properties (A & −B & C), A contributes to the object having M because it figures in a context partly made up by −B. In the second set of non-moral properties (A & B & C), A contributes to the object not having M, or contributes in neither way, because it figures in a context made up partly by B.[23]

5. SPECIFIC MORAL SUPERVENIENCE AND MORAL PRINCIPLES

According to the second main conception of particularism, it is the view that there are no true moral principles. Thus understood, particularism

properties, but does not have A, it does not have M. Similar formulations are available for the other cases described below.

[22] This can be understood in the following way. In this set of non-moral properties (A & B & C), A does not contribute to the object having M. Moreover, in this set, A does not contribute to the object not having M. In that case neither of the two conditionals mentioned in the text holds.

[23] The following objection can be directed against this account. The kind of set of non-moral properties that is claimed to make objects have a moral property might consist of a conjunction of non-moral properties. It is commonly assumed that properties are closed under conjunction in which case such a set of non-moral properties constitutes a non-moral property. Now, such a non-moral property contains all non-moral properties that are relevant to an object having a moral property. This means that it does not have any context that determines its relevance; the relevance of such a non-moral property is in other words not context-dependent. However, I think the account above is compatible with the basic idea in the view under consideration. The non-moral properties that defenders of this view appeal to in their examples, and whose relevance is claimed to be context-dependent, are simple properties, e.g. causing pleasure, being a promise and being considerate. But the non-moral properties that would make objects have a moral property if the relevance of simple non-moral properties as these is context-dependent would be complex, perhaps very complex, non-moral properties.

seems indeed incompatible with SMS. The reason is that SMS contains (**B**)—a necessary implication from a set of non-moral properties to a moral property—and such an implication seems to constitute a kind of moral principle.

However, it is important to observe that SMS is compatible with other sceptical views about moral principles that philosophers known as particularists may embrace. The kind of moral principle involved in SMS is primarily metaphysical since it concerns the way in which non-moral properties make objects have moral properties. Accordingly, it is reasonable to assume that it is compatible with the view that no other kinds of moral principles are true. For instance, as we will see in the next section, it is possible to argue that it is compatible with the view that there are no true moral principles that concern what reasons there are to perform actions.[24]

Dancy and other particularists sometimes suggest that the relevance of non-moral properties being context-dependent means that there are no true moral principles.[25] This view seems unfounded. As we saw in the previous section, the view that the relevance of non-moral properties is context-dependent is compatible with SMS. However, SMS involves (**B**), which, as just mentioned, seems to constitute a kind of moral principle. Hence, the relevance of non-moral properties being context-dependent is compatible with the truth of at least one kind of moral principle.

However, it might be argued that scepticism against this kind of moral principle is supported by a certain distinction. Dancy distinguishes between features that constitute reasons to perform actions, what he calls 'favourers', and features that do not constitute reasons themselves but merely enable other features to constitute such reasons, what he calls 'enablers'. As I mentioned earlier, Dancy assumes that what holds for the features that constitute reasons holds for the non-moral properties that make objects have moral properties as well. Thus, he distinguishes between non-moral properties that *make* objects have moral properties and non-moral properties that merely enable other non-moral properties to make objects have moral

[24] It is also plausible to assume that it is compatible with the view that there are no true moral principles of other kinds. For an overview of various conceptions of moral principles, see McKeever and Ridge (2006: 5–14). Dancy has characterized particularism in different ways over the years. In his recent book, he defines holism in the way mentioned above and particularism as the view that 'the possibility of moral thought and judgement does not depend on the provision of a suitable supply of moral principles' (Dancy 2004: 7). Dancy believes that holism implies particularism thus understood. If particularism is conceived in this way it does not entail that there are no true moral principles. However, this is not essential in the present context since he, as we will see, appeals to an important distinction which might be thought to have this result.

[25] See e.g. Dancy (1993: 66). See also e.g. Little (2000: 284), and Dancy (2004: 78–85).

properties, also called *enablers*.[26] It is particularly the latter distinction which concerns us here.[27]

Dancy argues for the distinction between what makes objects have moral properties and enablers partly by means of examples. I think this way of arguing is dubious. It is a matter of substantial normative argument whether the non-moral properties he appeals to are relevant at all to objects having moral properties, either as non-moral properties that make objects have a moral property or as enablers. Especially, I think many moral philosophers would object that the non-moral properties that are claimed to make objects have moral properties are too specific to have this function.[28] It is therefore controversial whether these examples support the distinction.

Nevertheless, Dancy's overall argument why a non-moral property that he classifies as an enabler does not make, say, an action right seems to be that it is not 'something for which we judge the action to be right'.[29] His general argument for claiming that a certain non-moral property is an

[26] See e.g. Dancy (1993: 22–6, 55–8, 77, 81), Dancy (1999*a*: 148–50), Dancy (1999b: 26–9), and Dancy (2004: 38–41, 45–52, 89–91, 95–9, 125–7). For a classification of various types of relevance of non-moral properties, see Sinnott-Armstrong (1999: 5).

[27] This distinction can be combined with the view that the relevance of non-moral properties is context-dependent. As a consequence, what is thought to make an object have a moral property varies accordingly. Return to the example in the last section. In one case a certain non-moral property—such as causing pleasure—may be considered to make an action right whereas another non-moral property—such as not being a punishment—may be considered to enable it to have that function, but in a different case these non-moral properties may be thought to have other functions. It should be observed, however, that the distinction between non-moral properties that make objects have moral properties and enablers is quite independent of the view that the relevance of non-moral properties is context-dependent. First, the distinction might hold even if the relevance of non-moral properties is not context-dependent. To see this, suppose that the relevance of, say, causing pleasure is not context-dependent but always contributes to actions being right. It is still possible to argue that for this non-moral property to make an action right, the action has to have other non-moral properties—such as not being a punishment—which enable it to have this function. Second, the relevance of non-moral properties might be context-dependent even if the distinction does not hold. To see this, recall the last section. I argued that a set of non-moral properties might consist of a number of non-moral properties and that within such a set the relevance of a non-moral property can be context-dependent, where the context is made up by other non-moral properties in the set. In Section 7, I will argue that, according to one use of 'make', what makes objects have a moral property might be such a set of non-moral properties and that there is no need for enablers. However, I will also argue that there is a pragmatic use of 'make' according to which a certain non-moral property might be selected as what makes an object have a moral property whereas another non-moral property is considered as an enabler.

[28] Cf. Shafer-Landau (1997: 590), Jensen and Lippert-Rasmussen (2005: 134–5), and Crisp (2007: 43–5).

[29] Dancy (1993: 81). See also e.g. Dancy (1981: 377) and Dancy (2004: 38–41).

enabler rather than something that makes objects have a moral property seems in other words to be that it does not constitute a reason to believe that an object has the moral property in question. Dancy admits, however, that he does not know how to draw the distinction and that it is easy to find examples of cases where it is not clear how to categorize a certain non-moral property.[30] He has, for instance, abandoned the idea that enablers always are negative non-moral properties.

Thus, Dancy argues for the view that a non-moral property which he classifies as an enabler cannot make objects have a moral property. However, he also embraces a further view: that a set of non-moral properties which *includes* such a non-moral property as a part cannot have this function either.[31] As far as I can see, Dancy does not always distinguish clearly between these views, and one gets the impression that he takes the first to support the second. However, it is the latter view which has consequences for SMS.

First, it means that SMS may refer to a kind of set of non-moral properties which cannot make objects have a moral property. A set of non-moral properties of the kind referred to in SMS may include a non-moral property that Dancy classifies as an enabler. According to Dancy's view, it follows that such a set of non-moral properties does not make an object have a moral property. Second, it means that a moral principle of the kind stated in (**B**) may turn out to be false. Such a moral principle concerns the way in which non-moral properties make objects have moral properties.[32] According to Dancy's view, a set of non-moral properties that includes a non-moral property which he classifies as an enabler cannot make objects have a moral property. It follows that if a version of (**B**) cites such a set of non-moral properties, it does not constitute a true moral principle.

In the next two sections, I will argue that these consequences can be avoided. Since the second consequence can be seen as an implication of the first, I will focus on the first.

6. COMPLEXES OF NON-MORAL PROPERTIES

Let us assume for the sake of the argument that Dancy is correct in his first view mentioned above: that a non-moral property which he classifies as an

[30] Dancy (2004: 51). See also Dancy (1999*a*: 148). Dancy's view of enablers is criticized by Lippert-Rasmussen (1999: 99–104), Sinnott-Armstrong (1999: 2–8), Raz (2000: 68–9), and Raz (2006: 103–7).

[31] See e.g. Dancy (1993: 76–7) Dancy (1999*b*: 26), and Dancy (2004: 38–40, 87, 89–91, 95–9, 125–7).

[32] Cf. Dancy (1993: 76–7), Dancy (1999*b*: 25), and Dancy (2004: 87).

enabler cannot make objects have a moral property. We may still refute his further view: that a set of non-moral properties which *includes* such a non-moral property cannot have this function either. It is evident that the first view does not entail the second. Even if a non-moral property which Dancy classifies as an enabler *by itself* cannot make objects have a moral property, it might be *part* of a set of non-moral properties which has this function.

There seems to be a rather simple way in which this is possible. Dancy distinguishes, as we have seen, between non-moral properties that make objects have moral properties and enablers. Now, there are complexes of non-moral properties which consist of combinations of non-moral properties of the first kind with non-moral properties of the second kind. To illustrate, consider one of Dancy's favourite examples of the distinction.[33] Suppose a person has promised to perform an action and that she ought to do it. In Dancy's view, the action being such that she has promised to perform it makes it have that moral property. According to the principle that ought implies can, unless the person is able to perform the action, it is not the case that she ought to perform it. In Dancy's view, the action being such that the person is able perform it is an example of an enabler. The combination of these two non-moral properties makes up a complex of non-moral properties. There might be further non-moral properties of either kind which can be added so as to obtain a more comprehensive complex of that kind.

It seems reasonable to argue that there are complexes of non-moral properties—that is, complexes which include both non-moral properties that Dancy believes make objects have moral properties and non-moral properties he classifies as enablers—that make objects have moral properties. There are at least three reasons for this view. First, a non-moral property which, in Dancy's view, makes an object have a moral property is combined with a non-moral property which, to exactly the same extent, is responsible for the object having the moral property. It then seems reasonable to maintain that, if we accept that the first non-moral property makes an object have a moral property, we should accept that such a complex of non-moral properties can have this function as well. In fact, we would seem to have reached a fuller account of what makes the object have the moral property. Second, if a single non-moral property is a reason to believe that an object has a moral property, such a complex of non-moral properties can clearly constitute such a reason as well. In that case it would seem to qualify as something that makes objects have a moral property, rather than being an enabler, according to Dancy's own argument. Third, it finds support

[33] Dancy (1993*a*: 148–9), and Dancy (2004: 39–40, 126–7).

in our way of thinking about what makes objects have moral properties. Suppose we ask why it is the case that a person ought to perform a certain action, say help another human being. One reasonable answer might be 'She promised to help her'. But another reasonable answer might be 'She promised to help and she's able to do so', or an answer which cites a more comprehensive complex of non-moral properties. These answers seem to differ mainly as regards how specific and complicated considerations they provide in support of the action having the moral property in question.

A complex of non-moral properties which, according to the reasoning above, makes objects have a moral property exemplifies a set of non-moral properties of the sort I had in mind earlier: one that includes what Dancy classifies as an enabler. The kind of set of non-moral properties referred to in SMS might constitute such a complex of non-moral properties. Hence, there is reason to believe that it can make objects have a moral property even if it contains a non-moral property which Dancy takes to be an enabler.

However, Dancy refutes what he calls 'the agglomerative principle' according to which a feature that constitutes a reason to perform an action in combination with an enabler makes up a more complex reason of that kind.[34] As already mentioned, Dancy believes that what holds for the features that constitute reasons also holds for the non-moral properties that make objects have moral properties. Consequently, he denies the agglomerative principle as regards the latter notion as well. On this view, a complex of non-moral properties which includes what he classifies as an enabler cannot make objects have a moral property.

Dancy's main argument against the agglomerative principle takes its point of departure in an example. Suppose a person has promised to perform an action and that she ought to do it. Suppose further that if the person had given her promise under duress, the action had not had this moral property. Formulated in terms of non-moral properties, Dancy seems to understand this case in the following way. The action being such that the person has promised to perform it constitutes a reason for her to do it. However, the action not being such that the promise to perform it was given under duress merely constitutes an enabler. Dancy then argues that it is mistaken to think that these two non-moral properties make up a complex—the action being such that the person *freely promised* to perform it—which constitutes a reason for her to do it.[35] He writes:

> [T]hose who recognize that their promise was deceitfully extracted from them often feel some compunction in not doing what they promised, even though they

[34] Dancy (2004: 39–41).

[35] Dancy (2004: 39). In discussing Dancy's example, I ignore certain problems which are due to how he describes the facts and properties in question.

> themselves recognize that in such circumstances their promise does not play its normal reason-giving role. I think their attitude would be different if what plays the reason-giving role were not that one promised but that one 'freely' promised (where to be free a promise must not be extracted by deceit). For on that hypothesis there would be no sign of a favourer in the case at all.[36]

Dancy's argument is, as I understand it, that the agglomerative principle cannot account for the following type of cases. Suppose a person has promised to perform an action but that the promise was given under duress. Then it is not the case that she ought to do it. The person may nonetheless feel compunction for not performing the action. The explanation seems to be that she believes that her having promised to perform the action still gives her a reason to do it. Especially, she believes this in spite of being aware that her promise 'does not play its normal reason-giving role' because it was given under duress and that she therefore may have a stronger reason not to do the action in question.[37]

As far as I understand, it would be difficult to argue that cases of the kind Dancy describes cannot occur. It is reasonable to maintain, however, that he is not in a position to appeal to such cases when arguing against the agglomerative principle since his own view on reasons seems unable to account for them.[38] According to Dancy's view, that a person has promised to perform an action is a reason for her to do it only if a certain enabler is in place, namely that it is not the case that her promise was given under duress. This view entails that if the promise *was* given under duress, her having promised to perform the action is no reason for her to do it. But the situation Dancy describes seems precisely to be one in which a person believes that she has a reason to do what she has promised to do *even if* her promise was given under duress.[39]

Nevertheless, it might still be asked how Dancy's argument in relation to the agglomerative principle should be responded to. Below I suggest

[36] Dancy (2004: 39–40).

[37] According to another interpretation, the explanation as to why the person feels compunction for not doing what she has promised to do is not that she believes that she has a reason to perform the action in question. Rather, the explanation is that a promise *normally* plays a reason-giving role, although she is aware that it does not do so in the case in hand. But then it is difficult to see how such cases can constitute an argument against the agglomerative principle.

[38] Cf. Raz (2006: 105).

[39] Of course, it might be argued that in such cases a person is mistaken in believing that she has a reason to perform the action in question. However, if she does not have such a reason, the kind of cases Dancy considers cannot constitute an argument against the agglomerative principle. In what follows I therefore leave open the possibility that she actually has a reason to perform the action. However, the two accounts I provide below are compatible with this not being so.

two different responses. However, I do not discuss which of them is preferable since that would take us too far.

(1) When Dancy discusses the agglomerative principle, he seems to presume that, on this principle, a single non-moral property cannot constitute a reason, but that only a complex of non-moral properties can do so. Thus, the single non-moral property of an action being such that a person has promised to perform it does not constitute a reason for her to do it, but this function is only had by a complex of non-moral properties, for example the action being such that she freely promised to perform it. This explains why Dancy thinks that the agglomerative principle cannot account for cases where a person believes that her having promised to perform an action gives her a reason to do it even if she is aware that she gave the promise under duress and that she therefore may have a stronger reason not to do it.

However, advocates of the agglomerative principle need not accept this presumption. They may maintain that a complex of non-moral properties (e.g. an action being such that a person freely promised to perform it) constitutes a reason for her to perform the action. Yet, they may maintain that a non-moral property which is part of the complex (e.g. an action being such that a person has promised to perform it) also constitutes such a reason. This view consequently leaves open the possibility that reasons vary in specificity and strength.

On this view, it is possible to account for the kinds of case Dancy appeals to in this argument. We might maintain that an action being such that a person has promised to perform it constitutes a reason for her to do it although she has a stronger reason not to perform the action since she gave her promise under duress and hence not freely.

If this reasoning is plausible, advocates of the agglomerative principle may uphold it as regards the features that constitute reasons. In that case there are, as far as I can see, no grounds for believing that it does not hold for the non-moral properties that make objects have moral properties as well. Thus, a complex of non-moral properties which involves what Dancy classifies as an enabler can make objects have a moral property.

(2) As we have seen, Dancy assumes that what holds for the features that constitute reasons to perform actions also holds for the non-moral properties that make objects have moral properties. Accordingly he denies the agglomerative principle as regards both notions. As a consequence, he believes that a complex of non-moral properties which contains what he classifies as an enabler *neither* can constitute a reason to perform an action *nor* make actions have a moral property. However, Dancy's assumption is open to doubt. We may accordingly deny the agglomerative principle

as regards the features that constitute reasons but accept it as regards the non-moral properties that make objects have moral properties. In other words, the following view suggests itself: a complex of non-moral properties which includes what Dancy classifies as an enabler does not constitute a reason for action but it may still make objects have a moral property.

It might first be noted that the view that there is such a difference between the features that constitute reasons and the non-moral properties that make objects have moral properties should not strike us as particularly surprising since they concern different matters: the first notion concerns what constitute reasons whereas the second concerns a primarily metaphysical make-relation. Moreover, this view does not need to be particularly radical since it is compatible with the notion that what constitutes the relevant kinds of reason and what makes objects have moral properties come apart only to a limited extent. On this view, there are certain parts of what makes objects have a moral property that are not part of what constitutes the reasons in question, namely what Dancy classifies as enablers. However, each non-moral property which constitutes such a reason may be part of what makes objects have a certain moral property.

One ground for accepting this view is that we seem to have different conceptions of what function non-moral properties have with respect to reasons for action and with respect to the make-relation. To see this, we might compare a view about reasons with the corresponding view about rightness. Consider first a simple internalist view of reasons according to which a person has a reason to perform an action in so far as she desires to perform it. We would presumably not say that, on this view, the person having a desire to perform a certain action needs to be part of her *reason* to perform that action.[40] Rather, her having such a desire is a standing precondition for something to constitute a reason to perform an action. In Dancy's terminology it can be characterized as a permanent enabler. Consider next subjectivism about rightness according to which an action is right for a person to perform in so far as she desires to perform it. Here we seem inclined to say that, on this view, the action being such that the person has a desire to perform it is something that *makes* it right. We would presumably have the same result if we considered other views of reasons and moral properties. Hence, there are grounds to believe that a non-moral property which is classified as an enabler is not part of a reason to perform actions but is part of what makes objects have a moral property. Another justification for this view is the notion that reasons, at least in favourable circumstances, figure in a person's practical deliberation about what to do,

[40] Cf. Persson (2005: 114), McKeever and Ridge (2006: 34), and Väyrynen (2006: 715).

whereas what makes actions have moral properties need not do so. It can then be argued that considerations such as that a person has not given a promise under duress or that she is able to perform an action are not, at least not normally, part of her practical deliberation and hence not part of her reasons.[41] However, they may still be part of what makes actions have moral properties.

On this view, it is possible to account for the kinds of case Dancy appeals to in his argument. We might maintain that what constitutes a person's reason to perform an action is that it is such that she has promised to do it, not a complex of non-moral properties which includes a non-moral property which Dancy classifies as an enabler, such as the action being such that she freely promised to do it. Hence, a non-moral property which Dancy classifies as an enabler (e.g. an action not being such that the promise to perform it was given under duress) is not part of a reason to perform an action. However, this view leaves open the possibility that although such a non-moral property is not part of a reason, it is relevant to the nature and strength of the reasons a person has to perform an action. In that case we can maintain that an action being such that a person has promised to perform it constitutes a reason for her to do it even if her promise was given under duress and that she therefore has a stronger reason not to perform the action.

This view might have consequences for what kinds of moral principle are true. If a non-moral property which Dancy classifies as an enabler cannot be part of a complex of non-moral properties which constitutes a reason for action, it might be the case that there are no true moral principles that state such reasons. However, if what he classifies as an enabler can be part of a complex of non-moral properties which makes objects have a moral property, there is ground for believing that there are moral principles which state what makes objects have such properties. Hence, it might be the case that there are moral principles concerning the make-relation but not concerning reasons.[42]

[41] Cf. Cullity (2002: 179–80).

[42] I have maintained that a complex of non-moral properties, such as the kind of set of non-moral properties referred to in SMS, can make objects have a moral property. Such a set of non-moral properties may be quite complicated. As a result, (**B**) in SMS states a moral principle that might be quite complicated as well. Dancy and other particularists argue against supervenience on this ground. See e.g. Dancy (1999*b*: 25–6), Dancy (2004: 87–8), and Little (2000: 285–6). I think this argument betrays a failure to distinguish between different kinds of moral principles. This is perhaps a plausible argument against some kinds of moral principle, e.g. principles that are meant to provide moral guidance. However, it is not effective against the kind of moral principles that concern us here. As long as such a principle describes what makes objects have a moral property, it is not a problem that it is complicated.

7. TWO USES OF 'MAKE'

In the previous section, I argued that a set of non-moral properties which involves what Dancy classifies as an enabler can make objects have a moral property. As I pointed out, there is therefore reason to believe that a set of non-moral properties of the kind referred to in SMS can have this function even if it includes such a non-moral property. In the present section, I will argue that this view is confirmed by considerations of two pertinent uses of 'make'.

Let us start by considering the kind of set of non-moral properties referred to in SMS. Such a set consists of those, but only those, non-moral properties that are sufficient for an object to have a certain moral property. Next, consider a part of such a set of non-moral properties. Each such non-moral property, or combination of non-moral properties, is a necessary part of a sufficient condition for an object to have a moral property. It is in other words a necessary condition for a particular set of non-moral properties to be a sufficient condition of the indicated kind.[43]

Now, I think it can be argued that when we claim that non-moral properties make objects have moral properties, we may have in mind either a part of such a set of non-moral properties or the set in its entirety.

The pragmatic use of 'make' According to the first use of 'make', what makes objects have a moral property might be a *part* of a set of non-moral properties of the kind referred to in SMS; it need not be an entire set of that kind.

If we understand 'make' in this way, we are in the position to account for Dancy's distinction between non-moral properties that make objects have a certain moral property and enablers. We may consider a certain part of a set of non-moral properties of the kind referred to in SMS as what makes an object have a moral property. And we may consider a certain other part of such a set as an enabler. To illustrate, take another of Dancy's examples. Suppose we believe that a person is good. We might want to claim that what makes her good is, say, her being considerate. Furthermore, we might believe that, if she had been cruel, she would not have been good. In Dancy's view, her not being cruel is an enabler.[44] In the above mentioned

[43] This has a counterpart in J. L. Mackie's notion of INUS-conditions; see Mackie (1974: 61–7).

[44] Dancy (1981: 377). For the sake of the argument, I accept Dancy's presumption that not being cruel is a non-moral property.

kind of set of non-moral properties, being considerate can then be regarded as something that makes the person good whereas not being cruel can be thought to constitute an enabler.

It might be asked why we consider a certain non-moral property as something that makes an object have a moral property whereas another non-moral property is considered as an enabler. Suppose we believe that a certain non-moral property is a necessary part of a sufficient condition for an object to have a moral property in the way just mentioned. Put abstractly, I think it is plausible to maintain that we regard such a non-moral property as one that makes the object have the moral property because we for some reason find it significant in consideration of the object having that moral property. Similarly, we consider another non-moral property of that kind as an enabler because we do not find it thus significant. Whether such a non-moral property is found to be significant or not depends, I think, on pragmatic factors.

According to the account of the pragmatic use of 'make' that I favour, a non-moral property which is considered significant in view of an object having a moral property is part of an explanation of a certain kind. Briefly put, the idea is this. When we claim that a certain non-moral property makes an object have a moral property, we put forward an explanation of what makes the object have that moral property. According to an established view of explanations, we select certain pieces of information as explanations because we find them significant as a consequence of being directed by various pragmatic considerations. Thus, in explaining what makes an object have a moral property, we refer to a non-moral property of the kind indicated above which we take to be significant as to why the object has the moral property in question. Moreover, whether we consider such a non-moral property as significant or not depends on broadly pragmatic considerations.[45] What these considerations consist of may presumably vary. However, put very generally, whether a non-moral property is considered as significant or not depends on the context at issue, in particular on the context represented by the beliefs of the people for whom the explanation is intended. It depends, for instance, on their beliefs about the relation between moral properties and non-moral properties and their beliefs about the circumstances in which they find themselves.[46] (Needless to say, often

[45] It might be argued that there is a parallel in the reason why a certain causal factor is selected as what causally explains an event. According to an influential view, we select a causal factor as causally explanatory because we consider it as salient in a certain context as a consequence of being directed by various pragmatic considerations. See e.g. van Fraassen (1980: ch. 5), Woodward (1984), Lewis (1986), and Lipton (1990).

[46] It should be noted that this kind of context differs from the one discussed in Section 4.

we are not aware that we select a non-moral property as what makes an object have a moral property on these grounds.)[47]

A very general reason why a certain non-moral property is regarded as significant in consideration of an object having a moral property is probably that, in a given context, it is presumed that it has certain other non-moral properties. As a consequence, a non-moral property that is thought to make an object have a moral property is considered as more significant as compared with these 'presumed' non-moral properties, some of which are classified as enablers. A non-moral property that is thought to make an object have a moral property stands thus out as remarkable against a background of non-moral properties which are taken for granted to belong to the object. To illustrate, suppose again that a person being considerate is considered as something that makes her good whereas her not being cruel is considered as an enabler. According to the present suggestion, the reason might be that these non-moral properties differ in significance since, in the context at issue, it is presumed that the person is not cruel, whereas it is not in a similar manner presumed that she is considerate.

This ground for distinguishing between non-moral properties that make objects have moral properties and enablers is presumably reinforced by the fact that in ordinary communication we are governed by pragmatic concerns (understood in a more narrow sense than above). Generally we only utter sentences that we believe are relevant to the people we are communicating with by providing them with information that we believe that they are not already familiar with.[48] As a consequence, we do not normally provide information to the effect that objects have certain features which we believe that people already presume objects to have. It is reasonable to assume that this affects what non-moral properties we select as those that make objects have moral properties. In particular, it prevents us from claiming that what makes an object have a moral property is a non-moral property which we believe that people presume belongs to the object.

One reason why people presume that an object has a certain non-moral property might in turn be that they believe that objects have to have it in order to have the moral property in question. That is, it might be thought that it holds for any object that if it does not have the non-moral property, it cannot have the moral property. One example might be the relation

[47] It might be asked whether this pragmatic account of the distinction between what makes objects have moral properties and enablers also applies to the distinction between what constitutes reasons for action—what Dancy calls favourers—and the corresponding enablers. Although I am not committed to this view, I think the pragmatic account is generalizable in this way. For related suggestions, see Raz (2000: 59), Broome (2004: 32–5), and McKeever and Ridge (2006: 72–5).

[48] See Grice (1989 (1975): 26–7).

between the non-moral property of not being cruel and a person being good. The same applies perhaps to the example above which rests on the principle that ought implies can. Another reason why people presume that an object has a certain non-moral property might simply be that they believe that objects belonging to the relevant kind normally have that non-moral property. As people's views about these matters presumably may vary, there are grounds to believe that what is considered to make objects have a moral property may vary accordingly.

Moreover, whether a non-moral property is considered significant might according to this account also depend on what people believe to be the case in the circumstances in which they find themselves. As a consequence, what is thought to make an object have a moral property may vary depending on what they believe about the circumstances in question. To illustrate, we might recall one of the examples I discussed in the last section. In Dancy's view, an action being such that a person has promised to perform it makes it such that the person ought to do it, whereas the action not being such that the promise was given under duress merely constitutes an enabler. We also saw that in Dancy's view the combination of these two non-moral properties does not make the action have the moral property in question. However, it is plausible to argue that whether this description is plausible depends on what is thought about the circumstances in question. Consider first a situation where an action has the two non-moral properties just mentioned. In this situation, we may suppose, people are not generally forced to promise to perform actions. As a consequence, it is generally presumed that promises are not given under duress. This is probably the circumstances we find ourselves in. Admittedly, in such a situation it seems plausible to describe the case in the way Dancy does. However, consider now another situation. In this situation an action has exactly the same non-moral properties that are had by the action in the first situation, including the two the non-moral properties just mentioned. However, here people quite generally *are* forced to promise to perform actions under duress. In such circumstances it cannot generally be presumed that promises are not given under duress; on the contrary, it can quite generally be expected that certain promises *are* given under duress. Now, that an action has the non-moral property Dancy classifies as an enabler seems in such a situation to stand out as remarkable against a background of other properties which the action is thought to have. It seems therefore plausible to assume that when people in this situation come to believe that the person in question ought to perform the action, they consider this non-moral property as significant. As a result, rather than taking it to be an enabler, they may take it as something that makes the action such that it ought to be done. In particular, it seems reasonable for them to hold that the above mentioned complex of non-moral properties—the action being such that the person freely promised to perform it—makes the action have the moral property.[49] On either alternative, what Dancy classifies as an enabler would contribute to what is thought to make an action have a moral property.[50]

[49] According to the pragmatic use of 'make', whether a certain non-moral property is considered to make an object have a moral property depends on non-metaphysical, pragmatic, factors. One consequence is, as we have seen, that negative non-moral properties, or at least complexes of non-moral properties of which such non-moral properties are part, can be claimed to make objects have moral properties in certain situations. As we will see below, according to the strict use of 'make' negative non-moral properties might also be part of what makes objects have moral properties. However, it might be maintained that only positive non-moral properties can make objects have moral properties whereas negative non-moral properties cannot have this function. In his early writings on particularism, Dancy seems attracted to this position; see e.g. Dancy (1993: 81). Recently Nick Zangwill has suggested a similar view; see Zangwill (2003) and his contribution to this volume. This is an interesting idea which I cannot do justice to here. However, I would like to make the following brief comments. (**i**) It is not entirely clear how this view should be understood. On a weak understanding it says that a negative non-moral property *by itself* cannot make objects have a moral property. On a strong understanding it says that a complex of non-moral properties of which such a non-moral property is *part* cannot have this function either. My view is compatible with the first view but not with the second. (**ii**) As I have just argued, in certain situations we seem prepared to claim that negative non-moral properties—or at least complexes of non-moral properties of which they are part—make objects have moral properties. (**iii**) Correspondingly, there are positive non-moral properties which contribute to objects having moral properties but which we in most situations would not claim make objects have moral properties. For example, an action being such that a person is able to perform it seems to be a precondition for the action to be such that she ought to perform it, but in most situations we would not claim that this non-moral property makes actions have that moral property. Hence, the distinction between non-moral properties that make objects have moral properties and those that do not have this function does not coincide with the distinction between the pertinent positive and negative non-moral properties. (**iv**) It might further be argued that, on this view, there are cases where there is no difference in terms of what makes objects have moral properties which can explain why two objects have different moral properties. The following illustrates what I have in mind. Suppose an action is such that it causes pain and that we believe that it is wrong. Suppose further that the action has certain negative non-moral properties which are preconditions for it being wrong, e.g. not being such that it causes pain as a consequence of medical treatment. Now, consider another action which is also such that it causes pain. However, suppose that we believe that this action is *not* wrong. The reason, we may suppose, is that it differs from the first action in the following way: it has the negative non-moral property of not being such that it causes pain to someone who cares about being in pain. According to one version of the view under consideration, negative non-moral properties are not part of what makes objects have moral properties. It follows that there is no difference between these two actions in terms of what makes actions have moral properties which explains why they differ morally.

[50] It is worth pointing out that in these two situations the actions in question have the same non-moral properties. Hence, that a non-moral property is considered as an enabler in the first situation whereas it is considered as part of something that makes an action have a moral property in the second situation cannot be explained in terms of the

As we have seen, the present account of the pragmatic use of ‘make’ is confirmed by the observation that what is thought to make an object have a moral property might vary. It is further confirmed by the fact that it is able to explain two phenomena I have alluded to earlier.

First, it is able to explain why many non-moral properties that Dancy classifies as enablers are negative properties (e.g. an action not being such that the promise to perform it was given under duress). According to the present suggestion, one reason why a non-moral property is considered as an enabler is that it is presumed that the object in question has it. Now, it is quite natural that we typically presume that an object has a negative property, that is, a property an object has in virtue of *not* having a certain property. This is so since we generally presuppose that objects lack properties in case we do not have any information to the effect that they *have* these properties.

Furthermore, it is able to explain Dancy’s view that there does not seem to be any clear distinction between non-moral properties that make objects have a moral property and enablers, and that it is easy to find examples of cases where it is not clear how to categorize a certain non-moral property in these terms. According to the present account, which category a non-moral property belongs to depends on the context of the object in question, where the context primarily is represented by people’s beliefs. Since an object’s context of this kind may vary, whether a non-moral property is thought to make the object have a moral property or merely is considered as an enabler may vary accordingly. Moreover, as it might be unclear what makes up the pertinent context of an object, it might be unclear how to categorize a given non-moral property.[51]

In the last section, I provided some general arguments to the effect that a set of non-moral properties of the kind referred to in SMS can make objects

actions having different non-moral properties. More precisely, the explanation cannot be that in these situations the non-moral property in question has different contexts, where the contexts are made up by other non-moral properties of the actions. For a different view, see Zangwill’s contribution to this volume.

[51] In Section 3, I proposed briefly the idea that the asymmetric relation between moral and non-moral properties can be explicated in terms of the latter being explanatory prior to the former. The pragmatic account suggested here provides yet an aspect in which this might be the case. Consider again the following statements: (1) ‘What makes actions right is that they maximize happiness’ and (2) ‘What makes actions maximize happiness is that they are right’. (1) seems more plausible than (2). One reason might be this. On the pragmatic account we understand statements to the effect that non-moral properties make objects have moral properties as explanations. Moreover, explanations figure in certain contexts represented by people’s beliefs. It is reasonable to assume that explanations of what makes objects have moral properties quite generally figure in contexts where it is presumed that moral properties are more problematic and hence in need of clarification than non-moral properties. As a result (1) seems more plausible than (2). Cf. e.g. van Fraassen (1980: 130–4) and Richardson (1995).

have a moral property even if it includes what Dancy classifies as an enabler. What I have said above about the pragmatic use of 'make' helps to support this view. As I have already indicated, I think it is uncontroversial to assume that, when we claim that non-moral properties make objects have moral properties, we have primarily in mind a metaphysical relation. Admittedly, we pick out certain non-moral properties as what makes objects have moral properties and consider others as enablers. However, we do so because, in explaining what makes objects have moral properties, we select certain non-moral properties as significant as a consequence of being directed by various pragmatic considerations. But from a metaphysical point of view, there is no relevant difference between these non-moral properties; each of them is a necessary part of a sufficient condition for an object to have a moral property. Hence it is in accordance with the pragmatic use of 'make' to maintain that such a set of non-moral properties can make objects have a moral property in spite of it including what Dancy takes to be an enabler.[52]

The strict use of 'make' According to this use of 'make', what makes objects have a moral property is a set of non-moral properties of the kind referred to in SMS in its *entirety*, not a part of it.[53] It follows that such a set of non-moral properties makes objects have a moral property even if it contains a non-moral property which Dancy classifies as an enabler.

We saw earlier that there are reasons to believe that it is correct to use 'make' in this way. The following considerations confirm this view.

First, it seems to find support in common parlance. Suppose someone claims that a person is good because she is considerate. Against this someone might object: 'But that doesn't *make* her a good person', stressing 'make'. To support this view, the objector might argue that if the person in question is cruel, she is not good. As far as I understand it, this is perfectly consistent with a correct use of 'make'. One reasonable interpretation of this objection

[52] It is worth observing that the pragmatic use of 'make' is compatible with the possibility that a complex of non-moral properties which includes what is considered as an enabler makes objects have a moral property even if it is not as comprehensive as a set of non-moral properties of the kind referred to in SMS. This can be accounted for in the following way. Suppose we select a non-moral property as something that makes an object have a moral property whereas another non-moral property is considered as an enabler because the first is thought significant whereas the second is not. Consider now a complex that consists of these two non-moral properties. Since the first non-moral property is considered significant, the whole complex is presumably considered so as well. We might perhaps add further non-moral properties of either kind with the same result. According to the pragmatic use of 'make', such a set of non-moral properties can then be claimed to make objects have a moral property.

[53] Cf. Crisp (2000: 32–40) and Raz (2000: 49–58, 61–70).

is that, for a set of non-moral properties to make a person good, it has to provide a sufficient condition for the person to be good. According to the objector, being considerate is not thus sufficient; it also has to involve the non-moral property of not being cruel and perhaps other non-moral properties as well.[54]

Second, it finds support in the fact that the use of 'make' at issue primarily refers to a metaphysical relation. Earlier I pointed out that from a metaphysical point of view there is no relevant difference between the various parts of a set of non-moral properties of the kind referred to in SMS. More precisely, each such non-moral property is a necessary part of a sufficient condition for an object to have a moral property. Thus what is metaphysically responsible for an object having a moral property is not a single part of such a set of non-moral properties but rather the set in its entirety. As I mentioned above, it seems uncontroversial to assume that when we claim that non-moral properties make objects have moral properties, we have primarily in mind a metaphysical relation. It is then reasonable to claim that a set of non-moral properties of the kind referred to in SMS in its entirety can make objects have a moral property. At least, it is difficult to deny that it is appropriate to use 'make' in this way, even if there also are other legitimate uses of the term.

Relatedly, the strict use of 'make' seems more fundamental than the pragmatic. As we have seen, we adopt a pragmatic use of 'make' because we feel a need to single out certain non-moral properties when we put forward explanations of a certain kind. However, since 'make' primarily refers to a metaphysical relation, the strict use of the term seems more fundamental.

Moreover, it might be argued that this view is supported by our notion of what is involved in explanations of what makes objects have moral properties. Above I maintained that when we put forward such explanations, we select certain non-moral properties that we find significant because we are directed by various pragmatic considerations. However, I do not think we would deny that a set of non-moral properties of the kind referred to in SMS in its entirety also can explain what makes an object have a moral property. Indeed, since such a set contains those, but only those, non-moral properties that together are sufficient for an object to have that moral property, it might be argued that it, in a certain sense, provides

[54] Moreover, according to the entry on 'make' in the *Oxford English Dictionary*, we sometimes use this term to refer to a relation that holds when something is a sufficient condition for something other. See 'Make', *OED* (1989: 234–7). Thus it seems difficult to deny that it is consistent with a correct use of 'make' to claim that a set of non-moral properties of the kind referred to in SMS makes an object have a moral property.

a complete explanation or '*the* explanation' of what makes the object have the moral property in question.[55]

8. CONCLUSION

In this paper, I have argued that particularism does not threaten the capability of supervenience to account for the notion that non-moral properties make objects have moral properties. There are basically two conceptions of particularism: holism and the contention that there are no moral principles. I argued that the view that the version of supervenience I developed, SMS, provides a basis for an account of the mentioned notion is compatible with the relevant version of holism, the view that the relevance of non-moral properties is context-dependent. However, SMS is incompatible with the view that there are no true moral principles. Particularists find support for this view in Dancy's distinction between non-moral properties that make objects have moral properties and enablers. According to Dancy's conception of this distinction, if SMS involves what he classifies as an enabler, it does not refer to a set of non-moral properties of a kind that makes objects have a moral property. As a result, the moral principle in question does not hold. I argued, however, that there is reason to believe that SMS refers to a set of non-moral properties of a kind which can make objects have a moral property even if it involves what Dancy takes to be an enabler. In particular, it does so according to two pertinent uses of 'make' of which the first is compatible with Dancy's distinction. Consequently, there might be at least one kind of true moral principle. It should be stressed again, however, that SMS is compatible with the view that there are no other true moral principles.

REFERENCES

Alm, David (2004) 'Atomism about Value' *Australasian Journal of Philosophy* 82: 312–31.

Blackburn, Simon (1993 (1985)), 'Supervenience Revisited', *Essays in Quasi-Realism*, (Oxford: Oxford University Press), 130–48.

Brink, David (1989) *Moral Realism and the Foundations of Ethics* (Cambridge: Cambridge University Press).

[55] Once more there might be a parallel in causal explanations. Although we select a certain causal factor as what causally explains an event, we would presumably not deny that a complex of factors which contains those, but only those, that together are sufficient for it to occur also explains the event. Indeed, it might be claimed that it constitutes a complete causal explanation or '*the* causal explanation' of the event in question. See e.g. Lewis (1986: 218).

Broome, John (2004) ‘Reason’ in *Reason and Value*, ed. R. Jay Wallace, Philip Pettit, Samuel Scheffler and Michael Smith, Oxford: Clarendon Press, 28–55.

Crisp, Roger (2000) ‘Particularizing Particularism’ in *Moral Particularism*, ed. Brad Hooker and Margaret Little (Oxford: Clarendon Press), 23–47.

____ (2007) ‘Ethics Without Reason?’ *Journal of Moral Philosophy* 4: 41–9.

Cullity, Garrett (2002) ‘Particularism and Presumptive Reasons’ *Proceedings of the Aristotelian Society* supp. vol. 76: 169–90.

Dancy, Jonathan (1981) ‘On Moral Properties’ *Mind* 90: 367–85.

____ (1993) *Moral Reasons* (Oxford: Blackwell).

____ (1999*a*), ‘On the Logical and Moral Adequacy of Particularism’ *Theoria* 65: 114–26.

____ (1999*b*), ‘Defending Particularism’ *Metaphilosophy* 30: 25–32.

____ (2004), *Ethics Without Principles* (Oxford: Clarendon Press).

Grice, Paul (1989 (1975)), ‘Logic and Conversation’, *Studies in the Ways of Words* (Cambridge, MA: Harvard University Press), 22–40.

Grimes, Thomas (1991), ‘Supervenience, Determination and Dependence’ *Philosophical Studies* 62: 79–92.

Jackson, Frank (1998) *From Metaphysics to Ethics* (Oxford: Clarendon Press).

Jensen, Karsten Klint, and Lippert-Rasmussen, Kasper (2005), ‘Understanding Particularism’ *Theoria* 71: 118–37.

Kim, Jaegwon (1993 (1984)), ‘Concepts of Supervenience’ in *Supervenience and Mind* (Cambridge: Cambridge University Press), 53–78.

____ (1993 (1990)), ‘Supervenience as a Philosophical Concept’ in *Supervenience and Mind* (Cambridge: Cambridge University Press), 131–60.

Kirchin, Simon (2003) ‘Particularism, Generalism and the Counting Argument’ *European Journal of Philosophy* 11: 54–71.

Lewis, David (1986) ‘Causal Explanation’ in *Philosophical Papers*, ii, (Oxford: Oxford University Press), 214–40.

Lippert-Rasmussen, Kasper (1999), ‘On Denying a Significant Version of the Constancy Assumption’ *Theoria* 65: 90–113.

Lipton, Peter (1990) ‘Contrastive Explanation’ in *Explanation and its Limits*, ed. D. Knowles (Cambridge: Cambridge University Press), 247–66.

Little, Margaret (2000) ‘Moral Generalities Revisited’ in *Moral Particularism*, ed. Brad Hooker and Margaret Little (Oxford: Clarendon Press), 276–304.

McKeever, Sean, and Ridge, Michael (2006) *Principled Ethics* (Oxford: Clarendon Press).

Mackie, J. L. (1974) *The Cement of the Universe* (Oxford: ClarendonPress).

____ (1977) *Ethics: Inventing Right and Wrong* (London: Penguin).

McLaughlin, Brian (1995) ‘Varieties of Supervenience’ in *Supervenience: New Essays*, ed. Elias E. Savellos and Ümit D. Yalçin (Cambridge, MA: Harvard University Press), 16–59.

McNaughton, David (1988) *Moral Vision* (Oxford: Blackwell).

Oxford English Dictionary (1989), ix (Oxford: Oxford University Press).

Persson, Ingmar (2005) *The Retreat of Reason* (Oxford: Oxford University Press).

Post, John (1999) ‘Is Supervenience Asymmetric?’ *Manuscrito* 12: 305–44.

Raz, Joseph (2000) ‘The Truth in Particularism’ in *Moral Particularism*, ed. Brad Hooker and Margaret Little (Oxford: Clarendon Press), 48–78.

____ (2006) ‘The Trouble with Particularism’ *Mind* 115: 99–120.

Richardson, Alan (1995), ‘Explanation: Pragmatics and Asymmetry’ *Philosophical Studies* 80: 109–29.

Robinson, Luke (2006) ‘Moral Holism, Moral Generalism, and Moral Dispositionalism’ *Mind* 115: 331–60.

Shafer-Landau, Russ (1997) ‘Moral Rules’ *Ethics* 107: 584–611.

____ (2003) *Moral Realism: A Defence* (Oxford: Clarendon Press).

Sinnott-Armstrong, Walter (1999) ‘Some Varieties of Particularism’ *Metaphilosophy* LLC: 1–12.

Strandberg, Caj (2004) *Moral Reality: A Defence of Moral Realism* (Lund: Lund University).

van Fraassen, Bas (1980) *The Scientific Image* (Oxford: Clarendon Press).

Väyrynen, Pekka (2006) ‘Moral Generalism: Enjoy in Moderation’ *Ethics* 116: 707–41.

Woodward, James (1984) ‘A Theory of Singular Causal Explanation’ *Erkenntnis* 21: 231–62.

Zangwill, Nick (2003) ‘Negative Properties, Determination and Conditionals’ *Topoi* 22: 127–34.

7

Robust Ethical Realism, Non-Naturalism, and Normativity

William J. FitzPatrick

Ethical realists have labored for the past thirty years now in the shadow of Mackie's "argument from queerness."[1] The result has been a familiar retreat by many realists from any metaphysical and epistemological commitments not already endorsed by the sciences. For at least some of us who are attracted to ethical realism, however, there is a sense that we have often gone too far in that direction, losing sight of concerns or intuitions that ultimately drew us to the idea of ethical reality in the first place. Ambitions to capture the categoricity of moral requirements, for example, or the irreducibly evaluative or normative nature of ethical facts and properties have often been sidelined or renounced altogether in the rush to find safe metaphysical ground.

Not everyone sees the dropping of such ambitions as a loss. Many contemporary realists are driven more in the end by aspirations to square ethics with metaphysical naturalism or to exploit developments in semantics to show how ethical terms could refer to real natural properties so that ethical properties might pull their weight within scientific explanations or predictions. It seems to me, however, that such approaches secure the 'reality' of ethical facts and properties only by turning them into something else and deflating them in the process.[2] My aim is therefore to explore and to motivate what is in any case a *more robust* yet still plausible ethical

I would like to thank Torin Alter, Tom Baldwin, Noell Birondo, David Brink, Sarah Buss, Tom Carson, Terence Cuneo, Jamie Dreier, David Enoch, Joshua Gert, Patricia Greenspan, Jim Klagge, Brad Majors, Tristram McPherson, Geoff Sayre-McCord, Mark Schroeder, other participants of the 2006 Metaethics Workshop, and two anonymous referees for helpful discussion of an earlier version of this paper.

[1] Mackie (1977: 38–42).

[2] Cf. Wiggins (1993*a*: 311): "When the naturalist reconstructs moral predicates, I suspect he loses hold of moral properties altogether."

realism, and to argue that this would require embracing non-naturalist commitments associated with normativity. The hope is to begin to carve out a viable and attractive middle position between the extreme version of realism portrayed in Mackie's caricature of it and the naturalistic versions typically offered as alternatives.[3]

Among contemporary metaethical theories, one of the closest to the view I want to explore is Shafer-Landau's ambitious and attractive take on realism.[4] While I am highly sympathetic to his project, however, I will argue that his non-naturalism does not go far enough. In particular, his view remains naturalistic with regard to the metaphysics of moral properties and facts, construing them as exhaustively constituted by natural properties and facts. In this respect, his view overlaps with Brink's non-reductive ethical *naturalism*.[5] I will try to show why a robust realist needs to posit a range of facts that cannot be so construed, particularly if he wishes (as Shafer-Landau does) to incorporate into his view both the intrinsic normativity of moral requirements and the stance-independence of moral standards. And I will argue that this leads to a non-naturalistic construal even of ordinary ethical facts and properties, though without threatening or obscuring their supervenience on natural facts and properties.

It is important to be clear at the outset about what I am and am not seeking to accomplish here. There are really two related projects. One, which will occupy the first half of the paper, is to illuminate what I take to be interesting and plausible motivations for what I am calling a more robust ethical realism. The point is not to devise a knockdown argument for robust realism, but simply to explore honestly what has driven me to it after having initially been attracted to some of the naturalistic realist views I now reject. The strategy will involve tracing the implications of certain intuitions and showing how such views seem unable to accommodate them, but I want to be clear that I do not thereby take myself to be establishing the truth of robust ethical realism through a process of elimination. For one thing, even if the intuitions are granted, much of what I say against naturalistic views could equally be accepted by expressivists. So those not otherwise attracted to realism might take many of the arguments to point not to robust realism but to expressivism (much as Moore's arguments were taken by many to support non-cognitivism rather than non-naturalism).

[3] Mackie's portrayal of realism—the view, as he puts it, that there are "objective values"—is a caricature insofar as he likens alleged objective values to transcendent Platonic forms with mysterious coercive powers, grasped by a special faculty. The more important and accurate part of Mackie's portrayal of realism, at least for the robust ethical realism I will explore, is the idea of objective prescriptivity, understood as the idea that moral judgments express or imply categorical reasons. (Cf. Smith 1993.)

[4] Shafer-Landau (2003).

[5] Brink (1989: chs. 6–7 and 2001: 157).

Since I shall not here take up a critique of expressivism, I cannot hope to be giving an unqualified argument for robust realism. Instead, my purpose is to clarify the intuitions and considerations that might lead some of us who are independently attracted to realism to adopt a more robust form of it. This will at least show why this debate matters, and what is at stake.

The second project is to establish a strong conditional claim of significant metaethical interest: if we do embrace realism, and accept the intuitions that lead to a robust form of it, then we are in fact committed to a non-naturalistic metaphysics of ethical facts and properties. Again, this would not show that non-naturalism is correct. But it would show that there are real pressures in that direction for those who accept realism together with certain other plausible assumptions: we may not be able to have our cake and eat it too when it comes to espousing realism and avoiding any metaphysical cost. For those who share the central intuitions on which I draw, it may therefore be time to reconsider the almost instinctive contemporary embrace of metaphysical naturalism. I will conclude with an exploration of a *dual aspect* version of non-naturalism as one form that a robust ethical realism might take.

1. ETHICAL REALISM AND ROBUST ETHICAL REALISM

It will help to begin by clarifying what we mean by 'ethical realism.' While this in itself is a matter of some controversy, it is agreed that ethical realists accept *at least* the following:

(1) Ethical claims purport to state facts (attributing ethical properties to actions, persons, policies, etc.), and so are straightforwardly true or false in the way that other purportedly fact-stating claims are, by accurately representing the facts or not; and

(2) At least some ethical claims, when literally construed, are true.

On this broadest characterization, ethical realism contrasts most obviously with non-cognitivism or expressivism (ruled out by the first claim) and with error theory (ruled out by the second claim).[6] It also, however, excludes Korsgaard's neo-Kantian constructivism, despite the fact that the latter combines cognitivism with a rejection of error theory. This is because

[6] Sayre-McCord (1988) and (2005). Ethical realism, as defined by the above two claims, also contrasts with "non-descriptivist cognitivism"—a view defended by Horgan and Timmons (2000). Their view is that while ethical claims express genuine beliefs, they are non-descriptive.

such constructivism construes moral truth very differently from how it is understood in claim 1 above. According to ethical realism, there are ethical facts to be discovered or recognized, which serve as truth-makers for the ethical claims that accurately represent them;[7] and insofar as there are correct procedures for answering ethical questions they count as correct because they track those ethical facts.[8] By contrast, according to neo-Kantian constructivism things are reversed: we begin with certain procedures that are authoritative not because they are correct (in the sense of tracking ethical facts) but simply because they are practically necessary for the exercise of agency, solving a general practical problem; and then whatever principles and answers they yield may be said to be 'true' simply in the constructed sense that they are so yielded by such agency-enabling procedures—not because they accurately represent any prior ethical facts. Truth and knowledge are derivative and secondary notions, with the real focus placed instead on normativity rooted in practical necessity. Such a view, then, is not a form of ethical realism.[9]

Nothing, however, has so far been said about the independence or objectivity of ethical facts. Some hold that these issues are of no direct relevance to ethical realism, which they take not to favor objective views over subjective ones or to require any kind of independence other than what is necessary to secure the second condition above.[10] Sayre-McCord, for example, takes the two minimal claims so far to be sufficient to characterize ethical realism, without any need for additional conditions of independence or objectivity. Others add at least the following claim:

(3) There are ethical facts that obtain independently of our actual ethical beliefs or attitudes or practices, both on an individual and a societal level, *at least* in the sense that such facts about what is right, or what is good for us, or what reasons exist are not simply a direct function of these things as they stand (as on crude subjectivism or relativism).

This condition makes explicit a commitment on the part of ethical realism to a certain kind of independence and a certain resistance to deflation. It is not that ethical standards need to be wholly independent of us or of facts about human life, but there are certain kinds of dependence that seem

[7] Shafer-Landau (2003: 15 n. 2).

[8] Korsgaard (1996: 36–7).

[9] I critique Korsgaard's view in FitzPatrick (2005). It is defended in Korsgaard (1996, 2003) and elsewhere.

[10] On the first point, see Sayre-McCord (1988: 5) and Brink (1989: 21). As Sayre-McCord puts it: "*Realism is not solely the prerogative of objectivists*" (1988: 16). On the second point, Sayre-McCord argues that in the characterization of realism, independence properly comes into play "when, *but only when*, [it is] relevant to whether the [claims in question], literally construed, are literally true" (1988: 6).

antithetical to the spirit of ethical realism if it is to be a useful and interesting category. For example, while a crude subjectivist or relativist view combines cognitivism with a rejection of error theory, and allows for straightforward truth for ethical judgments, by making ethical truths directly dependent on our actual beliefs or attitudes or practices, it hardly posits anything worth referring to as 'ethical reality.' If a claim to ethical reality can be made so cheaply, then one would be committing to so little in being an ethical realist that the category would lose any theoretical point, capturing everything from Platonism to crude subjectivism under one useless umbrella.

The best response for those, such as Sayre-McCord, who characterize realism only in terms of the first two conditions, is to agree that views such as crude subjectivism do not belong in the realist camp but to argue that we do not need a separate third condition to rule them out because they are already ruled out by the 'literal construal' clause in the second condition. That is, it might be argued that such views flout platitudes associated with our ethical concepts and discourse in such a way that they fail actually to engage these concepts, and so are not really accounting for the truth of ethical judgments literally construed, but are in a sense changing the subject; they therefore fail the second condition.[11] While this is plausible enough in simple cases; however, I believe there are problems with relying on this sort of response to carve out the appropriate space for realist ethical theories. But set that debate to one side. Whether we include the third condition explicitly or leave it implicit, a more substantive question arises about just how much independence and objectivity should be associated with ethical realism as such.

Again, many will stop with 1, 2, and 3 (or with 1 and 2 understood as already incorporating 3), so that ethical realism includes in its scope sophisticated subjective approaches wherein evaluative or normative facts are *constructed via some sort of idealization* from agents' beliefs, desires, responses, practices, agreements, and so on. Consider a naturalist view according to which facts about what is good for someone are facts about what her ideal self would desire her actual self to desire for herself, where her ideal self is imagined to have undergone deliberative procedures starting from

[11] Cf. Joyce (2002: 1–5), who puts the point about platitudes in terms of "non-negotiable parts of the discourse" in question, abandonment of which amounts to talking about something else altogether. Compare the reaction we might have to someone who, in a theological debate, offers an account of God as love and suggests that the existence of God is therefore much less controversial than has previously been thought. The principal objection to this is that he isn't actually talking about God: he may be fully realist about *something*, but that something is not what we were trying to address in theological inquiry. Thanks to Geoff Sayre-McCord for pressing this point. Cf. also Schroeder (2005: 1–4).

her actual desires and with full empirical information and no irrationality. This creates enough space for error and correction to satisfy 3 above, and although it makes truths about personal good a function of certain procedures and mental states, it need not run afoul of claim 1 as Korsgaard's constructivism does. It needn't involve any deflationary construal of what it is for an ethical claim to be true, but just a substantive view about what figures into the truth conditions for claims of goodness—namely facts about what a person would prefer under certain conditions.[12]

Many views widely regarded as paradigms of contemporary ethical realism fall into this broadly subjective camp, appealing to some preferred perspective—such as that of a fully rational, informed agent, perhaps taking up a certain point of view—and constructing ethical facts from the desires or responses yielded from that perspective.[13] In pointed contrast to this taxonomic inclusiveness, however, Shafer-Landau builds significantly more objectivity and independence into the characterization of realism in the form of *stance-independence*, with the addition of a fourth claim:

(4) There are moral truths that obtain independently of any preferred perspective, in the sense that *the moral standards that fix the moral facts are not made true by virtue of their ratification from within any given actual or hypothetical perspective*.[14]

The addition of this condition rules out ideal observer/advisor theories of the sort just described: for even if we could characterize an ideal advisor in such a way as to yield only morally correct results, ''the responses of the ideal [advisor] would not be constitutive of moral truth, but [would] merely bear a very close (perhaps perfect) correlation with a set of truths whose conditions may be fixed without any reference to such an [advisor].''[15] On this conception of realism, then, such naturalistic views as Railton's and Smith's strikingly fail to count as realist after all, despite their standard classification as such.

For simplicity, I will follow convention and count views that satisfy the first three conditions, including idealized subjective approaches, as broadly

[12] Someone who holds the view described above holds that there are abstract facts about my good that are there to be discovered, which make judgments about my good true insofar as they accurately represent those facts; it's just that those facts consist in facts about what I would prefer under certain hypothetical conditions that include the employment of certain procedures. And any *actual* procedure we employ in seeking to discover these important, complex facts—for example, the imaginative procedure you use in trying to figure out what my good would be—will count as correct or reliable only insofar as it tracks those facts.

[13] See e.g. Railton (1997), and Smith (1994).

[14] Shafer-Landau (2003: 15).

[15] Ibid. 16.

realist. Shafer-Landau, however, has a point in resisting that categorization, and even if we do not follow him in denying such views the realist label, we can see his addition of the fourth condition as articulating what is in any case a more *robust* ethical realism. Too see the point, consider what we would naturally think of as a realist theory of practical normativity. It seems doubtful that the first thing that should come to mind is a view such as Williams's neo-Humean internal reasons theory, where although there are facts about reasons for acting, they are constructed from and restricted by facts about each agent's desires: that is, there is a procedural motivational condition, tied to the agent's desires, on what can qualify as a reason, which is in fact taken to provide the very *content* for claims about reasons.[16]

Granted, such a view posits facts about reasons in a way that expressivist views of reasons discourse, for example, do not. So if we are contrasting such a view with expressivism, we might call it a broadly realist view of normativity. But there is also something odd about labeling as 'normative realism' a view that holds normativity hostage to people's contingent desires to such an extent (even given the extended reach provided by potential deliberative transformations of existing desires), and deflates it by reducing it to hypothetical facts about motivation.[17] This comes out more clearly if we contrast such a view with, say, a value-based external reasons view according to which we can be confronted by irreducibly normative, value-based reasons, the reality of which is undiminished by our motivational deficiencies. If we stick with a minimal construal of realism (just claims 1–3), then both of these views qualify as paradigms of normative realism,

[16] Williams (1981 and 1995). According to Williams, it is not clear what we could mean by saying that R is a reason for someone to ϕ except that "if he deliberated rationally, he would be motivated [by R to ϕ]" (1981: 109); and, for Williams, deliberation starts essentially "from the motivations that [an agent] has in his actual motivational set—that is, the set [S] of his desires, evaluations, attitudes, projects, and so on" (1995: 35).

[17] On the rejection of reductive analyses of what it is for something to be a normative reason, see Parfit (1997: 121–2), 2007, and forthcoming, and Scanlon (1998: 57 ff), among others. Though Smith holds a dispositional theory, he too rejects anti-rationalistic versions such as Williams's: for if the desires resulting from idealized reflection or deliberation were "a function of the desires [agents] actually have to begin with, the desires they were caused to have by the forces of socialization and enculturation that made them what they are," then they would still stand in need of vindication, as they stem ultimately from causal forces beyond our control. Instead, Smith adds a "rationalistic gloss" to his conception of reflection or deliberation, according to which rationality is able to neutralize the various differences in our initial desires as we reflect, resulting in a convergence of post-reflective desires in rational agents. See Smith (1997 and 1994: 164–74). If successful, such a rationalistic dispositional account would avoid worries about holding normativity hostage to contingent psychology. More on this later.

standing shoulder to shoulder in the realist camp. But their differences—in terms of the kind of reality they give to normativity—seem more significant than their similarities: one view deflates and reduces normativity, the other posits it as an irreducible reality largely independent of our desires, which also fits naturally with non-deflationary thinking about moral normativity and preserves the possibility of its categoricity. Focusing on this contrast, rather than on the one with expressivism, it seems much less attractive to categorize Williams's view as normative realism, despite its satisfying the first three conditions. The other view seems a much more natural candidate.

This illustrates, then, the pull in Shafer-Landau's direction, and even if we allow subjectively oriented views to count as broadly realist, we should recognize the important difference between realist views that meet only the first three conditions and those that meet the fourth as well. The latter are in any case more robust forms of realism, positing a *non-deflated, more independent, less derivative* ethical reality, as illustrated by the value-based external reasons view mentioned above, which is a plainly more robust form of normative realism than a view such as Williams's.[18] As a start, then, the robust ethical realism I shall be interested in includes Shafer-Landau's independence condition (4), or as I shall put it:

(4*) Ethical standards and facts are independent of us in the sense that they are not constituted by the actual or hypothetical results of *any ethically-neutrally specifiable set of conditions or procedures applied to our beliefs, desires, attitudes*, etc.

In what follows, I will examine some plausible motivation for realists to embrace 4 and 4*, along with four other ideas that fill out the characterization of *robust ethical realism* in my intended sense: the irreducibly evaluative or normative character of ethical properties and facts, the non-relativity of ethics (in a sense that still allows for significant pluralism), the autonomy of ethics, and the categoricity of moral requirements. I will then show why this robust ethical realism would commit one to a non-naturalistic metaphysics of ethical properties and facts that is incompatible with the view that they are exhaustively constituted by natural properties and facts.

[18] An internal reasons theorist such as Williams will of course object that the more robust realism about normativity is *too* robust for the subject matter, doing violence to the very phenomenon it is trying to capture by severing the connection to the agent's psychology that makes normativity intelligible. I try to answer this objection that external reasons theory is plagued by problems of normative alienation in FitzPatrick (2004).

2. AN ARISTOTELIAN MOTIVATION FOR 4*

What might drive some of us who are attracted to ethical realism toward a more robust form of it that embraces 1–4*, as against views that satisfy only 1-3, such as idealized subjective approaches?[19]

One reason we might take this course is skepticism about the prospects for capturing the full range of ethical facts by appeal to any set of hypothetical conditions and operations characterizable in an ethically neutral way, as by looking to the contents of people's psychologies as modified by more empirical information and formally specified deliberative procedures.[20] We might be significantly more confident in certain moral claims about rights or dignity than we are that those conclusions would necessarily be arrived at (or approved of, or desired to be conformed to) by any person, regardless of her experiential background, who is fed empirical information and proceeds to seek rational consistency and coherence; or we might be more confident that a marriage based on mutual respect as equals would be better for someone than we are that he would necessarily come to want his actual self to want this if only he had full awareness of the empirical facts and allowed deliberative processes to work on his desires in light of that awareness and without irrationality; or we might be more confident that considerations of fairness provide a genuine reason for a man to share in the care of his children than we are that he would come to be motivated by these considerations if only he were to deliberate with full empirical information starting with the elements of his actual subjective motivational set.[21] If so, we will obviously not rush to settle for any such philosophical account of the nature and source of ethical facts and properties.

[19] Enoch (2007*a*) offers a positive argument for robust realism about normativity, on the grounds that irreducibly normative truths are "deliberatively indispensable." I have doubts about whether there is any such argument to be found that would be convincing to opponents. My own strategy in this sphere, again, is to settle for illuminating plausible motivations for robust realism. While this is in one way less ambitious, however, I think it is also in another way stronger. Enoch says that if his one argument fails, he is prepared to give up robust realism (p. 4). I hope to show that there are a number of related, deep, and plausible motivations for embracing robust realism (at least given certain sympathies that at least some of those attracted to realism share), quite independently of his claims of deliberative indispensability, though these are related to his similar concern with the deliberative standpoint.

[20] For a critique of such views, though with a somewhat different emphasis from what follows, see Enoch (2005).

[21] Cf. Shafer-Landau (2003: 42) on greater confidence in certain substantive ethical views than in such procedures. I discuss the last example, in arguing against Williams, in FitzPatrick (2004). Cf. also McDowell (1995*a*).

Why might we be skeptical about the ability of such approaches to capture the full range of ethical facts? Even apart from dissatisfaction with the details of any particular view, we might take this stance because we suspect, with Aristotle, that getting correct results from deliberation depends crucially on having the right starting points—and that the *right* starting points cannot be correctly specified except from within a correct substantive ethical view about what constitutes a proper ethical upbringing, or which character traits embody proper sensitivities to relevant real values and other considerations, as developed through the right kinds of ethical experience.[22] I shall not here try to show that this is true, though I believe it is. What I want to notice is that *if* one does find this basic Aristotelian claim plausible then it constitutes an important motivation in the direction of robust ethical realism and sheds light on what it would involve. It implies that we cannot look to facts about idealized deliberation to provide an ethically external foundation for ethics, available to all parties to ethical disputes, with which to underwrite one substantive ethical view; we cannot in principle settle ethical disputes from a common, neutral perspective involving mere empirical information and procedural rationality.

Suppose we are thus led to give up on such ethically external foundations in idealized deliberation. If we maintain a commitment to ethical realism, then we will still hold that one substantive ethical view or range of views is nonetheless correct: it's just that it will be correct in a stance-independent way, better representing the ethical facts than rivals, rather than because it is ratified by some special independently specifiable deliberative procedure. This appeal to stance-independence may seem puzzling, given that a kind of stance-*dependence* was central to the above Aristotelian claim. But it is important not to confuse issues here. The Aristotelian claim was that there is no way properly to characterize the route to ethical truth except from within a correct ethical perspective: no one who lacked substantial ethical knowledge to begin with could lay down a set of conditions and procedures for arriving at ethical truth, because *there are no* such conditions and procedures that do not presuppose substantially correct starting points. But this does not imply stance-dependent *truth-conditions* for ethical claims.

[22] This has been a central theme in much of the work of McDowell and Wiggins (cited throughout)—though I want to separate this general idea from their further construal of values on the model of secondary qualities, which is a distinct and, I think, problematic move (for reasons given below). On the idea of right starting points involving character and relevant sensitivities, see the preceding note, as well as Rosati (1995) and Scanlon (1998: 57), among others. As discussed below, however, Rosati does not in her more recent work draw out the deeper implications of this point for open question arguments, which I believe leads her to miss the force of open question arguments against even forms of naturalism that meet her conditions of fit with agency.

There is no suggestion that what *makes* ethical claims true is that they would be endorsed by people deliberating from certain special starting points. In fact, that couldn't be so on this view, since among the true ethical claims are claims about what those proper starting points are. What makes ethical claims true for the realist we are now considering—who accepts the Aristotelian claim and rejects idealized deliberation views—are just a variety of ethical facts that *also* determine, among other things, what the right deliberative starting points are or what counts as a proper upbringing. This is a stance-independent view in the sense of 4 and 4*, positing an interconnected set of evaluative facts that cannot be cashed out in terms of uncontroversial empirical and rational refinements of our psychologies.

3. SENSIBILITY THEORIES, NON-RELATIVISM, AND OBJECTIVE VALUES

This goes importantly beyond even non-reductive, response-dependent sensibility theories, and for good reason.[23] By working within a secondary quality model of values, such theories have difficulty avoiding excessive metaethical relativism, and so lack the robustness I am concerned with. One can make semantic moves that might seem to get around such relativism by allowing for non-relativistic counterfactuals. For example, we could argue that 'good' works like 'red' in that 'red' *rigidly* designates the range of reflective qualities it picks out in the context of our *actual* color sensibilities, allowing us to say that had we instead had an inverted spectrum we simply would have seen red things as blue, rather than having to say that in such a case the colors of things would have been flipped; and something parallel could thus be said about goodness.[24] But this is a superficial gain. The bottom line is that such views about values allow for multiple practices within which different sets of evaluative concepts and claims are internally justified in relation to the relevant sensibility and its system of reasons, but no clear sense in which one stands *asymmetrically justified* above the others (i.e. justified above the others not merely as judged by its own standards, which every other view can—symmetrically—do for itself as well, but in a way that sets it apart), as having a special normative

[23] Sensibility theories have been developed and defended by John McDowell and David Wiggins. See esp. McDowell (1997*a*, *b*) and Wiggins (1997). For an excellent general discussion of sensibility theories, see Justin D'Arms and Daniel Jacobson (2006).

[24] See Wiggins (1997). For an argument against such a rigidifying move specifically in connection with intrinsic value or welfare, see Peter Railton (2003). Cf. also Lewis (1989).

claim to our practical allegiance.[25] A more objective justification of this sort would seem to require further realist commitments to more objective values going beyond any "sensible subjectivism," to give content to the idea that one sensibility or range of sensibilities is closer to getting things right than others, or that certain conditions within which moral responses take place are the truly 'suitable' ones—even though *our arguments about what those values are* or our attempted vindications of the reasons in favor of them will always take place from *within* an ethical sensibility, rather than from some external perspective, which is just the epistemological or methodological part of the Aristotelian point again.[26]

Michael Smith suggests an alternative, metaphysically cheaper way of solving the above problem by "completing" McDowell's dispositional account of moral concepts with a rationalist account of 'suitable conditions'. These will be the conditions under which desires and evaluative reasoning are "controlled by the particular norms of practical rationality to which moral norms reduce," allowing for a justification of the values associated with responses under those suitable conditions—giving a sense in which such value is a property that is "there to be experienced," and in which ascribing value is expressive of categorical reasons.[27] This is an interesting proposal, but again it puts a great burden on the appeal to rationality. In supposing that moral norms reduce to norms of rationality we would have to suppose that rational reflection, regardless of people's various experiential and affective starting points, has the power to neutralize those differences and to "lead us all to converge on the very same desires" about actions. This would be based on the idea that (1) "what we believe when we believe that it is desirable to perform a certain act in certain circumstances is that we would all converge upon a desire that we act in that way in those circumstances if we attempted to come up with a maximally coherent and unified desire set under the impact of increasing information," and (2) there would indeed be sufficient convergence of this sort to account for the full range of moral facts.[28]

Those sympathetic to the earlier Aristotelian claim, however, are unlikely to think that "what we believe" when we believe that an action would be good (or desirable) to perform is something about the convergence of

[25] See Darwall, Gibbard, and Railton (1997*b*: 22–3). As they note, mere "proprietary labeling" of the sort described above fails to address the real justificatory issue.

[26] I think that a similar problem arises for Scanlon's appeal to correct standards for judgments about reasons, as grasped from within substantive ethical reflection, while eschewing any metaphysical appeal to objective values to make sense of the intended non-relative notion of correctness. See Scanlon (1998: 63–4).

[27] Smith (1993: 250–2). On this view, genuine values would be natural features of acts, for example, that "elicit certain attitudes in us when our thinking is in accordance with [the norms of practical reason to which the norms of morality reduce]" (p. 251).

[28] Smith (1997: 317).

desires under idealized circumstances specified in some morally neutral way.[29] More importantly, they will be skeptical about the prospects for such convergence. After all, whether one is a response-dependent sensibility theorist or a more robust realist of the sort I have in mind, much of the point of appealing to rich practices embodying engaged moral sensibilities is a sense that ethical values and distinctions go far beyond what can be derived by disciplining desires with empirical information under constraints of coherence and unity. We could, of course, scale back our conception of the content of ethics to something fairly minimal, in which case it might become much more plausible that rationality could deliver the goods. But that again seems like a deflationary move, and I am interested in where we are led if we resist such deflation.

Again, my point is not to try to show that rationalism cannot be right, but just to flag and illuminate a reasonable motivation for looking to some other way of making sense of the special status of one sensibility or set of ethical standards over rivals. For the robust realist I have been describing, this points toward positing more objective evaluative facts—such as the fact that a certain kind of ethical upbringing or set of character traits or sensibilities counts as a proper one, fostering correct moral responses that track values determinative of facts about how it is good and right to live. On such a view, these facts are all connected, and there is no suggestion (for reasons further brought out below) that any can be extracted and cashed out from an ethically-neutrally specifiable perspective available to all parties to substantive ethical debates, providing some sort of external foundation for ethics.[30]

4. THE AUTONOMY OF ETHICS: THE DELIBERATIVE PERSPECTIVE AND EMPIRICAL INPUT

The robust realism I am exploring combines the broadly Aristotelian line of thought sketched so far with a closely related concern: a sense that

[29] For an insightful critique of Smith's views, see Enoch (2007*b*).

[30] Cf. McDowell's (1980) characterization of the proper way to understand Aristotle's claims about the connection between virtues and *eudaimonia*. He does not see Aristotle as engaged in the foundationalist project of appealing to an ethically-prior theory of human nature to underwrite a conception of *eudaimonia* and thereby derive a specification of the virtues as whatever qualities promote it. Instead, the virtues and *eudaimonia* are to be fleshed out together from within the standpoint of proper ethical reflection, with the teleological theory serving more as a conceptual framework than as an ethically-prior source of ethical substance. See also McDowell (1995*a*).

ethics is essentially *autonomous*. By this I mean that truths in ethics—truths about morality, reasons for acting, excellence of character, what is good for a person, and so on—can be accurately and justifiably arrived at only through engaged first-order ethical reflection and argument employing its own internal standards, and not from the outside through some other form of inquiry, such as psychology or biology.[31] This is not to deny that other forms of inquiry can and must inform ethical inquiry: obviously they must do so insofar as it is crucial to successful ethical reflection to get the facts of life straight. But the *significance* of what these other forms of inquiry contribute—such as facts about evolved psychological traits, or about what tends to make us feel satisfied, or to promote our survival, or to stabilize our social life—must still always be assessed through the lens of autonomous ethical reflection on our life and experience.

The reason for this is that nothing presented to a rational agent in any other way could be *authoritative* for her. We needn't accept anything as strong as a neo-Kantian constructivist idea of agent autonomy (as self-legislation of morality) to see this point. To take a crude example, suppose evolutionary psychologists tell us that a disposition for philandering is an adaptation in male humans just as it is in male Superb Fairy-wrens, present in them for similar evolutionary reasons, and philosophers add that on the best account of natural teleology any such adaptation has a proper biological function; philandering when possible is thus as much a part of the proper biological functioning of male *Homo sapiens* at the behavioral level as it is of male Superb Fairy-wrens—and refraining counts as defective. Even if this were all true, it should be plain that it could have no normative force as such for a human rational agent.[32]

A (practically) rational agent, after all, is characterized precisely by the ability to step back from given desires, inclinations, drives, dispositions, etc.—to establish a "reflective distance" from them—and to ask whether they are *worth* pursuing or *ought* to be pursued, with a view to conforming choices to those evaluative and normative judgments. This capacity is precisely what gives rise to the possibility of (and need for) reasons for acting.[33] But for such an agent, nature as such can have no authority: that something is natural, or even has a proper natural function, cannot in itself settle anything in its favor, since it belongs to the rational agent always to ask: "but is this a worthwhile way for me to live?" And this

[31] See Nagel (1979). McDowell and Wiggins have emphasized similar points in much of their work.

[32] See FitzPatrick (2000) for an extended critique of such appeals to nature in ethical inquiry.

[33] This is well brought out by Korsgaard (1998: 62 and 1996: 93). See also Allison (1990).

is not settled for such an agent by any underwriting from evolutionary theory, even as filtered through a naturalistic normative theory of proper functions.[34] He can always ask why he should care, and whether it is really good and worthwhile, and the only way to answer that question to the proper satisfaction of a rational agent is through ethical argument appealing to his understanding and experience of what makes for a good life. In the present case, one can acknowledge whatever is claimed about natural teleology—even that it provides non-subjective norms of a sort—and reasonably reject it as having no normative authority whatever for us, on the grounds that philandering is inconsistent with the goods and values one has found through experience and ethical reflection to be most important in life.[35]

The above example is deliberately simplistic, but the point applies generally. Suppose a credible psychological theory showed that human beings tend to be less satisfied if educated beyond a high school level, especially if this brings about a more realistic understanding of the disturbing complexities of life, undermining cultural or religious myths that once provided comfort. What is the significance of this finding for an understanding of a person's good, or of how one ought to live? My claim is that as a deliverance of empirical psychology, it so far has none. It gets any particular normative significance it may have in these respects only by way of becoming *input* for substantive ethical deliberation—which might well discount such facts in favor of values found within this reflective perspective to be more important than such satisfaction.[36]

This point about the role of the deliberative perspective is widely recognized and is not unique to robust realism. What sets robust realism apart

[34] Cf. Rosati (2003). The crude form of evolutionary naturalism sketched above is an example of what Rosati calls "brute naturalism," and her appeal to the autonomous evaluation and reflective distance associated with agency (2003: 506 ff.) in diagnosing why such views fail is very similar in important ways to my own. There are also important differences, however, in terms of how much work the appeal to autonomy-making motives and capacities is taken to do in itself, and whether forms of naturalism that have a better "fit" with agency thereby avoid the basic problems. More on this below.

[35] See McDowell (1995*a*). Foot (2001) has developed a form of teleological naturalism wherein natural teleology is entirely independent of evolutionary theory, making for a *prima facie* more plausible appeal to natural teleology in ethical theory. Her approach, however, still falls prey to McDowell's objections, and there are also deep problems with her ahistorical, welfare-based account of natural teleology, which I explore in FitzPatrick (2000). I argue that a satisfactory account of natural teleology must indeed appeal closely to evolutionary history (even though teleological explanation does not reduce to historical-causal explanation), making natural teleology plainly unsuited to shed any light on ethical standards and normativity.

[36] This may be getting at what Wiggins (1993*b*: 335) means when he says that "in a sense, moral science has here *swallowed* social science."

from various forms of constructivism here, however, is the commitment to the idea that *there is such a thing as getting such deliberation right or wrong, where this is not settled merely by standards of procedural rationality and adequate empirical information* (and is not to be understood merely on an expressivist analysis, and so on). This ties into the earlier Aristotelian point about right starting points in terms of experience and character: the deliberative perspective is the necessary channel through which empirical information must pass in order to acquire genuine normative significance for us, but this is not to say that it is sufficient, as if whatever results from rational deliberation on such information and on the contents of one's psychology is thereby correct; on the contrary, according to my robust realist, the ethical facts we are after have their standing independently of such ratification, and their proper recognition through deliberative reflection will at least often require that the starting points of the deliberation be of the (non-reductively specifiable) right sort.[37]

On the robust realist view, then, what has genuine normative force for us may well buck what perhaps seemed to be a normative result handed down to us from scientific inquiry, for reasons related to but going beyond those given by constructivists. So to take the above example again, psychologists may discover that more educated people are less satisfied than others by empirical measures, but if such facts would be found within properly informed, ethically sound reflection to pale in significance to other considerations (as they did for Mill in reflecting on Socrates and the pig), then they lack normative force for us and no amount of scientific backing will change that. In other cases, for example with respect to a psychological finding about the role in human satisfaction of being deeply understood by a friend or lover, we might find ourselves reflectively giving it great weight as a value. But for the robust realist it is not merely this reflective endorsement that matters: the hypothetical or actual exercises of agency involved are crucial to the existence or grasping of normativity, but they are not sufficient; the normativity is not reduced to the exercises of agency. Still, the point against the naturalist appeals sketched above is that it is only through such deliberative agency that a natural consideration's normative authority for us is manifested, assuming we are correct in our deliberations.[38]

[37] For a robust realist it will not, of course, be a condition of a consideration's having normative force that the agent actually deliberate in this way (though that will be crucial to the agent's *grasping* that normative force). It will be enough if the consideration is such that it would be recognized as normatively significant from a properly informed deliberative perspective, where this includes right ethical starting points.

[38] Similarly, if this is right, it is only from such a properly informed ethical perspective—rather than from that of the scientist—that we can adequately characterize

5. OPEN QUESTION ARGUMENTS AND AGENCY

It is precisely this point about the nature of rational agency and what can be authoritative for a rational agent that underlies the persistent and plausible appeal of what is often referred to as a kind of 'open question argument.'[39] It is by now well known that Moore's original version of the open question argument fails to hit its mark, not only against non-analytic or reforming naturalism but even against sophisticated forms of definitional or analytic naturalism.[40] But it is widely held that *something like* the open question argument nonetheless remains potent against a large range of views, revealing failures to capture the normativity of the claims in question. Here's Wiggins against Railton's naturalism:

> I think it is no accident that it is well-being that Railton seeks to connect with the obligatoriness of some act. For well-being may seem to promise to pull its weight in a social theory. But if that is the place from which the would-be vindicator proceeds, he cannot help but leave himself open to the retort: 'To do *A* may promote human well-being as naturalistically specified. But it is an open question—indeed doubly open—whether it is indeed obligatory.' First doubt: is the naturalistic version of well-being something we fully recognize as the proper object of all our striving? Secondly, can one here get from statements about it to statements of right and wrong?[41]

these goods and distinguish them from close imitations that may look the same from a scientific point of view. Compare distinguishing good from bad music from an aesthetic perspective, even though both might cause satisfaction in many people, of a sort that may look the same from the perspective of a research psychologist and her empirical metrics.

[39] See Rosati (2003: 506 ff.), for a similar view, though there are also important differences, discussed below.

[40] See e.g. Brink (1989: ch. 6 and 2001: 157 ff.), on the problems with the "semantic test of properties," or with reliance on the "descriptional theory of meaning" combined with the view that meaning determines reference. See also Smith (1994: ch. 2), and Rosati (1995). Sturgeon (2003) points out that Moore's open question argument also neglects the possibility that a term such as 'good' or 'right' stands for a natural property that happens not to be picked out by any *other* descriptive term D, which would thus equally account for the openness of any such question as "Granted that X is D, but is it good?" As Dancy (2006: 132) notes, a view such as Sturgeon's also escapes Parfit's "triviality objection," since there is no descriptive characterization D to use to demonstrate the problematic triviality or redundancy (as in "This act of D-ing [e.g. maximizing utility] is right," which would just be ascribing the same property twice).

[41] Wiggins (1993*b*: 335–6). Similarly: "An act may well 'contribute to aggregate well-being [or happiness], where this includes the alleviation of suffering,' but, even allowing that it does contribute in that way, it must still be an *open question* whether the act or practice is good (is, as Railton puts it, better from the moral point of view)" (Wiggins, 1993*a*: 304). See also Rosati (1995) and, specifically on attributions of reasons, Scanlon (1998: 58).

Why do the crucial questions remain open? Ultimately it is because for a rational agent, there is no automatic normative authority attached to any such natural fact as that an action will promote some naturalistically specified state of affairs—whether this has to do with natural teleology, or people's satisfaction, or something more complicated, such as (in this case) the satisfaction of the desires the people affected by the action would desire themselves to have if they were fully rational and informed. To be told that an action will have such an effect is at most to be given a fact (among others) to take account of within engaged ethical reflection, and the results remain to be seen.

It might seem that the special content of the facts in this case *should* settle things. After all, reflective endorsement is *built in* to the fact about well-being, for example, as described above: the desires in question are ones that would be reflectively endorsed under the specified conditions. So doesn't that accomplish the same thing, in terms of normative authority, as what I've been describing (unlike appeals to purely objective facts, such as facts about natural teleology)? It does not. Even if we just stick with the first-person case involving well-being, the point is this: what is presented to me as a rational agent occupying an ethically engaged deliberative perspective is still just a psychological fact about my hypothetical second-order desires under certain conditions specifiable in some ethically neutral way—a fact that, like any other presented to me, I must *still* reflect on from my actual, ethically engaged reflective perspective in order to assess its normative significance. And again, one thing I may sensibly wonder from that reflective and critical perspective is whether my *own* psychological starting points are sound.[42] Thus, not even being told what I would desire myself to desire if I deliberated with empirical information and without irrationality will *settle* things, because that misrepresents what we are after when we deliberate: we are not aiming at discovering and conforming to any such hypothetical deliberative results, even in connection with ourselves, but at discovering and conforming to the truth about what is good (or right, or what there is reason to do, and so on), as such.

This is closely related to, but goes importantly beyond, Connie Rosati's diagnosis of why some forms of naturalism fail. In "Agency and the Open Question Argument," she construes informed desire accounts more narrowly than I have above, such that they appeal only to information and rationality but not necessarily to any autonomous deliberative processes (operations of "autonomy-making motives and capacities") that might "improve on or idealize an individual's motivational system."[43]

[42] See Rosati (1995), esp. with the example of Sandy on 53–4.

[43] Rosati (2003: 514–15).

Her objection, then, is that this is why they leave a problematic open question and fail to capture normativity: they do not "capture the structure of or otherwise engage with or reflect autonomous evaluation"; a person can still "wonder whether what she would approve under these conditions [of rationality and information] reflects her autonomous evaluation or her purely eccentric features."[44] What I am claiming, however, is that the problem runs deeper: even if we construe the naturalistic account as incorporating such autonomous evaluation and deliberation (as Williams's account of reasons does, for example), there remains a problematic open question because of the earlier point about proper starting points: the mere fact that I would approve of something if I reflected autonomously and with empirical information, starting from my actual body of experience and set of character traits, does not settle things, because I can wonder whether those starting points were sound; perhaps I am missing something important, and would continue to miss it even after deliberating autonomously with an encyclopedia, because of some relevant impoverishment in my ethical background and experience that needs to be corrected. I therefore do not think that Rosati has fully captured the source of the open question worry in the issue of fit with "autonomy-making motives and capacities" that are constitutive of agency as such, and I do not think that the problem with naturalist accounts can in principle be solved just by securing such a fit within the account.[45]

For anyone sympathetic to the Aristotelian point about the limits of deliberative procedures and the need for substantively right starting points, *any* general account that tries to capture the evaluative or normative in

[44] Ibid. 115.

[45] I think that Rosati, following others such as Korsgaard, is trying to get too much from agency itself (i.e. from motives and capacities constitutive of agency). Her own focus, of course, is on the idea of personal good, and perhaps this makes it more tempting to rely so heavily on agency, since autonomous evaluation may seem especially authoritative with respect to one's own good. I believe, however, that the problems raised in the text apply even here (i.e. that a person can be mistaken about her own good even after autonomous evaluation with empirical information, if her ethical starting points are impoverished in certain ways). More importantly, even if Rosati's strong appeal to agency were sufficient for the case of personal good, there is little reason to suppose (as she suggests in 2003: 499 n. 27) that this would translate more generally to morality, for example, especially given her appeal to a quite liberal choice of self-ideals as part of an agency-fitting account (2003: 523). While some may find my worries about substantively impoverished starting points less compelling when it comes to the determination of one's own good, it is harder to dismiss them when it comes to moral values, where it is far less clear why empirically informed autonomous reflection should automatically carry the day (regardless of the experience and character one starts from). For an excellent critique of appeals to agency to explain normativity, see Enoch (2006).

terms of biological or psychological facts about us will be plagued by problematic open questions. This is true even if the account appeals to facts about our hypothetical responses under 'ideal' circumstances that include rich exercises of agency, at least as long as the 'ideal' is characterized in ethically neutral ways, which is the real root of the problem. The robust realist will avoid open question problems by refusing to adopt *any* such schema. At most, he will accept an innocuous variant where the idealization is ethically loaded, which therefore will not give rise to a compelling open question worry. For example, there is no compelling question along the following lines: "Granted, I would desire to desire X for myself if I were fully *ethically* (as well as empirically) informed, that is, appropriately sensitive to all relevant values and other considerations, and deliberated rationally, but is X *really* good for me?" Unlike all the accounts employing ethically neutrally characterized idealizations, including agency-oriented ones, this sort of account leaves no compelling open question; there is nothing left to point to that might be being overlooked.[46] But it is also trivial and uninteresting. For the robust realist, that is just as it should be: the whole point is that *while there are real ethical facts, they are not capable of being captured by any non-trivial general formula in other terms*—whether by appeal to the sciences or even by appeal to the nature of agency—and that there is thus nothing useful to be said along those lines. To call this a failing, or to insist that the robust realist come up with something along those lines by way of developing the account, is to miss the point and beg the question against robust realism.[47]

[46] By contrast, in my employment of the open question argument against naturalist views, I could point to the plausible worry that they overlook the possibility that my deliberative starting points may be substantively impoverished in ways that distort the results of otherwise 'ideal', empirically informed deliberation. This makes for a legitimate use of the open question argument, whereas merely showing that one can raise a question about an account does not constitute an interesting objection if one lacks anything to say about *why* that account seems inadequate to capture something important about the evaluative or normative concepts in question—a point nicely brought out by Rosati (2003). The present point, then, is that the account suggested in the text builds so much in that there is no room left for a compelling open question.

[47] Thanks to Sarah Buss for pressing me to explain why the open question argument cannot just be turned around and used against the robust realist. The point above is that the robust realist will avoid any genuinely compelling open question arguments by refusing to embrace any schema that tries to account for ethical properties by appeal to some non-ethically characterized condition or ideal—a project he takes to be misguided. Obviously this does not mean that no one can raise questions about his metaethical claims, as if his being correct about metaethics entailed its being transparently obvious that this is so. But the existence of 'open questions' in this sense does not constitute an open question argument against the view, any more than the existence of similar open questions surrounding any non-trivial normative ethical view automatically constitutes an argument against the view. There is another kind of open question argument directed

6. OBJECTIVE VALUES AND IRREDUCIBLE NORMATIVITY

Thomas Nagel expresses a central tenet of what I am calling robust realism when he writes: "If values are objective, they must be so in their own right, and not through reducibility to some other kind of objective fact. They have to be objective values, not objective anything else."[48] The idea is that any adequate account of values as objective must at the same time retain their evaluative or normative character; otherwise, we're just changing the subject. Of course, as Mark Schroeder has argued, this does not automatically show that values cannot be reduced to natural properties. To show that, we would need to establish further that the natural properties in question don't have the requisite evaluative or normative character.[49]

Schroeder gives short shrift to the open question argument as a source of doubts about the evaluative or normative character of naturalistically reduced properties or facts, noting the inadequacies of Moore's original version of it.[50] I have tried to show, however, that the open question worry is tied to central and important motivations for resisting such naturalistic reductions, which I think explains the "strident pessimism" many of us feel about such reductions in general, even apart from their details.[51] This is important because, among other things, it creates problems for his strategy for making such reductions seem more plausible. He argues that naturalistic reductions in ethics can in principle get around problems of normativity by going *through* the concept of reasons, which is what normativity is all about, rather than by reducing directly to something else. So, for example, if we reduce rightness not directly to something about utility maximization, but to something about reasons, and then reduce reasons, we will have captured normativity in the picture.[52] The problem, however, is that if the open question worries raised earlier seem compelling, they will apply just as powerfully to reductions of reasons. This is nicely brought out by Scanlon:

> Even if there were a true, nontrivial biconditional of the form "Something is a reason if and only if a person would regard it as one under conditions C," this would not provide a satisfactory reductive analysis of what it is for something to be a reason. This is because "R is a reason" expresses a substantive normative judgment, while the

against all forms of ethical realism by Korsgaard (2003), intended to show that ethical truths not having their source in the will cannot have normative force for rational agents. I develop an answer to that argument in FitzPatrick (2005), as does Parfit (forthcoming).

[48] Nagel (1986: 138).

[49] Schroeder (2005).

[50] Ibid. 4.

[51] Ibid. 8.

[52] Ibid. 12–17.

> right hand half of such a biconditional (where C is free of question-begging phrases like "responding in the right way") remains a mere prediction of my reactions. As long as C is free of such phrases, the question "I would not regard R as a reason even under conditions C, but is it a reason nonetheless?" will have an "open feel".[53]

If what I have said earlier is at all persuasive, this "open feel" is quite justified and we will be as skeptical about the reduction of reasons as we were about the reductions of goodness or rightness.[54] And if we thus find the reduction of reasons unsatisfactory precisely because it fails to capture the normativity of reasons, then we won't regard the overall reduction as having adequately captured normativity in the picture just because it has something to say about reasons.[55]

As before, I do not claim to have shown here that such reductions cannot possibly be right (though I believe they cannot), but only to have provided some plausible motivation for resisting them, for the purpose of seeing where this leads. The robust realist who continues down this path will therefore maintain with Nagel, and more recently Parfit, that ethical properties and facts are *irreducibly evaluative or normative* even while being relevantly objective.[56] I take this to mean that if we are to avoid deflating moral facts, for example, then we need to see them as facts with an inherent normative significance, providing reasons directly because of *their own evaluative and normative content*, and not merely because of how they might bear on something else that is taken to be inherently normatively significant, such as an agent's morally-neutrally characterized self-interest. After all, any kind of fact—even plainly empirical facts with no normative content at all—can be normatively significant in the sense of bearing on other things in such a way as to provide reasons for acting.[57] But a non-deflationary

[53] Scanlon (1998: 58).

[54] Parfit (forthcoming) provides a number of important arguments against naturalistic reductions of reasons or normativity.

[55] As noted earlier, Smith's rationalistic view may do significantly better in this regard, though again it is also extremely ambitious in the work it requires of rationality.

[56] See Parfit (forthcoming). This sort of irreducibility is not to be confused with the irreducibility emphasized by "non-reductive naturalist" views, i.e. the idea that the complex natural properties referred to by moral terms are not themselves picked out by terms from the empirical sciences, and have distinct explanatory roles to play (just as might be said of biological properties in relation to physics). See Boyd (1997) and Brink (1989: ch. 7).

[57] Schroeder (2005: 13) gives the following example: "the fact that by pulling the trigger of a gun while it is pointed at one's skull one will ensure one's own death ... has as a necessary consequence that there is a reason for one not to pull the trigger of the gun while it is pointed at one's skull," even though it is not itself normatively contentful. Parfit (forthcoming) has made the same point: as Dancy (2006: 137) puts it, "a fact that has normative significance need not for that reason be a normative fact."

view of the normative character of moral facts will see them as facts with an inherent normative significance, providing reasons *because* of their own evaluative and normative content. To try to account for their normative nature simply by showing them to bear on other things taken to have inherent normative significance is a deflationary move, compromising the normative stature of morality by making it derivative and contingent on other things.

It might help if the element of contingency were removed—for example, if moral requirements were *necessarily* connected with a sort of self-interest held to be automatically reason-giving. But this is still ultimately deflationary: why should some independent conception of self-interest be taken to be more fundamentally normative than morality, which thus has to piggyback on it? Does it not deflate the normative stature of facts about human rights, for example, to see them as normatively authoritative for one only insofar as they happen to connect up in the right way with one's own interests? If human rights are so important as such, then it seems they ought to have non-derivative normative force in their own right. What I am interested in, in any case, is a realism that maintains such direct *categoricity* of moral requirements within the robustly realist framework I have been exploring (as opposed to a constructivist or Smith-type rationalist framework).

For those of us sympathetic to such ambitions, not only will subjective naturalist accounts such as Railton's (which explicitly disavows categoricity[58]) be unsatisfying, but so will more objective accounts, such as Boyd's, that are driven more by semantic and explanatory concerns than by a focus on capturing the evaluative or normative nature of ethical properties and facts. Consider a parallel, making similar semantic and explanatory moves. People use the term 'cool' to refer to a natural homeostatic cluster property—one that causally regulates our use of 'cool' and is thus tracked by it; and there are social practices of reason-giving and correction that make for significant objectivity: it is not merely subjective or based on individual feeling. (It is a matter of fact, insofar as there are any objective facts at all, that Miles Davis was cool and Gomer Pyle was not.) Moreover, the property of coolness figures into causal explanations: Why did Fonzy get more dates than Potsy? Because Fonzy was (much) cooler. Nonetheless, we can reasonably deny that we have any reason to care about and pursue this property, giving it weight in our deliberations. This property doesn't have any claim to merit our concern. By contrast, goodness, if it is not to be deflated, must be understood as a property that does *merit* our concern.

[58] Railton (1997: 140).

That's part of *what it is to have* a normative and evaluative character in the realm of ethics. So the semantic and explanatory points don't take us very far by themselves.

Of course, any tempting natural specification of the referent of 'good' will focus on something such as human needs, in which we naturally take an interest. It may then be pointed out that goodness, so understood, is something we *do* generally take an interest in (perhaps for evolutionary reasons), unlike coolness, and this might seem to take care of the issue with normativity.[59] But it doesn't. The claim above was not just that we do take an interest in goodness, but that goodness is the sort of property that merits our interest: we *ought* to be concerned to promote it, or to give it a certain kind of deliberative weight; if someone happens not to, then it remains true that he ought to. But that sort of thought is utterly missing here. Indeed, it is hard to see how it could possibly be captured within an objective naturalist picture: what objective natural fact or facts would such a fact about a natural cluster property's *meriting a certain practical response* consist in? The result thus seems to be just what Nagel complained about: 'goodness' has been given a clear objective referent and gainful explanatory employment, but only at the cost of being turned into something else.[60]

Moreover, the more robust realist I have been describing will not think that the naturalistically characterized needs figuring into the cluster property can do the required work as such. They are certainly relevant as *input* into substantive ethical reflection, but it is only through that reflection that we can see their normative force and relative weights in the determination of the good (assuming that our reflection is sound, in a sense that isn't reducible to any external, naturalistic criteria), and not through any kind of scientific inquiry. Perhaps if we stuck with basic biological needs this would be unnecessary, but of course that would be an impoverished view of the good for human beings (and an implausible one if it is meant to have any authority for us as rational agents). And once we allow talk of needs

[59] Even apart from the more important problem raised for such a suggestion in the text that follows above, we might note the limitations of that psychological point in any case. While it's true that we take a natural interest in our own needs, and typically some interest in others' needs, we *don't* generally take the sort of impartial interest in people's needs that goes with the consequentialism Boyd endorses, making it even harder to see how the relevant normativity is captured at all: not only is there no sense, on Boyd's view, in which we *should* care about some cluster property in which other people's needs figure impartially, but we don't tend to anyway.

[60] My appeal to the need for an apparently non-natural sort of fact about a property's *meriting* a certain practical response is very similar to Dancy's (2006: 137) appeal to non-natural "metafacts," such as *the fact that* a certain fact about an action *makes it the case that one ought* to perform that action. See also Parfit (2006 and forthcoming).

more generally—entertaining claims of need for intellectual stimulation, intimacy, artistic expression, social status, power, and so on—we are back to requiring substantive ethical reflection to determine their relative significance and normative force. Merely empirical, scientific inquiry is no help here.

7. NON-NATURALISTIC COMMITMENTS

This completes my elucidation and motivation of robust ethical realism as I understand it. What I want to emphasize now are the non-naturalistic commitments of such a view, which those of us who share the above motivations must be prepared to accept.[61] It is, of course, always difficult to characterize the divide between the natural and the non-natural, even within a given sphere of inquiry, and I will not try to settle how best to do this for ethics. It will be enough for my purposes to bring out the clear ways in which this view departs from common conceptions of naturalistic views and their commitments.[62]

While robust realism is not a stance-dependent view (it does not tie the truth conditions for ethical claims to the results of deliberations from some standpoint favorably characterized by some set of ethically neutral criteria), it does insist that ethical truths can be justified and grasped as such only

[61] Noell Birondo has pointed out that it may seem odd, given the deep McDowellian strains in my arguments in the text, that I am so quick to embrace talk of non-naturalism, while McDowell himself considered his view to be simply a different sort of naturalism (1995*a*). I hope to make clear in what follows why it seems best to me in the context of contemporary metaethics to conceive of robust ethical realism as a form of non-naturalism, though there is a sense in which it could also be seen as a re-thinking and broadening of the metaphysics of 'the natural world.'

[62] As an example of the complications with delimiting the natural and the non-natural, consider Crisp's (1996: 117) characterization of ethical naturalism as the view that "moral properties—those which would be identified by the best moral theory—are natural properties—those that would be identified by the best scientific theory, and which can be described in the conceptual terms available to a being occupying some non-local point of view on the world"—where a non-local point of view signals lack of substantive ethical engagement. In some ways this is close to my focus, but the characterization of natural properties is problematic. Even where naturalists make an identity claim for ethical and natural properties, there need be no claim that the best scientific theory will contain predicates that pick out moral properties, or recognize such properties as natural scientific kinds. As Brink (1989: 157, 194–7) emphasizes, the claim is only that for any moral property "there is a natural predicate that is *constructible* (e.g., by operations of conjunction and disjunction) from predicates that do designate natural kinds from the point of view of the natural sciences." Brink's own naturalistic view involves only a constitution claim, rather than an identity claim.

from within an engaged ethical viewpoint of the right sort, where "the right sort" cannot be reductively cashed out in a naturalistic way. This is in part a non-naturalistic *methodological* and *epistemological* claim: empirical inquiry as such (including empirical inquiry into the semantics of ethical discourse) is incapable of settling the answers to ethical questions. More importantly, however, the above claim—particularly that "the right sort" of ethical viewpoint cannot be reductively cashed out in a naturalistic way—implies a non-naturalistic *metaphysical* commitment with respect to certain facts. According to the robust realist as so far characterized, it is an objective, irreducibly evaluative fact that:

F Some forms of ethical upbringing, or sets of sensibilities, are *better than* others, constituting the *right starting points* for ethically accurate deliberation.

But what natural, empirical fact(s) could a fact like *this* consist in? We might expand on F by adding that what makes one upbringing superior to others is that it fosters the responses that track the values determinative of ethical facts, such that through this grasp in conjunction with sound ethical reasoning one is led to correct ethical judgments. But this is of no help to naturalism unless there is in turn an independent, naturalistic way of picking out those values and ethical facts, or of specifying the correct ethical judgments—for example, correct ethical judgments are those that conform to natural teleological norms, or to standards concerning the promotion of survival, longevity, general satisfaction, etc. Yet that is just what the robust realist has found reason to doubt, as discussed earlier: demarcating the ethical facts or correct ethical judgments seems to him already to require appeal to the idea of a *properly informed ethical standpoint* (one embodying the right sensibilities as substantive starting points, as well as empirical information), to which any relevant natural facts have to be submitted as input for engaged reflection before we get any objectively ethical results amounting to authoritative conclusions about what the ethical facts are or which ethical judgments are correct. We therefore seem to have an irreducibly evaluative circle here, which resists naturalistic inroads on either side: F seems to be a non-natural fact, as do corresponding facts about what genuine values there are, what the genuine ethical facts are, or which ethical judgments are in fact correct; and attempts to break into this circle from the outside just lead to the kinds of problems examined earlier. This circle is not a problem, however, and does not tell against the truth of F: all it implies is that F resists explication in naturalistic terms. It is a non-natural fact.[63]

[63] Non-natural facts are facts that cannot be cashed out in empirical terms, as by appeal to facts of psychology or biology, or to complex facts constructed entirely from

Similarly, consider another kind of fact I have argued the robust realist is committed to:

G The moral wrongness of an act is a property that does not merely tend to *cause* a feeling of disapproval in us but *merits* such a reaction, along with the act's dismissal from deliberative consideration.[64]

What kind of natural, empirical fact could this be? If we accepted one of the stance-dependent naturalistic views considered earlier, we could understand it along some such lines as this: moral wrongness is a natural property that merits such practical reactions from us insofar as we *would* disapprove of or deliberatively dismiss actions with that property if we were fully informed, instrumentally rational, etc. But if we have resisted such views, what is left among objective empirical facts that could possibly capture a fact such as G? Certainly the natural facts we cite in explaining why the act is wrong (facts about the natural features by virtue of which the act is wrong) do not as such constitute a fact about *meriting* a certain practical reaction, which is a further normative fact *about* those facts—that is, the normative metafact that such facts about an action make it merit such practical reactions. So commitment to facts such as G constitutes a theoretical commitment to non-natural facts.[65]

We can actually extend the above points to ordinary ethical facts themselves—facts about something's being good, or right, or a reason for acting—and even to ethical properties, all of which turn out to be non-natural in an important sense given the other commitments of robust realism, or so I shall argue. This may seem puzzling, since ordinary ethical facts and properties obviously stand in intimate relations to the natural facts and properties *by virtue of* which they obtain or are ascribed. Aren't ethical facts and properties obviously just *constituted by* those natural facts and properties, which is what makes sense of supervenience? And if they are so constituted by natural facts and properties, doesn't that clearly give us a naturalistic metaphysics of ethical facts and properties? So there is a puzzle about how ordinary ethical facts and properties could be non-natural. I turn now to resolving this puzzle.

such facts (as illustrated by the various naturalistic views considered earlier). In Section 10 I will qualify talk of nature or the natural vs. the non-natural, and note a way in which this can be misleading. This will clarify what I mean by insisting that facts such as F are non-natural.

[64] Alternatively, G could be taken to be the *metafact that* the fact that the action has the property of wrongness makes it the case that it should be dismissed from deliberative consideration, etc. See Dancy (2006: 137).

[65] This point will come up again later.

8. EVALUATION AND STANDARDS

We may begin by considering the non-ethical case of attributing an evaluative property to an artifact, which is comparatively simple. My computer counts as a *good* computer *by virtue of* its possession of certain natural features: it boots up quickly, rarely crashes, runs at a high speed, has a large memory, and so on. Let 'XYZ' stand for this set of these natural features we cite as reasons for attributing goodness to the computer—the features that *make* it good. The computer's goodness, we may say (following Dancy), is a resultant property, and XYZ is the resultance base.[66] But what exactly is the relation between the resultance base and the resultant property?

A natural thing to say is that the computer's goodness just *consists in* these natural features XYZ. Once we've fully enumerated the features in XYZ, what more could be left to add in order to capture the goodness? How could it be anything *over and above* XYZ? But while there is surely something correct in this thought, there is also something misleading about it. This comes out if we think about the *fact* that the computer is good. What does this fact consist in? Does it consist simply in the fact that it has XYZ? One clear indication that it does not is that any number of things could have XYZ without being good. This is generally true for good-making properties of artifacts: sharpness makes for goodness in a knife, for example, but not in a pair of glasses. This is important not because this point itself translates directly to the ethical case—it does not, as at least some ethically good- or bad-making properties may plausibly have that status invariantly (e.g. certain character traits will be bad-making in any rational being)—but because it directs us to the more complex structure of facts about goodness, which does have an analog in ethics. In the present case, the point is that the fact that this computer is good consists not only in the fact that it has XYZ, but in this *together with* the fact that XYZ is such as to satisfy the *standards of goodness* S for computers.[67]

Though the point here is not merely an epistemic one but a claim about the metaphysics of facts, it helps to notice that if we are introduced to an unfamiliar artifact, we can discover that it has a certain set of features

[66] See Dancy (1993: 73–7 and 2004).

[67] It is even more clear that what it is for the computer to be *good because of XYZ* is not merely for it to have XYZ, but for it to have XYZ *and* for XYZ to satisfy the standards of goodness for a computer. But since the computer *is* in fact good, if it is good at all, precisely *because* of XYZ, the same point seems to extend to the fact *that it is good*, as claimed in the text. Again, recall that other things of different kinds can have XYZ without being good.

without thereby discovering whether it is good or bad; and this is because the fact of its being good or bad consists not just in its having those features, but in that together with facts about how those features relate to the appropriate standards of goodness for such a thing. To put it another way: even with respect to artifacts, the act of *evaluating* something is not just the act of attributing certain natural properties to it, and the content of an evaluation is not just the content of an ascription of natural features.

Now I am not proposing to defend non-naturalism about the goodness of computers. The point is just that the evaluative fact is not just the fact of possessing the features that make up the resultance base. This doesn't cause any problems for naturalism here, though, because we can plausibly give a naturalistic reduction of facts about standards of goodness for computers in terms of their proper functions, understood in terms of such things as the intentions of designers, social conventions, and so on (though this will almost certainly prove more difficult than it may initially appear to be). If this story appeals exclusively to natural properties and facts, then there is no threat to a naturalistic understanding of evaluative properties and facts concerning artifacts.

Moving to the moral case, however, things get more complicated. Consider the judgment that Claggart's behavior toward Budd was morally bad. This badness, again, is a resultant property: Claggart's behavior was bad *by virtue of* the fact that he lied about Budd's loyalty to Vere and falsely accused him of mutiny, knowing the devastating effect this would have on Budd and doing it for that very reason, out of jealous loathing of him. Shall we then say that the badness of Claggart's behavior just *consists in* those natural features in the resultance base? Again, in a sense this seems right, since it is not as if there is anything else *of the same sort* that we have to add in order to answer the question: 'by virtue of what was the behavior bad?' Once we have cited the relevant natural features making up the resultance base, we are done with *that particular line of explanation*.

As we have seen, however, that's not the whole story, and it is misleading just to say that the badness consists in those natural properties, as if *evaluating* the behavior were just the same thing as attributing those natural features to it (which is precisely what makes some forms of naturalism seem to turn a variety of moral judgments into trivial or redundant claims[68]). The fact that the behavior is bad consists not merely in the fact that it exhibits those natural properties as such, any more than in the case of the computer's goodness and XYZ. It consists rather in that fact together with the facts that there are appropriate standards of goodness S for human

[68] For a good discussion of Parfit's triviality objection, see Dancy (2006: 131–2).

behavior and that actions that exhibit the features in the resultance base in question—the deception and intent to hurt, driven by jealousy and loathing, etc.—violate those standards. This is what we are concerned with in evaluative judgment.[69]

As before, then, the next question is how to understand these other facts, and this brings us back to the commitments of robust realism as I understand it. Even Mackie allowed that there can be evaluative truths relative to conventional standards (e.g. at dog shows),[70] and others might propose standards rooted in something like natural teleology, or non-conventional standards rooted in facts about deliberatively modified desires or attitudes given empirical information, or in facts about basic social needs. But we have seen that there are significant reasons for resisting such accounts of the appropriate standards of goodness for human action. Robust realists, in any case, will insist on *objective standards that are appropriate for rational agents, having a legitimate claim on our attention and practical response*. But I argued earlier that (given certain plausible motivating assumptions) this idea of correct, objective standards of goodness that can simultaneously be authoritative for us as agents is really just the idea of the *standards bound up with proper ethical sensibilities*, where the fact that a certain set of sensibilities counts as a proper one is a *non-natural*

[69] It might seem that we could avoid this by making use of 'thick' concepts as an intermediary. For example, we might say that the fact that the behavior is bad just consists in its being *cruel*, and that the cruelty just consists in the various natural features in the resultance base. But this doesn't really make things any easier. On the one hand, suppose we exploit the rich descriptive aspect of 'cruel' that can be recognized by any linguistically competent speaker in central cases. Then we can say directly that the cruelty consists simply in the various natural features associated with that descriptive aspect, but we've accomplished nothing more than supplying a different descriptive term to capture those features; the earlier point just arises again in connection with that term: if we evaluate the action as *bad* by virtue of its cruelty, then this is again to appeal not just to the fact that it is cruel, but also to the fact (however obvious it may seem) that cruelty violates the appropriate standards of goodness for actions. On the other hand, suppose we use a morally loaded notion of cruelty, taking 'cruelty' to function as the name of a vice (such that it is no longer a term that just any linguistically competent person can reliably employ, since at least hard cases will require substantive moral understanding in order to distinguish genuine cruelty from things that might resemble it, such as hurting someone's feelings in cases where there turn out to be good reasons for it). This also doesn't help, because although now the connection between the badness and the cruelty is automatically secured, we now also need to appeal to the fact that the various features in the resultance base are *vice-making*, which brings us back to the original claim about violating the relevant standards of goodness. All of this suggests that Shafer-Landau's (2003: 74–6) arguments appealing to the example of a person's *generosity* move a bit too quickly, and that the exhaustive constitution view (discussed further below) misses something important.

[70] Mackie (1977: ch. 1, §5).

fact, resisting cashing out in either subjective or objective naturalistic terms. That was F and the evaluative circle from Section 6. So if F—a fact about the existence of proper sensibilities—is a non-natural fact, then so too is the related fact that the set of ethical *standards* S bound up with that proper set of sensibilities is irreducibly and objectively correct.

It follows from this that for the robust realist ordinary ethical claims, such as the claim that Claggart's behavior was bad, involve implicit appeals to non-natural facts. Recall that the fact that his behavior is bad does not consist merely in its possessing the natural features we normally cite when explaining what makes it bad: It consists in this *together with* the facts that there is some correct set of standards of ethical goodness S for human beings, and that his behavior violates S.[71] But we have just seen that, given the line of argument so far, the fact that S constitutes a correct set of standards of ethical goodness for human beings is a non-natural normative fact. Therefore, the fact that Claggart's behavior was bad requires for its full explication an appeal to a non-natural normative fact. And that is to say that the fact that Claggart's behavior was bad is not itself merely a natural fact, but has a non-natural element. Indeed, I take this to be crucial to capturing the idea that when we call his behavior bad we are *evaluating* it—not merely describing it again—and doing so according to *objective* standards of the sort that could be *authoritative* for rational agents.[72]

[71] We can also describe the latter two facts as the non-natural fact that these natural features are morally *bad-making* qualities, where such badness is such as to *merit* our disapproval and avoidance.

[72] Copp (1995) has developed a rigorous and interesting *naturalistic* and largely *objective* standard-based account of morality that is more promising than any based on natural teleology, for example. According to Copp, a set of moral standards is justified relative to a society (roughly) if it would be instrumentally rational for the society to adopt it, by virtue of its tending to serve (independently specified) basic social needs better than other codes. On such an account, the fact that a set of standards is justified is a natural fact, reducible to instrumental facts about how well social codes meet basic social needs. Much of what I have said above—particularly in Sections 4 and 6—indicates why I find such an approach problematic, though it deserves a much more careful treatment than I can give it here, where just a few remarks will have to suffice. One worry is that any such reductive, instrumental approach will leave something important out of the account of important moral values, making it seem deflationary in an important way. To take one of his own examples, the wrongness of slavery on this view turns out to be a relational matter pertaining to how a code permitting slavery fails best to serve a given society's basic needs overall, where these include, for example, the need to "ensure the continued existence of the population that it is," a "need for peaceful and cooperative relations with neighboring societies," and the need "to ensure that at least a sufficient number of its members are able to meet their basic needs" (1995: 219, 193). Not only is it questionable whether carefully selective oppression must thwart

9. SUPERVENIENCE WITHOUT EXHAUSTIVE CONSTITUTION

Given robust realist commitments, then, we cannot say—as Shafer-Landau is willing to concede to ethical naturalists—that ethical facts are ''exhaustively constituted'' by natural facts.[73] Not only are there some irreducibly evaluative or normative facts such as F and G that are non-natural and not constituted by natural facts, but even ordinary ethical facts are non-natural, that is, at least not completely constituted by natural facts. This is because while they are partly constituted by the natural facts in the resultance base, they are *also* partly constituted by non-natural facts such as F and G, about the appropriate standards of goodness, or equivalently about certain natural features being *good- or bad-making* where such goodness or badness is such as to *merit* our approval or disapproval—which again are not themselves constituted by natural facts. Robust realists must therefore reject the naturalistic exhaustive constitution thesis and instead adopt a non-naturalistic metaphysical view of ethical facts.

Now if the *fact* that Claggart's behavior is bad is ultimately a non-natural fact (though partly constituted by natural facts), then what are we to say about the *property* of badness? Can we say that the badness simply consists in the various natural features that make up the resultance base, or that the latter exhaustively constitute the former? It is hard to see how this could be anything but misleading, again covering over a deep difference between a robust ethical realist and a non-reductive ethical naturalist. If the *fact* that the behavior is bad is a non-natural fact, then it would be odd to say that the behavior's *property* of badness is a purely natural one, as suggested by the exhaustive constitution claim.

It is not that we need to add something spooky to the natural features in the resultance base, standing right alongside them—as if to say that Claggart's behavior was bad by virtue of being dishonest, nasty, cruel, and Z, where Z is some mysterious non-natural property thrown in for good

such basic social needs, but even if it does it is not clear that we should be looking to such instrumental facts in order to understand the moral outrage of slavery. This seems better captured by a direct appeal to the violation of the dignity of persons, where this is taken to be a rich moral concept that cannot be cashed out instrumentally in relation to naturalistically construed basic societal needs, and the normative force of which is not adequately captured by his remarks about categoricity (1995: 224–6). Again, however, these brief remarks are not meant to constitute a satisfactory critique, but only to indicate why a robust realist will likely not be happy with such an instrumentalist and reductive approach to morality.

[73] Shafer-Landau (2003: 74–9).

measure. The natural features are indeed *complete* when it comes to the 'by virtue of' explanation giving the reasons why the action was bad, in the ordinary sense. Nor is it being suggested that badness is non-natural in any sense implying independence from the natural features in its resultance base. Surely the badness is *partly constituted* by the natural features in the resultance base, and similarly with other ethical properties, in such a way that the relevant natural features and facts *fix the ethical properties and facts*. Indeed, as just noted, in *this* respect—the subject of the 'by virtue of' story—the partly constituting natural features and facts *are* exhaustive and complete: they leave nothing out of that particular story.[74]

These points of partial concession to naturalism are in fact precisely what captures and explains the supervenience of ethical facts and properties on natural ones.[75] This is because partial natural constitution in the above sense is sufficient to explain this supervenience, and this is not affected by the claim that there is also a crucial non-natural element to ethical facts and properties of the sort I have described. Consider, for example, that even in the case of the computer the fact of goodness is not exhaustively constituted by the possession of XYZ as such, but also involves further facts about standards and relations to them, as argued above; yet this does not in any way threaten the supervenience of goodness for artifacts on the class of natural properties to which the XYZ properties belong, because an artifact still *meets* or *fails to meet* the relevant standards *simply by virtue of its natural properties*, so that two objects cannot differ in their evaluative properties without differing in their natural properties. Now I have claimed that for the robust realist the further facts about standards in the case of ethics are non-natural, unlike in the case of artifacts. But that does not in any way affect the present point about supervenience. Whatever the particular details of the facts about the relevant standards, including the metaphysical status of those facts (i.e. whether they are natural or non-natural), we still have the very same structure of individuals meeting or failing to meet those standards by virtue of the natural features they possess, and that is the structure that explains supervenience—in ethics no less than for artifacts. The non-naturalism I have introduced does not threaten supervenience.[76]

[74] This answers one objection raised by Jackson (1998: 127–8), discussed further below.

[75] Cf. Dancy (1993: 79): "That the moral properties supervene upon the natural ones is entailed by the idea that a moral property cannot exist on its own, but must result from some other (probably natural) properties."

[76] This is meant to address the problem Shafer-Landau (2003: 78) raises for any denial of exhaustive constitution. Note that I have so far been addressing intra-world

To return, then, to the metaphysical status of ethical properties: If the *fact* that an act is bad is a non-natural fact about it, and therefore the property it has of *being bad* is likewise non-natural for the same reason, then it is hard to see how we can say that the *badness* itself is a property exhaustively constituted by natural properties, and so is a natural property itself. If it weren't for the non-natural aspect of the fact that the behavior is bad, that fact wouldn't exist as it is (there would be only a watered down substitute), and so neither would the property of badness exist as it is, that is, *as a property with the distinctive evaluative and normative nature and force that it has*. This would be missed by saying that the badness just consists in the natural properties. The badness is more than that, though again not by virtue of something merely tacked on alongside the other natural properties in the resultance base.

It is therefore a mistake to be led by the *completeness* of the resultance base in answering the 'by virtue of' question to the conclusion that the natural features in the resultance base *exhaustively constitute* the resultant property in the sense that suggests a naturalistic *metaphysical* view of ethical properties. The former completeness is compatible with a non-naturalistic metaphysics of ethical properties, which goes hand in hand with the non-naturalistic metaphysics of ethical facts to which robust realism seems to be committed.[77] In my view, then, Shafer-Landau should therefore take a decisive step further away from ethical naturalism, distancing himself from it not merely on methodological and epistemic grounds, but also on metaphysical grounds. While he may or may not accept all of what I've built in to robust ethical realism as I understand it, he does clearly endorse central elements of it: in particular, the stance-independence of moral facts

supervenience. I will discuss global supervenience in the next section. For a lucid defense of realism against Blackburn's objections involving supervenience, see Klagge (1988). I take my account here to be compatible with Klagge's approach, and to help explain how a realist can make sense of ascriptive as well as ontological and descriptive supervenience.

[77] The view I've sketched would count as non-naturalistic by any criterion I am aware of. For example, Brink (1989: 159) characterizes ethical naturalism as the view that "moral facts and properties ... are constituted, composed, or realized by organized combinations of natural and social scientific facts and properties. The former are, then, in a certain sense nothing over and above the latter." Since the robust realist view rejects this last claim, it clearly counts as non-naturalist. Brink does later say that there is also a sense in which "constituted facts are (after all) something over and above their bases" (1989: 193–4), but by this he means only that on the constitution view (as opposed to the identity view) we can distinguish between structural and compositional *aspects* of the facts in question (1989: 196), and recognize that the causal or explanatory power of facts often depends on their structural aspect. This is still thoroughly naturalistic, though non-reductive, and is not the same sense in which the robust realist claims that ethical facts are something "over and above" their bases.

and standards, and the intrinsic normativity of morality.[78] This is enough to make problems for an exhaustive constitution view of ethical facts and properties, as we have seen.

Even apart from the complicated story above involving F and its background role in ordinary ethical claims, we can see that there is little hope of combining the stance-independence thesis and the intrinsic normativity of morality within an exhaustive constitution metaphysical framework. What purely natural, empirical facts could exhaustively constitute *the fact that* an action's being bad *merits* my refraining from giving the considerations in its favor weight in deliberation, or constitutes a *reason* for me not to do it? (Cf. G above.) I have given reasons for rejecting stance-dependent approaches, but at least such an approach could offer a ballpark candidate for such exhaustively constituting natural facts; for example, the fact that an action's being bad merits my dismissing it from deliberative consideration might be claimed to be exhaustively constituted by the fact that my ideally rational self would want or advise my actual self to dismiss bad actions from deliberative consideration. But Shafer-Landau has (rightly, I believe) rejected such views as part of his more robust realism. And if such views are unavailable, it is hard to see what *other* kinds of natural empirical fact could even begin to fit the bill for exhaustively constituting such a normative fact. The objective, natural facts in the resultance base *make* the action bad (it is bad by virtue of them), and in doing so they *make* it merit deliberative dismissal. But this doesn't solve the problem. Our question was not to identify natural facts that make the action bad or even that make it merit a certain treatment in deliberation, but rather to identify natural facts that could be said to exhaustively constitute *the fact that* certain natural facts make the action bad and thereby make it merit a certain treatment in deliberation. And that fact is not itself just the original natural facts in the resultance base.

Compare: The empirical fact, E, *that smoking causes cancer*, is a *reason* to quit smoking. Now consider the following metafact, M: *The fact that* the fact that smoking causes cancer is a reason to quit (or equivalently: *the fact that* smoking's causing cancer is a reason to quit). What kind of fact is M? It is not merely E, the fact that smoking causes cancer. It is a normative fact about E. We therefore need an account of it, and stance-dependent naturalistic views, such as Williams's account of reasons, offer such accounts. But if we reject such accounts, as the robust realist does, then what is left in the way of naturalistic resources for understanding (meta)facts such as M? I am claiming that there is nothing plausible to look to here for a story of natural exhaustive constitution, which is precisely

[78] Shafer-Landau (2003: 15, 190, 205, 211).

why we need to reject naturalism and accept that (meta)facts such as M commit us to a non-naturalist metaphysics, according to which there are some facts that are not exhaustively constituted by natural facts.[79] My claim, then, is that given his commitments to at least the central tenets of robust realism, Shafer-Landau should part company even with non-reductive naturalists such as Brink when it comes to the metaphysics of ethical facts and properties. And the same is true for anyone else who shares the motivations and concerns I have tried to show push in this direction.[80]

10. A DUAL ASPECT VIEW OF EMPIRICAL AND ETHICAL REALITY

In fleshing out robust ethical realism, I have appealed to the idea of objective ethical standards reflecting a structure of objective values that resists explication in naturalistic terms. This raises the question: Where, then, do these values and related standards have their source? If, on the one hand, we are to avoid supernaturalism (the idea that they have their source in God) or even transcendental ethical realism (such as an appeal to Platonic Forms), as I propose to do, and yet on the other hand we have also rejected naturalism, then it might seem utterly mysterious how we are conceiving of and locating these real values and objective standards. If they do not spring either from nature or from the supernatural or the transcendent, where *do* they come from?[81]

The problem here lies in an ambiguity in talk of nature or of natural properties and facts. I have resisted the idea that ethical standards can come from nature *as empirically investigated*, or that they can be cashed out in terms of natural properties and facts *as grasped by the sciences*. But according

[79] My points here are closely related to those made by Parfit (forthcoming), although he does not understand his non-naturalism about normative properties to involve any positive metaphysical commitment.

[80] Thanks to Jamie Dreier for pressing me to show why it is not open to Shafer-Landau simply to reassert the exhaustive constitution claim involving the original natural properties in the resultance base to explicate the (meta)facts I have claimed resist such moves.

[81] Shafer-Landau (2003: 47–8, 96–7) rejects any demand for an explanation of the source of the correct moral standards: they are just true and there is nothing that *makes* them true. As I will bring out below, however, I think this cuts off explanation prematurely. While there are, on my view, brute facts about value that are not further explicable, there will be a rich story to be told about how these values ground the ethical standards that in turn partly explain particular ethical facts. The standards are not themselves generally basic.

to the kind of robust ethical realism toward which I am inclined (the development of which I can only sketch here), the objective values that are determinative of correct ethical standards are nothing other than objective, irreducibly evaluative or normative aspects of *this same world*, though they are not visible as such from the point of view of empirical inquiry. The property of being painful, for example, is something we ordinarily consider a natural property capable of empirical investigation. And so it is. But I take it *also* to be an evaluative and normative fact that pain is typically bad and to be avoided or mitigated, though this fact and property do not show up in the psychology or neurology lab. On the picture I have sketched, these facts and their normative implications cannot be properly grasped and fleshed out except within properly informed ethical reflection. And part of what we are doing in such reflection is placing such facts within a comprehensive structure of values and norms, which may be seen as an articulation of ethical standards—standards for human good, for right action, for what is to count as a reason, and so on. The *badness of suffering* informs correct ethical standards, for example, by shaping those standards to prohibit callous and cruel behavior, as fleshed out in substantive ethical reflection. Similarly, the property an utterance may have of being deceptive is a natural property capable of empirical investigation, but it is also something we can recognize in ethical reflection to be inherently problematic. And this informs our conception of what it is to live well—our specification of the standards of ethical goodness.

The metaphysical claim, then, is that many familiar facts and features of human life, behavior, and experience, which can be the subject of empirical investigation, are also *inherently value laden*, and as such are the source of objective standards of goodness for us—though these standards can be properly grasped and specified only through correct ethical reflection on these facts and features as we live, experience and reflect on human life and the best possibilities it has to offer. This means that the very features we have been calling 'natural' all along—features in the resultance base of ethical properties and facts, such as something's being painful or deceptive—were never *merely* natural to begin with. Indeed, it is just the mistake of thinking of them that way that creates the air of mystery that seems to surround non-naturalism: how can we possibly start with a bunch of value-free, natural properties and then somehow have them give rise to a resultant property that has a genuine and non-natural evaluative or normative nature? How can that gap possibly be bridged? The answer is that there was never a real metaphysical gap to begin with. The natural features in the resultance base are value laden to begin with, this value contributing both to the proper standards of ethics and to the evaluative nature of the resultant property or fact—though how this happens is a matter of how the base features taken together relate to the overall standards, which will be shaped by many other features besides the base features in a particular case.

To call the various base features *natural* is therefore just to focus on them *insofar as they are capable of being investigated and grasped empirically*, screening off any evaluative or normative aspect they may have. It is not to say that they are in themselves value-free, which would indeed make it hard to make sense of the resultance of robust ethical properties and facts from natural ones. I am therefore proposing a kind of *dual aspect* view of the world of human experience: the very properties and facts we typically refer to as natural are also inherently value laden, and their value laden aspect is the source of objective ethical standards. Talk of 'natural' properties and facts in connection with resultance base features thus turns out really to be just a way of referring to the empirical dimension of an essentially value laden world. And to reject naturalism is just to reject the idea that we can account for ethical standards and facts simply by appeal to that empirical dimension of reality as such, as by appealing to the actual or hypothetical results of some ethically-neutrally specifiable set of conditions or procedures applied to empirical facts about satisfaction, pleasure, desires, social needs, etc. It is not, then, to deny that objective ethical standards are rooted in familiar facts and features of human life, but only to insist that *they are rooted in the irreducibly evaluative dimension of these facts and features of human life, grasped through ethical reflection by people suitably acquainted with those values along with relevant empirical facts*. Such an approach can therefore remain broadly Aristotelian (as long as problematic forms of naturalism are avoided, as McDowell stresses) as opposed to Platonic or supernaturalist, and so remains, I believe, attractively moderate despite its embrace of non-naturalism.[82]

[82] Patricia Greenspan has pointed out that there are actually a variety of ways in which a robust realism could be developed here. One way, which I favor, is liberal about the kinds of fundamental, irreducibly evaluative or normative properties there can be as part of the value laden-ness of the relevant parts of reality. I am happy to include the (pro tanto) badness of pain, goodness of love, inappropriateness of deception, inviolability of persons, and value of persons as ends, among others. Others, however, limit the fundamental value or normative laden-ness to the category of reasons (e.g. Parfit), or to facts about human flourishing, deriving all the rest from them, and so might be 'less robust' in certain respects than what I have proposed. I myself follow McDowell in rejecting the idea of deriving moral facts from pre-moral facts about flourishing, for example, and I am also skeptical about the prospects for deriving evaluative facts from facts about reasons, since I think the latter presuppose the former. For a different view, defending normative realism without value realism, see Birondo (2006).

11. DUAL-ASPECT NON-NATURALISM, GLOBAL SUPERVENIENCE, AND THE ROLE OF STANDARDS

This dual aspect view of the facts and features of human life also raises its own questions, however. I have said, for example, that the non-naturalist, standard-based view I have defended does not undermine supervenience or make it mysterious: as long as individuals meet or fail to meet the standards that apply to them simply by virtue of their natural properties, it makes no difference what the metaphysical status of those standards is; once the standards are in place, supervenience will be preserved, just as it is for artifacts. But now I have said that the standards are themselves just a function of a structure of values inherent in the same single reality. One question this raises is whether I have just shifted the mystery by appealing to a *fundamental association* of certain value features with certain natural features. For example, I said above that the empirical property of being painful is also value laden, pain being something bad and typically to be avoided or mitigated. I take this to be a fundamental metaphysical association: it's just part of what pain is that it has this value laden aspect. Similarly with deception. So if these sorts of connections are among those that people find mysterious when they speak of supervenience, then it is true that I have done nothing to lessen that mystery. My view is that certain elements of the world just are value laden in this way, as a basic metaphysical fact about them, and that there may not be anything more for philosophy to say here.

It is important, however, not to overstate the point. While I bite the bullet about fundamental metaphysical facts of value laden-ness at this basic level, such as the fact of the badness of pain or the fact of the value of persons (e.g. the inviolability of rational agents), this is not to declare supervenience in general mysterious. While I may not be able to give any deep explanation of the value laden-ness of pain, I have not left it mysterious why supervenience holds in connection with our derivative judgments about actions being wrong due to their being cruel, for example, such that it is a constraint that actions with the same natural profile must be judged similarly. This is not mysterious because again it is simply an implication of the standard-based structure of evaluation that I have advocated for complex cases, just as it is in the case of artifacts. There is no pervasive mystery, then, but at most some mystery at the level of certain fundamental connections pertaining to the value laden-ness of certain elements of reality, from which the complex standards relevant to the vast majority of our ethical judgments are derived, in turn explaining all the rest of supervenience with which we are usually concerned.

Indeed, while I have so far spoken explicitly only of *intra*-world supervenience (that is, given a single set of standards within a world, derived from the evaluative dimension of relevant facts and features of that world, any two actions that are identical with respect to their natural properties will be identical with respect to their ethical properties, since they will similarly satisfy or violate those standards), the view I have proposed equally entails *global* supervenience, described thus by Jackson:

(S) For all w and w*, if w and w* are exactly alike descriptively, then they are exactly alike ethically.[83]

This is because if w and w* are exactly alike with respect to their natural properties and facts (i.e. "descriptively"), and if, as I have proposed, the standards of ethics are derived from the evaluative dimension of some of those natural properties and facts, then the same ethical standards will obtain in w and w* and they will be alike ethically. Or at least this follows on the assumption that the fundamental associations I have posited between the natural and the evaluative within the value laden dual-aspect reality are metaphysically necessary. I believe that to be true, and so my view entails that "the descriptive nature of complete ways things might be settles ethical nature."[84]

This, however, leads to another set of questions. If my view preserves global supervenience in this way, does it do so at the cost of undermining the very appeal to standards that I used to argue for non-naturalism in the first place, suggesting that they are superfluous after all?[85] I have said that the ethical standards are derived from the set of values inherent in a subset of the properties and facts we also identify as natural ones, and this means that, once the natural properties and facts are fixed, so are the ethical ones. Why, then, even bring in the detour through standards at all? They seem to have been short-circuited, so that the story can go directly from natural properties and facts to ethical ones, even across possible worlds: since the same standards will exist in all possible worlds with the same natural properties, they seem just to drop out as players in determining the ethical facts. Indeed, if ethical properties are *necessarily coextensive* with natural ones in the way I've suggested—consistently with Jackson's logical equivalences between ethical properties and "possibly infinitely disjunctive descriptive properties"—doesn't that imply that they are just identical?[86] If so, naturalism is vindicated after all.

[83] Jackson (1998: 118 ff.)

[84] Ibid.

[85] Thanks to Tom Baldwin, David Brink, Terence Cuneo, and Jim Klagge for pressing versions of this objection.

[86] Jackson (1998: 124).

Let me start with this last point, which represents Jackson's view. Such a move relies on the metaphysical assumption that *necessarily coextensive properties are identical*, and in particular on the assumption that if we can in principle make every distinction using descriptive vocabulary that can be made using ethical vocabulary, then ethical properties *are* natural properties. I follow Shafer-Landau in rejecting the "coextension test of property identity," and find his and others' objections to it more compelling than Jackson's defense of it.[87] But even apart from those objections, much of what I have attempted to do in this paper is to provide principled reasons for regarding ethical facts and properties as interestingly non-natural *despite* their obviously intimate relations to natural facts and properties, which Jackson has fleshed out in the most abstract terms using possible worlds metaphysics. So even if Jackson's response to common counterexamples were persuasive—for example, that the apparent distinction between being equilateral and being equiangular is really just a distinction in modes of representation, not in the properties themselves—this would not undermine the case I have made for the distinctness of ethical and natural properties unless the earlier motivations for non-naturalism can all be revealed in fact to be about nothing more than modes of representation. I have tried to show, however, that much more is at stake concerning the metaphysics of ethical properties and facts: the points were *not* just about our modes of representing such properties and facts to ourselves, but about the need (given robust realist motivations and concerns) to posit properties and facts or metafacts that cannot be captured within a naturalist framework. Those arguments are, of course, open to various possible objections. But they cannot be neutralized simply by an exercise in possible worlds metaphysics bringing out general extensional equivalences of the sort Jackson focuses on. And since my earlier arguments are apparently consistent with the view I have proposed for the source of ethical standards, and so with the necessary coextension of ethical and natural properties in Jackson's sense, they provide substantial grounds for rejecting Jackson's insistence on the identity of necessarily coextensive properties.

Jackson does offer some positive reasons for taking necessarily coextensive properties to be identical, but these are unconvincing. One is that "it is hard to see how we could ever be justified in interpreting a language user's use of, say, 'right' as picking out a property distinct from that which the relevant purely descriptive predicates pick out, for we know that the complete story about how and when the language user produces the word 'right' can

[87] Shafer-Landau (2003: 90–5). See also below. Thanks to David Enoch for encouraging me to make this commitment explicit.

be given descriptively."[88] But it is not the robust realist's claim that we discover the non-natural referents of ethical terms by scrutinizing language users' ethical judgments. We discover this instead through philosophical reflection on the apparent inadequacy of naturalistic accounts, exactly as I have tried to do here.

Another objection Jackson offers is that the non-naturalist seems committed to taking seriously "someone who says, 'I see that this action will kill many and save no one, but that is not enough to justify my not doing it; what really matters is that the action has an extra property that only ethical terms are suited to pick out'."[89] Again, this is a caricature of the position. As I have emphasized, the claim is *not* that there is some extra non-natural property injected right alongside the natural resultance base properties as part of the 'by virtue of' justificatory story. That story is already complete, just as Jackson insists. The claim, however, is that the way that those base properties (e.g. being destructive of life while saving no one) result in a moral property is by making the action stand in a certain relation to standards of goodness for human action (e.g. violating them), where the facts about those standards are not merely natural or empirical ones, which means that neither is the resultant ethical property or the fact that the action has it. That is the non-naturalism I have defended, and it does not commit the non-naturalist to taking seriously any bizarre claims about ethical justification, as in Jackson's example.

Finally, Jackson wonders what principled basis there could be for determining when such "duplication" or "twinning" takes place, assuming that not every natural property has a necessarily coextensive non-natural shadow property: "what is special about the descriptive properties that have twins from those that do not?" and why think that our ethical vocabulary coincides with this?[90] The answer I have suggested is that the "twinning" is just a matter of some natural properties and facts being value laden, and it is hardly a great mystery why we might take something like the property of being painful or deceptive to be value laden, while denying that the property of being flat or larger than the sun is not. While there is no general formula for determining which aspects of nature are value laden in the way I have proposed, a plausible start is to suggest that this is limited to the sphere of sentient and rational experience, capacities, interactions, and so on. This is still too broad, of course, but reflection on our own substantive experience of value can guide us here in ways that are no more arbitrary than our use of ethical language itself is. And if ethical language has been developed and refined as a way of thinking and speaking evaluatively and normatively

[88] Jackson (1998: 127–8). [89] Ibid. [90] Ibid.

about the value laden dimension of our world, and of its implications for how to live, then it is equally unsurprising that our ethical language should coincide with the natural properties for which "twinning" occurs.

Let us return now to my appeal to the role of standards in the determination of ethical facts, and to the worry that they drop out of the picture. While it may be true that the natural facts *fix* the ethical facts—that "the descriptive nature of complete ways things might be settles ethical nature," so that Jackson's equivalences hold—this does not by itself provide a *metaphysical explanation* of how ethical facts are determined by the natural facts, unless one just begs the question in favor of naturalism. Part of my argument has been that a proper metaphysical explanation of ethical facts, and of how they are determined by natural facts, involves appeal to the notion of something's measuring up to, or failing to measure up to, an appropriate set of standards of goodness for things of that kind. This is not just about *how we know* whether something is good or bad, or how we represent that fact to ourselves: it is about *what it is* for something to be good or bad, which should already be familiar from the case of artifacts.[91] This metaphysical claim is not voided simply by establishing broad entailments between natural and ethical properties and facts. If all we were interested in were the *fixing* of ethical facts by natural ones, of the sort revealed by Jackson's entailments, then the appeal to standards might be superfluous. But this is not all we are interested in. Far more interesting is the question *how* ethical facts are determined by natural ones, and I have argued that standards play a crucial role here. The fact that the same standards show up across possible worlds with the same natural properties does not undermine their structural role in the metaphysical explanation of *what it is for something to be good or bad*. They do not, then, merely drop out of the picture.[92]

It is also important to remember that although on my view the standards themselves have their source in the same set of facts and properties that we also regard as natural ones, it is not in their empirical aspect that such facts and properties determine ethical standards, but in their irreducibly evaluative or normative aspect, as grasped only through ethical reflection proceeding from adequate starting points in ethical experience and character. Our metaphysical account of the standards cannot, then, short-circuit the appeal to such standards once we have the natural facts and properties *empirically considered* (i.e. by the sciences), as if the standards

[91] Thanks to Mark Schroeder for helping me to emphasize that my point here is about metaphysical explanation.

[92] Cf. Dancy's (2004) related complaint that Jackson's naturalistic framework does not capture or explain the right-making relation, which I think is exactly right.

were determined by them as such. That is how it is for at least many naturalist views (though not all—see Copp (1995) for a naturalist view that retains an irreducible appeal to standards[93]), but it obviously would not be warranted for the view I have proposed.

Finally, the idea of standards of goodness does not drop out of the picture because, even though the particular resultance base features in a given case are value laden, they give rise to the resultant ethical property or fact only by way of a relation to the overall standards that are informed by many other value laden features and facts. So the story is more holistic than anything we get focusing just on the particular features in the resultance base of a given ethical property or fact. A particular action, for example, may be bad by virtue of causing someone pain. But the link between this feature and the action's badness is not simple and direct: it goes by way of the fact that *all things considered*, this feature of the action (in its empirical aspect) makes it violate the set of ethical standards that has the content and structure it does because of the implications of the whole set of evaluative and normative properties and facts inherent in the sphere of the world with which ethics is concerned. Again, none of this is eliminated by Jackson's equivalences, and the lesson, I think, is that possible worlds metaphysics is ultimately of limited value for understanding ethical facts and properties.[94]

12. CONCLUSION

I began with a complaint about the rush to avoid the metaphysical and epistemological worries Mackie raised, and the resulting impoverishment of the forms of ethical realism thought to be safe in these respects. My primary aim has been to try to understand what it is that drives some of us to think we need more, what is at stake for our understanding of ethics, and what the desiderata of a robust ethical realism ultimately commit us to. Someone might, of course, wonder whether this has just been an extended exercise in spelling out what Mackie has already showed us we cannot have,

[93] But see n. 72 above for why I nonetheless find Copp's naturalist approach problematic.

[94] For reasons similar to the above, we should resist the suggestion that Shafer-Landau's or Brink's exhaustive constitution claim turns out after all to be true in a sense on my view, as long as the dual aspect nature of the 'natural properties and facts' is taken into account. Such appropriation of talk of 'exhaustive constitution' would be misleading, since the exhaustive constitution view has always been understood to employ the traditional metaphysics of natural properties rather than the value laden account I have suggested, which represents a deep departure. It would also leave out the important structural point about the role of standards in the determination of ethical facts.

amounting to a *reductio* of robust ethical realism. Such a concern obviously needs to be addressed through a careful critique of Mackie's argument from queerness, which I think is far less powerful than it is usually taken to be. Some of what I have said here already speaks to such a critique, by showing, for example, that robust realists need not be committed to the kind of Platonism Mackie attacked, with transcendent entities boasting a coercive motivating force. The real question is whether the dual aspect view of a value laden reality that I have sketched and defended is inconsistent either with genuine scientific results or with metaphysical or epistemological claims that are compelling without reliance on scientistic assumptions. I believe it is not, though I cannot argue for that here. My purpose has been to provide enough motivation for such a view to suggest that it is time for that question to be taken more seriously again, and for robust ethical realism in some form to have a respectable place at the table in contemporary metaethics.

REFERENCES

Allison, Henry E. (1990) *Kant's Theory of Freedom* (Cambridge: Cambridge University Press).

Birondo, Noell (2006) 'Moral Realism without Values' *Journal of Philosophical Research* 31:, 81–102.

Boyd, Richard (1997) 'How to Be a Moral Realist' in Darwall, Gibbard, and Railton (1997), 105–36.

Brink, David (1989) *Moral Realism and the Foundations of Ethics* (Cambridge: Cambridge University Press).

—— (2001) 'Realism, Naturalism and Moral Semantics' *Social Philosophy and Policy* 18: 154–76.

Copp, David (1995) *Morality, Normativity, and Society* (Oxford: Oxford University Press).

—— (ed.) (2006) *The Oxford Handbook of Ethical Theory* (Oxford: Oxford University Press).

Crisp, Roger (1996) 'Naturalism and Non-Naturalism in Ethics' in Sabina Lovibond and S. G. Williams (eds.), *Identity, Truth and Value* (Oxford: Blackwell), 113–29.

Cullity, Garrett, and Gaut, Berys (eds.) (1997) *Ethics and Practical Reason* (Oxford: Oxford University Press).

Dancy, Jonathan (1993)*Moral Reasons* (Oxford: Blackwell).

—— (2004) 'On the Importance of Making Things Right' *Ratio* 17: 229–37.

—— (2006) 'Nonnaturalism,' in Copp (2006), 122–45.

D'Arms, Justin, and Jacobson, Daniel (2006) 'Sensibility Theory and Projectivism,' in Copp (2006), 186–218.

Darwall, Stephen, Gibbard, Allan, and Railton, Peter (eds.) (1997*a*) *Moral Discourse and Practice* (Oxford: Oxford University Press).

—— —— ——(1997*b*) 'Toward *Fin de Siécle* Ethics: Some Trends,' in eid. (1997*a*), 3–47.

Enoch, David (2005) 'Why Idealize?' *Ethics*. 115/4. (July), 759–87.

—— (2006) 'Agency, Shmagency: Why Normativity Won't Come from What is Constitutive of Action' *Philosophical Review* 115/2: 169–98.

—— (2007*a*) 'An Outline of an Argument for Robust Metanormative Realism' in Shafer-Landau (2007).

—— (2007*b*) 'Rationality, Coherence, Convergence: A Critical Comment on Michael Smith's *Ethics and the A Priori*' *Philosophical Books* 48/2: 99–108.

FitzPatrick, William (2000) *Teleology and the Norms of Nature* (New York: Garland).

—— (2004) 'Reasons, Value and Particular Agents: Normative Relevance Without Motivational Internalism' *Mind* 113/450: 285–318.

—— (2005) 'The Practical Turn in Ethical Theory: Korsgaard's Constructivism, Realism, and the Nature of Normativity' *Ethics* 115/4 (July): 651–91.

Foot, Philippa (2001) *Natural Goodness* (Oxford: Oxford University Press).

Haldane, John and Wright, Crispin (eds.) (1993) *Reality, Representation, and Projection* (Oxford: Oxford University Press).

Horgan, Terry, and Timmons, Mark (2000) 'Non-Descriptivist Cognitivism: Framework for a New Metaethic' *Philosophical Papers* 29: 121–53.

Jackson, Frank (1998) *From Metaphysics to Ethics* (Oxford: Oxford University Press).

Joyce, Richard (2002) *The Myth of Morality* (Cambridge: Cambridge University Press).

Klagge, James C. (1988) 'Supervenience: Ontological and Ascriptive' *Australasian Journal of Philosophy* 66/4: 461–70.

Korsgaard, Christine (1996) *The Sources of Normativity* (Cambridge: Cambridge University Press).

—— (1998) 'Motivation, Metaphysics, and the Value of the Self: A Reply to Ginsborg, Guyer, and Schneewind" *Ethics* 109/1 (October): 49–66.

—— (2003) 'Realism and Constructivism in Twentieth Century Moral Philosophy' in *Philosophy in America at the Turn of the Century*, APA Centennial Supplement, *Journal of Philosophical Research* 99–122.

Lewis, David (1989) 'Dispositional Theories of Value' *Proceedings of the Aristotelian Society* suppl. 63: 113–37.

McDowell, John (1980) 'The Role of *Eudaimonia* in Aristotle's *Ethics*' in A. O. Rorty (ed.), *Essays on Aristotle's Ethics* (Berkeley: University of California Press), 359–76.

—— (1995*a*) 'Might there be External Reasons?', in J. E. J. Altham and H. Ross (eds.), *World, Mind and Ethics* (Cambridge: Cambridge University Press) 68–85.

—— (1995*b*) 'Two Sorts of Naturalism' in R. Hursthouse, G. Lawrence, and W. Quinn (eds.), *Virtues and Reasons* (Oxford: Oxford University Press), 149–80.

—— (1997*a*) 'Projection and Truth in Ethics' in Darwall, Gibbard, and Railton (1997*a*), 215–26.

—— (1997*b*) 'Values and Secondary Qualities' in Darwall, Gibbard, and Railton (1997*a*), 201–14.

Mackie, J. L. (1977) *Ethics: Inventing Right and Wrong* (Harmondsworth: Penguin).

Nagel, Thomas (1979) 'Ethics without Biology' in *Mortal Questions* (Cambridge: Cambridge University Press): 142–6.

—— (1986) *The View From Nowhere* (Oxford: Oxford University Press).

Parfit, Derek (1997) 'Reasons and Motivation' *Proceedings of the Aristotelian Society* suppl. 71: 99–130.

—— (2006) 'Normativity' in Shafer-Landau (2006), 325–380.

—— (forthcoming) 'Normativity, Naturalism and Non-Cognitivism' appendix A of *Climbing the Mountain*.

Railton, Peter (1997) 'Moral Realism' reprinted in Darwall, Gibbard, and Railton (1997*a*), 137–66.

—— (2003) 'Red, Bitter, and Good' in *Facts, Values and Norms* (Cambridge: Cambridge University Press), 131–47.

Rosati, Connie (1995) 'Naturalism, Normativity, and the Open Question Argument' *Noûs* 29/1, 46–70.

—— (2003) 'Agency and the Open Question Argument' *Ethics* 113 (April): 490–527.

Sayre-McCord, Geoffrey (1988) 'The Many Moral Realisms' in Geoffrey Sayre-McCord (ed.), *Essays on Moral Realism* (Ithaca NY: Cornell University Press), 1–26.

—— 'Moral Realism' (2005) *The Stanford Encyclopedia of Philosophy (Winter 2005 Edition)*, ed. Edward N. Zalta, *http://plato.stanford.edu/archives/win2005/entries/moral-realism/*.

Scanlon, Thomas (1998) *What We Owe to Each Other* (Cambridge, MA: Harvard University Press).

Schroeder, Mark (2005) 'Realism and Reduction: The Quest for Robustness' *Philosopher's Imprint* 5/1 (February): 1–17.

Shafer-Landau, Russ (2003) *Moral Realism: A Defence* (Oxford: Oxford University Press).

—— (ed.) (2006) *Oxford Studies in Metaethics*, i (Oxford: Clarendon Press).

—— (ed.) (2007) *Oxford Studies in Metaethics*, ii (Oxford: Clarendon Press).

Smith, Michael (1993) 'Objectivity and Moral Realism: On the Significance of the Phenomenology of Moral Experience' in Haldane and Wright (eds.) (1993), 235–56.

—— (1994) *The Moral Problem* (Oxford: Blackwell).

—— (1997) 'A Theory of Freedom and Responsibility'' in Cullity and Gaut (1997), 293–320.

Sturgeon, Nicholas (2003) 'Moore on Ethical Naturalism' *Ethics* 113: 528–56.

Wiggins, David (1993*a*) 'Cognitivism, Naturalism and Normativity: A Reply to Peter Railton' in Haldane and Wright (1993), 301–14.

—— (1993*b*) 'A Neglected Position?' in Haldane and Wright (1993), 329–36.

—— (1997) 'A Sensible Subjectivism?' reprinted in Darwall, Gibbard, and Railton (1997*a*), 227–46.

Williams, Bernard (1981) 'Internal and External Reasons' in *Moral Luck* (Cambridge: Cambridge University Press), 101–13.

—— (1995) 'Internal Reasons and the Obscurity of Blame' in *Making Sense of Humanity* (Cambridge: Cambridge University Press), 35–45.

This page intentionally left blank

8

Constructivism about Reasons

Sharon Street

1. INTRODUCTION

Do valuable things possess their value independently of our valuing them? Or does their value always depend, at least ultimately, on our taking them to be valuable? I have argued elsewhere that Darwinian considerations settle this debate in favor of the latter, antirealist view.[1] Things are valuable ultimately because we take them to be. But the *ultimately* in this statement is important. While there are, ultimately, no normative truths that hold independently of our evaluative attitudes—while normative realism is false, in other words—it does not follow that it's impossible to go wrong with one's normative judgments. On the contrary, there is still a robust sense in which normative judgments can be, and often are, in error. This is the important core of truth in realism. The truth in antirealism, however—and what makes antirealism the right view in the end—is that the standards of correctness that determine what counts as an error are ultimately set by our own normative judgments. To put the point another way: A person does not have a normative reason merely in virtue of taking herself to have it; it's easy to go wrong about one's reasons, and we do so all the time. At the same

For their helpful comments on earlier versions of this paper, I am indebted to Melissa Barry, Stephanie Beardman, Paul Boghossian, Ronald Dworkin, David Enoch, Matt Evans, Michael Gill, Elizabeth Harman, Mark LeBar, James Lenman, Tristram McPherson, Thomas Nagel, Derek Parfit, Mark Schroeder, Matthew Silverstein, Sebastian Watzl, and especially Christine Korsgaard, T. M. Scanlon, Nishi Shah, and David Velleman. Thanks also to two anonymous referees from Oxford University Press, and to audiences at the NYU Colloquium in Legal, Political and Social Philosophy and the Third Annual Metaethics Workshop at the University of Wisconsin.

[1] "A Darwinian Dilemma for Realist Theories of Value" *Philosophical Studies* 127/1 (January 2006): 109–66.

time, however, the reasons a person has are always *ultimately* a function of the reasons she takes herself to have; any mistakes she makes will, in the end, be on her own terms.

This is a sketch of the view I'll call *metaethical constructivism*. Constructivism is increasingly mentioned and discussed in the contemporary metaethical debate, but there are relatively few developed statements of the view, and persisting puzzlement about exactly what constructivism is and whether it constitutes a metaethical position, much less a distinct one, at all.[2] The goal of this paper is to offer a systematic statement of constructivism that responds to these questions. I propose a general characterization and taxonomy of constructivist views in ethics, and then present the main outlines of the constructivist metaethical view I favor.

2. TWO KINDS OF CONSTRUCTIVISM IN ETHICS

We may begin by distinguishing two kinds of constructivism in ethics: what I'll call versions of *restricted constructivism*, on the one hand, and versions of *thoroughgoing* or *metaethical constructivism*, on the other.[3] On the general understanding of constructivism that I propose, the central, distinguishing feature of all constructivist views in ethics—whether restricted or thoroughgoing—is this:

Constructivist views in ethics understand the correctness or incorrectness of some (specified) set of normative judgments as a question of whether those judgments withstand some (specified) procedure of scrutiny from the standpoint of some (specified) set of further normative judgments.

[2] Important statements of constructivism by its supporters include John Rawls's "Kantian Constructivism in Moral Theory," reprinted in John Rawls, *Collected Papers*, ed. Samuel Freeman (Cambridge, MA: Harvard University Press, 1999), 303–58, Christine M. Korsgaard's *The Sources of Normativity* (Cambridge: Cambridge University Press, 1996), and Korsgaard's "Realism and Constructivism in Twentieth Century Moral Philosophy," *Journal of Philosophical Research*, APA Centennial Supplement (2003): 99–122. Important statements of constructivism by its critics include David Brink's discussion in *Moral Realism and the Foundations of Ethics* (Cambridge: Cambridge University Press, 1989), and Russ Shafer-Landau's discussion in *Moral Realism: A Defence* (Oxford: Oxford University Press, 2003). Finally, important expressions of puzzlement about whether constructivism is a (distinct) metaethical position at all include Stephen Darwall, Allan Gibbard, and Peter Railton's discussion on pp. 12–15 of "Toward *Fin de Siècle* Ethics: Some Trends," reprinted in *Moral Discourse and Practice: Some Philosophical Approaches*, ed. Darwall, Gibbard, and Railton (Oxford: Oxford University Press, 1997), 3–47, and Nadeem Hussain and Nishi Shah's "Misunderstanding Metaethics: Korsgaard's Rejection of Realism," in *Oxford Studies in Metaethics*, i, ed. Russ Shafer-Landau (Oxford: Clarendon Press, 2006), 265–94.

[3] Korsgaard makes a similar distinction in section 6 of "Realism and Constructivism in Twentieth Century Moral Philosophy."

According to constructivism, in other words, for a normative judgment in the first set to be correct is for it to stand up to the specified sort of reflective scrutiny; the normative judgment's correctness is *constituted* by the fact that it withstands this scrutiny. To speak more metaphorically (in language that I give specific sense to in Section 7), the standards that determine the correctness or incorrectness of the normative judgment in question are thought to be "given from within," or "legislated by," some further practical standpoint: to be correct is to withstand scrutiny from that standpoint.[4] *Restricted* versions of constructivism make these claims about the nature of correctness with respect to some particular, restricted subset of normative judgments. *Thoroughgoing* or *metaethical* versions of constructivism, in contrast, make these claims about the nature of correctness with respect to *all* normative judgments; they say that this is all correctness or incorrectness ever is in the normative domain. Thus, in a way that I'll explain, metaethical constructivism takes the account of correctness that is proposed by restricted versions of constructivism with respect to certain limited sets of judgments about reasons, and argues that the same type of account applies across the board.[5]

3. RESTRICTED CONSTRUCTIVISM IN ETHICS

Since restricted versions of constructivism are the more familiar, it will be useful to begin with a sketch of their distinctive features, and then to introduce metaethical constructivism as a way of extending certain prominent strands of those theories into full-fledged metaethical territory. We may give the following characterization of restricted constructivism as a general type of view in ethics:

Restricted constructivism in ethics specifies some particular, restricted set of judgments about reasons, and says that the correctness of a judgment about reasons falling

[4] A *practical standpoint* is the standpoint of one who makes normative judgments. One occupies the practical standpoint whenever one judges that some things provide practical reasons, or are valuable, good, bad, required, worthwhile, and so on.

[5] In this paper, I sketch a constructivist account of the nature of *practical* reasons. While I believe that constructivism also provides the best account of the normativity of epistemic reasons, I will not discuss that here. Note also that in this paper I employ the term *reason* (always in the sense of a *normative* reason) as a catch-all normative term, using expressions such as *normative judgments* and *judgments about reasons* interchangeably. Nothing substantive hinges on this choice of language; my points could also be couched in the language of *value*, *should*, *ought*, *goodness*, what *makes sense*, what's *rational*, *worthwhile*, and so on. In some contexts, differences between such normative expressions are important. For my purposes here, they are not. What I am interested in is practical normativity in general.

within that set is constituted by the judgment's withstanding a certain (specified) procedure of scrutiny from the standpoint of some (specified) set of further judgments about reasons.

As this characterization indicates, particular versions of restricted constructivism will vary with respect to their specifications of three main elements: (1) the restricted set of normative judgments to which their account of correctness is meant to apply; (2) the procedure of reflective scrutiny that is involved; and (3) the set of normative judgments from the standpoint of which the procedure of reflection is undertaken (some of which may be "embedded" in the procedure of reflection itself—more on this below). I will refer to these three elements, respectively, as the *target set of normative judgments*, the *procedure of construction*, and the *grounding set of normative judgments*. Two further pieces of terminology will also be useful: I'll refer to the reasons that are the subject matter of the target set of normative judgments as the *results of construction*,[6] and I'll refer to the reasons that are the subject matter of the grounding set of normative judgments as the *materials of construction*.

To illustrate, consider how these main elements are specified by two of the most prominent examples of restricted constructivism in ethics, Rawls's political constructivism and Scanlon's contractualism. Take Rawls's view first—focusing, for the sake of simplicity, on his later writings. As presented in *Political Liberalism*, Rawls's proposed conception of justice, called *justice as fairness*, specifies the central elements of restricted constructivism as follows. The *target set of normative judgments* are judgments about social and political justice in a liberal democratic society. Among these are judgments at all levels of generality about the just distribution of rights, liberties, opportunities, income, and wealth, as well as the political virtues. The *results of construction* are Rawls's two principles of political justice, and the reasons of justice that these principles specify.[7] Among such results of

[6] I choose the term *results of construction* because I think it best captures the important constructivist idea that, once the materials and procedure of construction are adequately specified, the result is often quite determinate. Other terminological options—such as *objects of construction* or *targets of construction*— are potentially misleading, because they inaccurately suggest that the "object" or "target" (i.e. a normative reason) is something that one has free rein to create or not as one wishes, rather than something that one is actually (at least in many cases) *forced to* by one's other normative commitments. For related discussion, see pp. 123–5 (esp. §3.7.5) of Rawls's *Political Liberalism* (New York: Columbia University Press, 1993), where Rawls discusses the idea of *possibilities of construction*, which I take to be equivalent to what I'm calling the *results of construction.* I've chosen to avoid Rawls's term as well, because I think that to the ordinary ear it calls to mind too much of an idea of a wide open freedom to create whatever reasons one wants.

[7] See pp. 121–3 of *Political Liberalism* for a discussion of how the principles of justice "identify which facts are to count as reasons." As Rawls notes, it is important to

construction, for example, are "that slavery is unjust, and that the virtues of toleration and mutual respect, and a sense of fairness and civility, are great political virtues."[8] The *procedure of construction* specified by justice as fairness is the famous procedure of the original position, in which rational agents, conceived as representatives of citizens of a liberal democratic society and subject to what Rawls calls "reasonable conditions," select principles of justice to regulate the basic structure of society. The main *materials of construction* are two normative conceptions: the conception of persons as free and equal, and the conception of society as a fair system of cooperation over time. Rawls takes these normative conceptions to be implicit in the public political culture of liberal democratic societies, and our[9] judgments endorsing these conceptions constitute the *grounding set of normative judgments*. As Rawls explains, in justice as fairness, the materials of construction are "embedded in, or modeled by, the constructivist procedure," such that "the form of the procedure, and its more particular features, are drawn from those conceptions [of the person and of society] taken as its basis."[10]

Drawing together the various elements, then, justice as fairness, as a version of restricted constructivism, claims the following:

> The correctness of judgments concerning social or political justice in a liberal democratic society is constituted by their being in accordance with principles that withstand the scrutiny of the original position procedure (embedded within which are fundamental normative judgments implicit in the public political culture of a liberal democratic society).

In Rawls's later writings, it is an important additional feature of justice as fairness that the standard of correctness in question is a standard of reasonableness rather than truth. Justice as fairness is meant to serve as the focus of what Rawls calls an *overlapping consensus* of reasonable comprehensive doctrines, many of which may involve conflicting positions when it comes to the controversial question of the nature of moral truth. For this reason, justice as fairness avoids taking any position on this

recognize that constructivism does not say that the facts that *count* as reasons are results of construction. This would be implausible. What is a result of construction is not those facts themselves, but rather *their status as normative reasons*—or, to put it another way, the "further fact" that these facts *count* in favor of one thing or another. For example, the fact that some policy would benefit the wealthy at the expense of the poor is not a result of construction, but the *status of this fact as a reason* to reject the policy *is*. To put it another way, the "further fact" that "the fact that a policy would benefit the wealthy at the expense of the poor is a reason to reject it" is a result of construction.

[8] Ibid., 123.

[9] The "we" in question being reasonable members of liberal democratic societies, who, according to Rawls, endorse most or all of the grounding set of normative judgments.

[10] *Political Liberalism*, 103.

matter.[11] It offers no account of the truth of judgments about justice in a liberal democratic society, but rather limits itself to proposing an account of the reasonableness of these judgments. For a judgment about justice to be reasonable, on this account, is for it to be in accordance with the principles that withstand the procedure of scrutiny. The reasonableness of the judgment that slavery is unjust, or that toleration is an important political virtue, for example, is constituted by the fact that these judgments are implied by principles that withstand the scrutiny of the original position.

It is possible to challenge this reading of Rawls. I read him as putting forward an account of the reasonableness of judgments about justice as *constituted* by the outcome of the constructivist procedure. But one might argue that Rawls in *Political Liberalism* does not even want to go that far; one might argue that he is claiming only that the reasonableness of judgments about justice is *indicated* by the outcome of the procedure, but not actually *constituted* by it. This raises a thorny interpretive question, with textual evidence on both sides. But the important point for our purposes is this: If Rawls is *not* making the constitutive claim, then justice as fairness as presented in *Political Liberalism* does not qualify as a genuinely constructivist view. For notice that *any* view in ethics can say that the results of reasoning according to a certain procedure are correct (here, reasonable). What makes a view constructivist is its claim that the results of reasoning according to a certain procedure are correct *because they issue from that procedure*—that to be correct *just is* to issue from that procedure. In other words, what is distinctive about constructivist views is that they understand correctness to be *constituted* by emergence from a certain procedure, and not merely as *coincident* with it. Finally, regardless of whether Rawls is making the constitutive claim in *Political Liberalism*, it is clear that he is making it in his earlier "Kantian Constructivism in Moral Theory," so his view there still stands as an example of restricted constructivism.

Now consider Scanlon's contractualism as another example of restricted constructivism.[12] The *target set of normative judgments* in Scanlon's view is

[11] On this point, see esp. § 3.8 of *Political Liberalism*.

[12] *Contractualism* and *constructivism* are not synonymous. *Contractualist* views, roughly, are accounts of morality that give pride of place to some kind of (usually hypothetical) agreement or contract. *Constructivist* views need not do that at all, even though historically they have been closely associated with such views. On my proposed understanding of constructivism, one could be a restricted constructivist and a utilitarian—if, for example, one held that the correctness of judgments about right and wrong is constituted by their withstanding scrutiny from the standpoint of an impartial observer who is concerned with maximizing the general happiness. What's crucial to constructivism is not the idea of a contract or agreement, but rather the idea that the correctness of a given normative judgment is constituted by its withstanding scrutiny from the standpoint of further normative judgments.

judgments about right and wrong—in particular, judgments concerning the part of morality that Scanlon calls "what we owe to each other," which includes duties to others such as duties not to kill, harm, or deceive, duties to keep one's promises, and duties of rescue and beneficence.[13] The *results of construction* in Scanlon's view are an indefinite number of moral principles, which specify what facts are permissibly counted as reasons when we deliberate about what we owe to each other. For example, one such principle might identify minor inconvenience as an insufficient reason for breaking a promise; another such principle might identify the fact that a course of action would lead to someone's death as a (normally) conclusive reason against it (pp. 199–200). The *procedure of construction* in Scanlon's view is the contractualist formula according to which "an act is wrong if its performance under the circumstances would be disallowed by any set of principles for the general regulation of behavior that no one could reasonably reject as a basis for informed, unforced general agreement" (p. 153). As for the *grounding set of normative judgments*, the most central of these judgments is "embedded" in the contractualist procedure itself: it is the judgment that we have reason "to live with others on terms that they could not reasonably reject insofar as they also are motivated by this ideal" (p. 154). But equally crucial among the grounding set of normative judgments are the judgments that we must call upon to apply the contractualist procedure, namely, judgments about what counts as a valid reason for rejecting a proposed principle. Such "reasons for rejection" are not completely specified in advance in Scanlon's view (p. 218), but he makes it clear that our judgments about them have moral content (p. 217), that they are not restricted to considerations about gains and losses in well-being (p. 214), that they are the reasons of individuals rather than groups (p. 229), and that they do not include what Scanlon calls *impersonal reasons* (pp. 219–20). The *materials of construction* are (as always) the reasons that are the subject matter of the grounding set of normative judgments—in this case, our reason to live according to principles that others could not reasonably reject insofar as they share this commitment, and the set of adequate reasons for rejecting proposed principles of this kind.

Pulling these elements together, Scanlon's contractualism, as a version of restricted constructivism, claims:

> The correctness of judgments about right and wrong (or "what we owe to each other") is constituted by their being in accordance with principles that withstand the

[13] For a discussion of the domain of "what we owe to each other" as distinguished from morality broadly speaking, see § 7 of ch. 4 of *What We Owe to Each Other* (Cambridge, MA: Harvard University Press, 1998). All page references in this paragraph and the next are to this book.

scrutiny of the contractualist procedure (which is undertaken from the standpoint of the judgment that we have reason to live with others on terms they could not reasonably reject, and judgments concerning what would constitute adequate grounds for such rejection).[14]

Scanlon's theory, unlike Rawls's, has no special reason to avoid taking a position on the nature of moral truth. What is important, in Scanlon's view, is that there are standards of correctness such that it is possible to be mistaken in one's judgments of right and wrong. Scanlon has no objection to talking about these standards in terms of truth and falsity, so long as it is understood that we are not talking about truths in the sense of truths "about the natural world outside us or about our own psychology" (p. 60).

4. SIX OBSERVATIONS ABOUT RESTRICTED CONSTRUCTIVISM

With these examples of restricted constructivism in ethics in front of us, I'll now make six observations about this type of view. Later, in Section 8, I develop my portrait of metaethical constructivism by comparing and contrasting its corresponding features.

First, it is important to note that in restricted constructivist views, the question to be asked when assessing the correctness of a "target normative judgment" is not whether anyone *thinks* the judgment withstands the specified form of scrutiny; the question is whether it *does*. In Rawls's view, for example, reasons of justice are not determined by what principles you or I or anyone else *thinks* it would be rational to choose in the original position. Similarly, in Scanlon's view, reasons of right and wrong are not determined by what principles you or I or anyone else *thinks* could not reasonably be rejected. Restricted versions of constructivism do not identify normative facts with these or any other non-normative facts. Instead, they identify some normative facts with others—identifying justice, for example, with what it would be *rational* to accept in a choice situation that is designed to embody certain *normative* conceptions of the person and society (in Rawls's view), or identifying moral wrongness with what would be disallowed by principles for the general regulation of behavior that no one could *reasonably* reject (in Scanlon's view).

[14] Scanlon makes it clear that he means to be affirming the constitutive claim when he says that "the property of moral wrongness can be *identified with* a certain normatively significant property" (p. 12, my emphasis), and that the "contractualist formula ... is intended as an account of *what it is* for an act to be wrong" (p. 391 n. 21, my emphasis).

Second, this approach of characterizing some reasons in terms of others naturally raises worries about circularity and uninformativeness. Ultimately, each version of restricted constructivism must mount its own defense against such charges.[15] Here it suffices to note that there is no reason in principle why a version of restricted constructivism must be circular. In particular, the mere fact that a restricted constructivist view characterizes some reasons in terms of others does not imply that it is circular, for the *goal* of restricted views—and this point is critical to understanding them—is not to provide a characterization of reasons *in general*, but rather to provide a characterization of some limited *subset* of reasons, such as reasons of political justice or reasons of right and wrong. In offering such a characterization, it is legitimate to refer to further reasons, so long as these further reasons do not include the very same ones that one is trying to characterize. Thus, to avoid circularity, a version of restricted constructivism must simply avoid including the results of construction among its materials of construction; it must ensure that there is no overlap between its target set and its grounding set of normative judgments. This is perhaps easier said than done; yet the point remains that there is no reason in principle why a version of restricted constructivism must be circular or uninformative.

Third, in restricted constructivist views, it is thought that at least in many cases (though not all) the question whether a given target normative judgment stands up to the specified procedure of scrutiny has a determinate answer. So, for example, Rawls asserts that there is no possibility that the parties in the original position would agree to a principle permitting slavery,[16] and Scanlon takes his formula definitively to imply results such as the conclusion that it would be wrong to kill someone for the sake of boosting one's income, and wrong to break a promise in order to avoid a trivial inconvenience. Thus, according to restricted constructivist views, once the materials and procedure of construction are adequately specified, certain results are decisively entailed.

Fourth, the idea of radical choice—understood roughly as a choice based on no reason at all—plays no significant role in restricted constructivist views.[17] As should be clear, the principles, reasons, and values that constitute the results of construction are not objects of radical choice: rather, they are affirmed on the basis of their issuing from the procedure of construction,

[15] Scanlon defends his view against this charge in *What We Owe to Each Other*, ch. 5. See esp. pp. 194–5 and 215–18.

[16] *Political Liberalism*, 125.

[17] In "Kantian Constructivism in Moral Theory," Rawls makes this point as it applies to the specific version of restricted constructivism outlined in those lectures. Here I generalize the point to apply to all versions of restricted constructivism.

which itself embodies further judgments about reasons. And what about these grounding judgments, in turn? Their acceptance is not a matter of radical choice either, in restricted constructivist views. In Rawls's view, for instance, although justice as fairness deliberately remains silent about the deeper justifications of its grounding normative judgments, this does not mean that those judgments are affirmed for no reason. Rather, it is assumed that the citizens of a liberal democratic society will each have, as part of their wider comprehensive doctrines, their own reasons for affirming these grounding judgments. Similarly, in Scanlon's view, we are by no means thought to make radical choices concerning what does or does not count as a sufficient "reason for rejection." Rather, we arrive at these grounding normative judgments via reflection based on yet again further judgments about reasons.[18]

The fifth observation concerns the role of *reflective equilibrium* in versions of restricted constructivism. Broadly speaking, the method of reflective equilibrium is a procedure of reflection in which one begins with one's considered convictions (including, of course, one's normative judgments) at all levels of generality, and works back and forth between general principles and more particular judgments—pruning, revising, considering alternatives, and seeking eventually to reach a state of equilibrium in which one's considered convictions at all levels of generality are fully in line with one another.[19] In restricted constructivist views, the method of reflective equilibrium is employed throughout the process of fashioning the overall restricted constructivist view—for example, in identifying the relevant set of grounding normative judgments, in formulating the constructivist procedure itself, and in assessing the results of the procedure and further modifying the view in light of those results. The precise role of reflective equilibrium in Rawls's view is a difficult interpretive question that I will not try to address here.[20] The role of reflective equilibrium in Scanlon's view is clearer.[21] Scanlon's version of restricted constructivism assigns reflective equilibrium the following significance: the fact that a normative judgment stands up to the method of reflective equilibrium is a fact of *epistemological*

[18] Scanlon describes the general method by which we assess the status of a consideration as a reason on pp. 64–72 of *What We Owe to Each Other*. Throughout ch. 5 Scanlon discusses the kinds of considerations that underlie reasoning about what counts as a sufficient "reason for rejection."

[19] On the method of reflective equilibrium, see pp. 19–21 and 48–51 of Rawls's *A Theory of Justice* (Cambridge, MA: Harvard University Press, 1971).

[20] On how to interpret Rawls's views on the method of reflective equilibrium, see T. M. Scanlon's "Rawls on Justification," in *The Cambridge Companion to Rawls*, ed. Samuel Freeman (Cambridge: Cambridge University Press, 2002), 139–67.

[21] For the (rough) equivalent of the method of reflective equilibrium in Scanlon's view, see pp. 64–72 of *What We Owe to Each Other*.

relevance; it serves as a good *indication* that the normative judgment is correct. But the fact is accorded no additional significance beyond that. Later we will see how thoroughgoing constructivism differs on this point.

The sixth and final observation concerns the extent to which restricted constructivist views constitute positions in metaethics.[22] This issue is not straightforward. On the one hand, these views often *sound* metaethical, taking a position as they do on what it is for a given type of normative judgment to be correct, and frequently being couched in other metaethical-sounding terms.[23] On the other hand, by definition as restricted, these views speak only to the question of what it is for a limited set of normative judgments to be correct. And as we have seen, in explaining what it is for their target set of judgments to be correct, they presuppose, but do not themselves speak to, notions of correctness as applied to the normative judgments in their grounding set. These latter normative judgments are built right into the account of correctness for the target set, and nothing is said about what constitutes their correctness in turn.

Because restricted constructivist views say nothing about what it is for their grounding set of normative judgments to be correct, they are in principle compatible with any number of competing metaethical views on that question. And indeed, this is exactly one of the aims of Rawls's political constructivism, which is specifically designed to be able to fit as a "module" within different reasonable comprehensive doctrines.[24] But it is interesting to note the way in which all versions of restricted constructivism may be seen as "modules" capable of being embedded within more thoroughgoing metaethical views. Their silence on the status of their materials of construction is what enables them to do this. Scanlon's restricted constructivist account of right and wrong, for example, is compatible with any number of views on the nature of reasons in general.[25] In principle, a realist, an expressivist, and a metaethical constructivist could

[22] In this paper, I work with a rough and ready understanding of the distinction between normative ethics and metaethics, according to which normative ethics investigates what reasons we have whereas metaethics investigates metaphysical, epistemological, and semantic questions about reasons and normative language. Constructivist positions in ethics may be seen as ultimately breaking down this distinction, but there is not space to discuss this issue here.

[23] While Rawls deliberately avoids metaethical questions in his later work, in earlier work such as "Kantian Constructivism in Moral Theory," metaethical-sounding language is rampant. Scanlon is more sparing in his use of such language, but it nevertheless crops up repeatedly in *What We Owe to Each Other*, for instance when Scanlon describes himself as giving an "account of the property of moral wrongness itself," and draws an analogy with the concept of a natural kind, such as gold, and the property of being gold (p. 12).

[24] The term *module* is Rawls's; see pp. 12 and 145 of *Political Liberalism*.

[25] It's worth emphasizing that it's just Scanlon's contractualism—i.e. the restricted constructivist part of his view—that's compatible with any number of views on the nature

all arrive at an "overlapping consensus" about the nature of right and wrong—with Scanlon's restricted constructivism serving as the focus of that consensus—while nevertheless disagreeing about how to understand the reasons which constitute the materials of the contractualist construction.

In this way, restricted constructivist views might seem merely to push off all broader metaethical questions. Another way to put the point is that in spite of their tendency to sound metaethical, they seem capable of *disappearing* as metaethical views. This happens when one steps into the position of someone reasoning practically. It happens most strikingly of all when one steps into the practical standpoint that is constituted by an acceptance of the grounding normative judgments of the restricted constructivist view in question. For example, if I am a citizen of a liberal democratic society who endorses the relevant conception of persons as free and equal and of society as a fair system of cooperation, then Rawls's constructivism may seem simply to be an argument to me about what principles of justice I should accept. Similarly, if I am someone who accepts that I have reason to live in accordance with principles that others could not reasonably reject, then Scanlon's constructivism may seem simply to be an argument to me about what in particular I owe to others. If, on the other hand, the grounding normative judgments are ones I don't share, then these views may seem, in a frustrating way, not to address me at all, or merely to beg all kinds of questions. In this way, restricted versions of constructivism can appear to be straightforward exercises in normative reasoning: they address those of us who endorse the relevant grounding set of normative judgments, and argue that we have reason to accept the target judgments; they identify certain reasons or values that we, the audience, accept, and try to show us that from these materials, certain results follow.[26] Viewed in this way, restricted versions of constructivism fall squarely in the realm of normative ethics, presenting themselves as explorations of what normative conclusions follow from what normative premises.

In light of this, one might conclude that restricted constructivist views are nothing but first-order normative views, of no metaethical interest

of reasons in general. His overall view in *What We Owe to Each Other* is not compatible with any number of views on the nature of reasons in general. For although Scanlon is a constructivist about the reasons of right and wrong, he is best read as a non-naturalist realist about the reasons which constitute the materials of the contractualist construction. See ch. 1, esp. §§ 11–12, for Scanlon's discussion of the nature of reasons in general.

[26] Rawls is quite explicit in his later work that with his theory of justice, he is intending to address only citizens of liberal democratic societies; see e.g. "Justice as Fairness: Political not Metaphysical," reprinted in his *Collected Papers*, 388–414. Scanlon, meanwhile, is explicit that his intention is not to address the amoralist, but rather those of us who already care about right and wrong (*What We Owe to Each Other*, 147–8).

whatsoever. But this would be a mistake, in my view. After all, these views do say *something* about what constitutes the correctness of normative judgments—albeit a restricted set. In particular, they insist on understanding the correctness of those normative judgments as a matter of their *withstanding scrutiny from the standpoint of further normative judgments*. In insisting upon this, these views give at least limited expression to what I regard as a key metaethical insight. They express the insight—at least so far as their target set of normative judgments is concerned—that standards of correctness for normative judgments should not be understood as set by an independent order that holds apart from agents' normative commitments, but rather should be understood as set by, or from within, the standpoint of someone who accepts *further* normative judgments. Rawls puts the point this way:

> Kantian constructivism holds that moral objectivity is to be understood in terms of a suitably constructed point of view that all can accept. Apart from the procedure of constructing the principles of justice, there are no moral facts. Whether certain facts are to be recognized as reasons of right and justice, or how much they are to count, can be ascertained only from within the constructive procedure, that is, from the undertakings of rational agents of construction when suitably represented as free and equal moral persons.[27]

Here, as elsewhere, Rawls is taking the freedom and equality of persons for granted—a substantive normative position—and in so doing, pushes off a great deal of metaethics (quite deliberately, it's worth emphasizing, since this is a *restricted* constructivist view). But there is nevertheless a broader metaethical point suggested by such passages. Rawls writes elsewhere:

> Observe that in constructivism the objective point of view is always understood as that of certain reasonable and rational persons suitably specified .... [I]n justice as fairness, it is the point of view of free and equal citizens as properly represented. Thus, in contrast to what Nagel calls the "impersonal point of view," constructivism both moral and political says that the objective point of view must always be from somewhere.[28]

This I regard as the fundamental metaethical insight of restricted constructivist views—this idea that determinations of the correctness of normative judgments must always be made *from somewhere*—and in particular, from some practical point of view, constituted by the acceptance of further normative judgments. From nowhere, according to constructivist views in ethics, there are no normative facts. While such points require further development before we have a full-blown metaethical position, I believe such a position is in the offing.

[27] "Kantian Constructivism in Moral Theory," 307.
[28] *Political Liberalism*, 116.

So are restricted constructivist views metaethical views after all? The right answer, in my view, is that they are not—certainly not in any thoroughgoing way—but that they contain the seeds of one. They aren't thoroughgoing metaethical views because they take for granted, without explaining, the notion of correctness as it applies to the normative judgments in their grounding set. But they contain the seeds of a thoroughgoing metaethical view because the thought naturally suggests itself: what happens if we take the account of correctness proposed by these views for their limited target set of normative judgments, and try applying it across the board?

5. METAETHICAL CONSTRUCTIVISM INTRODUCED: EVOLUTION, REALISM, AND CONSTRUCTIVISM

Restricted constructivist views specify a conception of correctness for their target set of normative judgments, but say nothing about what it is for their grounding judgments to be correct. It's here that metaethical constructivism enters in and says that, when we turn to that question, the answer has to be of the same general type. What constitutes correctness in the case of the grounding judgments is the same basic thing that constituted correctness in the case of the target judgments—namely that these grounding normative judgments, when held up for examination in their turn, stand up to scrutiny from the standpoint of yet again *further* normative judgments. And the same will be said for *those* normative judgments in their turn, and so on, all the way down. (This raises the obvious question of how all this bottoms out. Different versions of metaethical constructivism part ways on this issue, and I return to it below.) According to metaethical constructivism, it is a mistake to ask about the correctness of *any* normative judgment in the utter abstract, without making at least implicit reference to a standpoint constituted by some further set of normative judgments. This is because there are no judgment-independent standards of correctness in the normative domain; the only standards of correctness that exist are those set from within the practical point of view itself.

Metaethical constructivism is thus premised on a rejection of realism. The goal of this paper is not to argue for the rejection of realism; I do that elsewhere.[29] But the rough intuitive picture may be presented with a brief evolutionary thought experiment.[30] For billions of years after the Big Bang,

[29] See "A Darwinian Dilemma for Realist Theories of Value."

[30] This thought experiment also has another purpose. Traditionally, constructivists in ethics (such as Rawls, Scanlon, Korsgaard, and Kant, if you read Kant as a constructivist) do not place much emphasis on considerations drawn from empirical science. But one

the universe was devoid of life. Gradually, however, on one or more planets, life evolved, at first only in non-conscious forms. This much is no thought experiment, of course; it's what actually happened. Now for the thought experiment. One day, let's suppose, the first two valuing creatures *ever* were born—remarkably, as it happened, in a fairly sophisticated form.[31] Until that moment, nothing had ever consciously valued anything. These two creatures were of the same species, let us suppose—perhaps even born of the same parents—but due to a random difference in their genes, they happened to take different things to be valuable. As it so happened, the first valued its own survival and nothing else, whereas the second valued its own destruction and nothing else. The first creature, whenever it saw that something would promote its own survival, enthusiastically sought to do it, feeling elation whenever it succeeded and anxiety whenever it didn't. (Imagine this creature leaping out of the way of an oncoming boulder, managing to dodge it just in time, and rejoicing in its intact state afterwards.) In an exactly parallel way, the second creature, whenever it saw that something would promote its own destruction, enthusiastically sought to do it, feeling elation whenever it succeeded and anxiety whenever it didn't. (Imagine this creature leaping *into* the path of an oncoming boulder, managing to throw itself underneath just in time, and rejoicing in its crushed state afterwards.)

I don't need to say which of these two creatures left more descendants, and which creature's genetic material was wiped out pretty much as soon as it appeared. Now for the intuitive thought behind metaethical constructivism: When the first creature judged that its own survival was good, and the second creature judged that its own survival was bad, the first was not recognizing some normative truth that the second was somehow missing. The first creature *survived*, of course, but this isn't because its judgment was *true*, but rather simply because that creature tended to do what promoted its survival. The second creature, in contrast, *didn't* survive, but that's not because it failed to be sensitive to any normative reality; it's rather simply because that creature sought to destroy itself and succeeded. Those with strong realist intuitions may wish to insist that the second creature was making some kind of mistake. After all, one might think, look where it led him. But if you're tempted to think this, ask yourself *why* you're tempted to think this. No doubt it's much easier to *relate* to the first creature's

attractive feature of metaethical constructivism—indeed, an important part of what I think ultimately forces us to the view—is its compatibility with a naturalistic worldview. One point of the following thought experiment is to gesture at this.

[31] This entirely unrealistic detail is one among many that make this a mere thought experiment rather than an accurate description of how the course of evolution actually proceeded. For a more thorough discussion of relevant issues, see §§ 3 and 4 of "A Darwinian Dilemma for Realist Theories of Value."

normative judgment. But why is that? The answer is not that this creature was *right*, but rather that you're *his* descendant.[32] The second creature didn't leave any. But that doesn't mean the second creature was making an *error* of any kind. Indeed, he got exactly what he valued—his own elimination; it's just that he left no descendants to carry on the evaluative tradition.

The constructivist intuition about this thought experiment is that these two creatures' normative judgments—about the goodness and badness of their own survival, respectively—were neither true nor false. They were instead mere instances of valuing, born of chance alone, not properly called correct or incorrect. No independent standards existed (nor do any exist now) to give any sense to the notion of truth or falsity when it comes to these two creatures' values; their normative judgments merely popped into existence in a universe which until that moment had been utterly devoid of standards.

With these two judgments, however, two very limited standards came into existence, according to the constructivist. (*Life* sets standards, on this view—consciously valuing life in particular.[33]) Since each creature valued one thing, each—in virtue of his single value—could now make *instrumental* mistakes.[34] Imagine the two creatures coming across a plant that, unbeknownst to them, was poisonous. If the first creature, thinking it would promote his survival, judged that he had reason to eat the plant, he would be making a mistake according to the standards set by his single non-instrumental value. Similarly, if the second creature, also believing that it would promote his survival, judged that he did *not* have reason to eat the plant, then he would have been making a mistake according to the standards set by *his* single non-instrumental value. When it came to these two creatures' judgments about what was non-instrumentally valuable, however, according to the constructivist, there were (and are) no standards in place to determine their truth or falsity. Neither was correct or incorrect.

The thought experiment isn't finished yet, though, for note that a constructivist thinks that when it comes to us living now, talk of correctness

[32] Not literally, of course, since this is a thought experiment.

[33] Before these two creatures arrived on the scene, there was teleologically organized life, but no consciously valuing life. This was not enough for standards to enter the world, in my view. Plants, fungi, and single-celled animals, for example—which are teleologically organized but in no way conscious—cannot make *mistakes*, any more than a lamp or a bicycle. With the emergence of conscious valuing—probably first in a rudimentary form of pleasure and pain—standards began to enter the world, for there was now a sense in which creatures could make mistakes. When a creature did something that was painful to it, it was, at least in an extremely primitive sort of way, making a mistake—doing something bad—as determined by the standards set by its own perspective. For discussion of pain as a form of valuing, see §9 of "A Darwinian Dilemma for Realist Theories of Value."

[34] I explain this in depth in §7 below.

and incorrectness with regard to our non-instrumental values has a place. What changed? Imagine now a third creature—perhaps a descendant of the first. Whereas the first two creatures each valued one and only one thing (non-instrumentally), the third, let us suppose, valued *two* things non-instrumentally: its own survival and the survival of its offspring. So take this creature's judgment that "My survival is valuable." The constructivist intuition is that with this third creature, talk of truth and falsity with respect to this judgment at least *starts* to get a foothold, because now a further standard is in place to determine its correctness—in particular, in this case, the standard set by its own *other* non-instrumental value. If, for example, the third creature's offspring depend on it for sustenance, then its survival is necessary for theirs, and in this sense the third creature is *correct* (as judged from the standpoint of its judgment that its offspring's survival is valuable) to judge that its own survival is valuable. The constructivist view is that if the *first* creature—the one who valued its own survival and nothing else—woke up one day and no longer took its survival to be worth pursuing, this would in no sense be an error; the creature merely would have changed. If, on the other hand, the third creature woke up one day and no longer took its survival to be valuable, there is sense to be made of the idea that it would be making a mistake. "Of course your survival is valuable," a sibling might properly say, "your children need you." The third creature would be making a mistake on his own terms.

All this is a mere sketch. But the goal here is simply to convey the rough intuitive picture driving metaethical constructivism. The picture is this: The possibility of evaluative error entered the world with consciously valuing life. There are no standards of correctness determining whether an agent's values are correct or incorrect except those set by her own further values. If she has no others, then she can't be making a mistake, though neither can she be getting anything right.

6. METAETHICAL CONSTRUCTIVISM STATED MORE FORMALLY

We may now state metaethical constructivism more formally, in a way that highlights its continuity with restricted constructivist views:

> According to *metaethical constructivism*, the fact that *X* is a reason to *Y* for agent *A* is constituted by the fact that the judgment that *X* is a reason to *Y* (for *A*) withstands scrutiny from the standpoint of *A*'s other judgments about reasons.

As this statement of the view makes explicit, metaethical constructivism involves a certain relativism. It is important to be clear right away what kind of relativism this is. Here two points need highlighting.

First, I said in Section 4 that the central metaethical insight suggested by restricted constructivist views is that standards of correctness for normative judgments are set by, or from within, the standpoint of someone who accepts further normative judgments. But this formulation leaves vague the answer to an important question: *whose* further normative judgments set the standards of correctness for *which* other judgments? According to metaethical constructivism, in other words, judgments of truth and falsity in the normative domain must always relativize to a particular practical point of view. But relativize in what way? There are two main possibilities. One option is to understand the truth of "*X* is a reason to *Y* for agent *A*" as a function of the normative judgments of *the person judging* whether *X* is a reason to *Y* for agent *A*—for example, my normative judgments if I'm the one making the judgment about *A*'s reasons, your judgments if you're the one making the judgment about *A*'s reasons, and so on. A second option is to understand the truth of "*X* is a reason to *Y* for agent *A*" as a function of the normative judgments of *the person whose reasons are in question*—that is, of *A* herself. Metaethical constructivism selects the second route. The standards of correctness determining what reasons a person has are understood to be set by *that person's* set of judgments about her reasons.

The constructivist selects this route for the simple reason that it accords much better with our overall usage. On the first account, it impossible for you and me sensibly to disagree about whether *X* is a reason to *Y* for *A*, since the answer might be "yes" for me but "no" for you. On the constructivist account, in contrast, you and I and everyone else, including *A* herself, can all sensibly disagree about what reasons *A* has and be talking about the exact same thing, for there's a common question that we're all disagreeing about, namely what withstands scrutiny from the standpoint of *A*'s normative commitments. Thus, according to metaethical constructivism, facts about reasons are judgment-dependent in the sense that a person's reasons depend on the reasons *she* judges herself to have, but they are *not* judgment-dependent in the sense of depending on the reasons *others* take her to have; the truth of "*X* is a reason to *Y* for agent *A*" relativizes not to the speaker's normative commitments, but rather to *A*'s. Thus, even though *A*'s reasons ultimately depend on what she takes them to be, all of us—including *A* herself—can be mistaken about what those reasons are. This can happen, for example, if we're all unaware of some non-normative fact that, in concert with *A*'s set of values, implies that there is reason for *A* to do *Y*—for instance, to look under the refrigerator for her keys (since unbeknownst to us all, they're there), or to give up trying to be a writer (since unbeknownst to us all, it will bring her nothing but ill health and misery).

Second, as the formal statement of the view suggests, according to metaethical constructivism, *X* is never a reason in favor of *Y* absolutely, full stop; *X* can only be a reason in favor of *Y for some agent A*, as determined by the standards set by the normative judgments of *A* herself. Just as the question "Is the Empire State Building taller?" is ill-formed in the absence of any (at least implicit) answer to the question "Taller than what?", so, according to metaethical constructivism, the question "Is *X* a reason to *Y*?" is ill-formed in the absence of any (at least implicit) answer to the question "For whom?" In the absence of such specifications, one has failed to point to the standard that makes the question make sense; no absolute standard exists. It's important to note, however, that nothing said so far rules out the possibility that for a given *X* and *Y*, *X* is a reason to *Y for every agent*—that is, from every practical point of view—no matter who *A* is, no matter what his or her particular normative judgments. This will be the case if it turns out that no matter what set of normative judgments one accepts, it's implied by those judgments that *X* is a reason to *Y*.[35] Thus, a very strong sort of universalism about practical reasons is fully compatible with metaethical constructivism, for all that has been said so far. I return to this issue in Section 9. The point for now is that, according to metaethical constructivism, *if* there are any universal truths about reasons, then they must be of the following kind: they must be "legislated" from within the standpoint of every creature who takes anything at all to be valuable. They are universally correct because the standards set by the normative judgments of every agent say they are correct.

Let us now see how metaethical constructivism specifies the various elements characteristic of all constructivist views in ethics. The *target set of normative judgments* in this case is the set of all normative judgments, and the *results of construction* are all reasons. The *procedure of construction* in metaethical constructivism requires an extended treatment; I return to this in Section 7. In brief, though: In contrast to restricted constructivist

[35] Korsgaard defends such a view in *The Sources of Normativity*, and Kant may be read as taking a similar view. These are what I call later (in § 9) *substantive* versions of metaethical constructivism. Note that there is also another, significantly weaker way in which it could turn out that for every agent, some *X* is a reason for some *Y*. This could turn out to be contingently true if all existing agents, as a matter of contingent fact, just happen to endorse sets of normative judgments such that some further normative judgment stands up to scrutiny from every one of their standpoints. This point is important because, in my own view, this sort of contingent universalism is all (if any) there is. My view is what I call later a *formalist* version of metaethical constructivism.

views, the procedure of construction in metaethical constructivism is given a purely formal characterization. As we saw, in restricted constructivist views, the applicable standard of correctness is given a particular substantive characterization; the correctness of certain normative judgments (those in the grounding set) is simply assumed. In contrast, in metaethical constructivism, the applicable standard of correctness is set by the relevant agent's other normative judgments, *whatever those may be*. The particular content of the standard is not specified in advance, but rather is given by the agent's own normative commitments. Notice that the specification of the procedure *must* be formal in this way: if metaethical constructivism stipulated a procedure of construction with a particular substantive content, then it would be just another version of restricted constructivism, accepting some particular normative standard as a given and pushing off the question of what it is for that further standard to be correct. (Of course one might worry that metaethical constructivism still hasn't managed to avoid this; I address this worry in Sections 7 and 8.)

Turning now to the *grounding set of normative judgments* in metaethical constructivism, this is the set of *all* of the relevant agent's normative judgments, minus the normative judgment whose correctness is in question. Here again notice the contrast with restricted versions of constructivism, which give the grounding normative judgments a substantive characterization. In metaethical constructivism, the grounding judgments are specified formally—without reference to any particular content—as whatever other normative judgments the relevant agent endorses. As for the *materials of construction* in metaethical constructivism, the term as we have been using it does not have a place in this context. Restricted versions of constructivism, as we've seen, define some reasons in terms of others, with these latter reasons being what I've called the "materials of construction." However, since metaethical constructivism is proposed as an account of all reasons, it would of course be illegitimate for it to define reasons in terms of further reasons. In metaethical constructivism, therefore, substantive assumptions about reasons drop out of the definiens, as they must if the account is to be informative. Reasons are not defined in terms of further reasons, but rather in terms of what the relevant agent *takes* or *judges* to be reasons, and how these judgments withstand scrutiny in terms of one another.[36] Since metaethical constructivism makes no substantive assumptions about a person's reasons (not presupposing the value of freedom, equality, survival, or

[36] One might worry that in defining reasons in terms of what agents *judge* or *take* to be reasons, the account builds the very concept it is trying to explain—the concept of a reason—into the definiens, thus rendering the account circular. I address this worry in §8.

anything else), it is not properly said to employ any materials of construction in the sense we've been using that expression, according to which these are the actual reasons referred to by the grounding normative judgments. Metaethical constructivism's "materials," to use the term in a broader sense, are simply the relevant agent's normative judgments themselves: metaethical constructivism explains how all reasons are ultimately "constructed"—or, to put it less misleadingly, *entailed* or *given*—from within the standpoint of creatures who take themselves to have reasons.

7. WITHSTANDING SCRUTINY

I have appealed repeatedly to the notion of one normative judgment's "withstanding scrutiny" from the standpoint of others, and have spoken metaphorically of the way in which every normative judgment may be seen as "setting" or "legislating" standards of correctness for other normative judgments. It's now time to say more about this.

Imagine that someone tells you he is a parent. When you ask how many children he has, he says he doesn't have any. This response will be met with confusion. Assuming we cannot locate the source of the problem elsewhere, we will conclude that he does not understand the concept of parenthood if he persists in saying the following two things, in full consciousness of both at once:

(1) I am a parent.

(2) I have no children.

Now imagine that someone tells you she has conclusive reason to get to Rome immediately, and that flying is the only way to do so. When you ask whether she has bought her plane ticket yet, she says "What are you talking about? I have no reason to get on a plane." This response will similarly be met with confusion. Assuming we cannot locate the source of the problem elsewhere, we will conclude that she does not understand the concept of a reason if she persists in saying the following three things, in full consciousness of all three at once:

(3) I have conclusive reason to get to Rome immediately.

(4) Getting on a plane is the only way to do so.

(5) I have no reason to get on a plane.

A "parent" who has no children is not a parent. Similarly, someone who "judges" that she has conclusive reason to Y, but who (at the same time, in full consciousness) "judges" that she has no reason whatsoever to take what she recognizes to be the necessary means to Y, is not making a normative judgment.

Just as it is constitutive of being a parent that one have a child, so it is constitutive of taking oneself to have conclusive reason to Y that one also, when attending to the matter in full awareness, take oneself to have reason to take what one recognizes to be the necessary means to Y. One *cannot* take oneself to have conclusive reason to Y without taking oneself to have reason to take the means to Y, where the force of the *cannot* here is not rational—as when one says a parent cannot rationally wish her child to be injured—but rather analytic or conceptual—as when one says that a parent cannot be childless. If someone "judges" that she has conclusive reason to Y, while simultaneously and in full awareness also "judging" that she has no reason to take what she recognizes to be the necessary means to Y, then she isn't making a *mistake* about what reasons she has; rather, she simply doesn't count as genuinely making the first "normative judgment" (or for that matter the second) at all. She's not doing what's constitutively involved in taking oneself to have a reason.[37]

[37] This account owes a great deal to Korsgaard's account in "The Normativity of Instrumental Reason," in *Ethics and Practical Reason*, ed. Garrett Cullity and Berys Gaut (Oxford: Clarendon Press, 1997), 215–54. The most important difference between Korsgaard's account and mine is that mine proceeds in terms of what's constitutively involved in *judging* or *taking something to be a reason* rather than in terms of what's constitutively involved in *willing*. This difference is more significant than it might at first seem, for the concept of *willing* is subject to a confusing ambiguity in a way that the concept of *taking something to be a reason* is not. The ambiguity is illustrated by the puzzling way in which the principle "Whoever wills the end wills (what he recognizes to be) the necessary means to that end" seems to be analytically true in one sense, and yet capable of being false in another sense (viz. in cases of instrumental irrationality). In contrast, the principle "Whoever takes himself to have conclusive reason to Y takes himself to have reason to take (what he recognizes to be) the necessary means to Y" is not subject to any similar ambiguity: it is straightforwardly analytic.

It is important to note that this latter principle in no way rules out the possibility of instrumental irrationality. On the contrary, it provides a clear way of thinking about it. A case of instrumental irrationality arises when a person is not sufficiently motivated to go ahead and do what, in virtue of taking herself to have conclusive reason to Y, she already necessarily takes herself to have reason to do—namely, to take the necessary means to Y. In this way, she fails to do what her own normative judgment says she should do. Note that normative judgments are by their very nature motivating, on my view, such that if one judges that one has reason to Y, then one is thereby necessarily at least *somewhat* motivated to Y. But this of course does not mean that judging that one has reason to Y necessarily involves a degree of motivation *sufficient to result in action in every case*: the opposing motivational obstacles (for instance, in the form of fear, depression, temptation, or laziness) may simply be too great. To put the point in terms of an example that Korsgaard considers in "The Normativity of Instrumental Reason" (p. 238): Tex, who lives in Civil War times and is aware that he will die unless he has his leg sawed off (without the benefit of anesthetic), takes himself to have conclusive reason to have the leg sawed off; it's just that his horror of the procedure overwhelms

There are other, similar claims to be made about what is constitutively involved in making judgments about reasons. For example, in the same, conceptual sense of "cannot," someone who judges that X is a reason to Y cannot also (simultaneously, in full awareness) judge that X is *not* a reason to Y. And someone who judges that only facts of kind X are reasons to Y, and who recognizes that Z is not a fact of kind X, cannot also (simultaneously, in full awareness) judge that Z is a reason to Y. These are purely formal statements about what is involved in the very attitude of taking something to be a reason. They make no substantive assumptions about what reasons there are; they merely state what is involved in taking something to be a reason in the first place. If someone "violates" them, then she is not making an *error*; she is merely not taking anything to be a reason. This is similar to the way in which a child who pretends that a pawn is riding a knight is not making a mistake; she is just not playing chess.

We are now in a position to see the sense in which every normative judgment "sets up standards" by which at least some other normative judgments may be judged. We have just seen how if one judges oneself to have conclusive reason to Y, and one is aware that Z is a necessary means to Y, then one cannot, simultaneously and in full consciousness, also judge that one has no reason to Z. But now suppose that one is *not* aware that Z is a necessary means to Y. What the observation about constitutive involvement shows is that if one genuinely judges oneself to have conclusive reason to Y, and it is a fact (of which one is not aware) that Z is a necessary means to Y, then *by one's own lights as someone who genuinely judges herself to have conclusive reason to* Y, one has a reason to Z, even though one is not currently aware of this. In other words, simply by judging yourself to have reason to Y, you're *thereby*—as a constitutive matter—also judging yourself to have reason to take the means to Y, whatever those may be. So even if you don't know that Z is a means to Y, and think you have no reason whatsoever to Z, you *do* have a reason to Z—*according to you*. Your very own normative judgment says so; *it* sets the standard according to which you are making a mistake if you think you have no reason to Z.

It is in this sense that to make a normative judgment is to "give laws to oneself." As soon as one takes anything whatsoever to be a reason, one thereby "legislates" standards according to which, by one's own lights as a

the motivation involved in his acceptance of this normative judgment. The concept of *willing* confuses matters here because there's a sense in which Tex *does* will to have his leg sawed off (in particular, he takes himself to have conclusive reason to undergo the procedure) and a sense in which he *doesn't* will it (in particular, the overall thrust of his motivations—many of which are not rational, but entirely understandable—is to resist the procedure).

valuing agent, one is making a mistake, whether one knows it or not, if one endorses certain other normative judgments. To return to our evolutionary thought experiment, the moment the first valuing creature sprang into existence and took itself to have reason to do what would promote its survival, it thereby set a standard according to which, by its own lights as a valuing agent, it would be making an error if it took itself to have reason to eat a plant which, unbeknownst to it, was poisonous. The creature's taking itself to have one reason constitutively involved taking itself to have other reasons, whether or not it was presently aware of those reasons, or ever would be. For one normative judgment to *withstand scrutiny* from the standpoint of other normative judgments, then, is for that judgment not to be mistaken as determined by the standards of correctness that are constitutively set by those other normative judgments in combination with the non-normative facts.[38]

Before continuing, it is worth noting that on this account, the attitude of *normative judgment* or of *taking something to be a reason* is importantly different from both the attitude of belief and the attitude of desire. Normative judgments are different from *beliefs* in that they are by their nature motivating, such that if one judges that one has reason to Y, then one is thereby necessarily at least somewhat motivated to Y. (This tie with motivation is crucial to understanding why attitudes of this kind evolved in the first place: from an evolutionary point of view, the function of normative judgment is to get us to respond to our circumstances in ways that are adaptive—a job that involves *moving* us.[39]) Yet normative judgments are different from *desires* in virtue of the kinds of constitutive involvements I've been sketching. For example, whereas *taking oneself to have reason* to live constitutively involves *taking oneself to have reason* to undergo the leg amputation that one knows is necessary, the attitude of desire is

[38] While I have focused mainly on the example of "legislating" *instrumental* standards for oneself, it is worth emphasizing that this is just one type of case. For example, if Lois Lane judges that the fact that Superman is back is a reason for celebration, then she thereby legislates a standard according to which she's making a mistake if she thinks that the fact that Clark Kent is back is no reason for celebration. Similarly, if someone judges that only things over which a person has control count as reasons for praise or blame, then she is legislating a standard according to which she's making a mistake if (in a foul mood at 3 a.m.) she finds herself blaming her child for coming down with the flu the night before an oral argument in an important case. Finally, if someone judges that it is never acceptable to profit off the unsafe working conditions of others, then he is legislating a standard according to which his investment portfolio (to which he pays little attention) may be morally problematic. Such examples are important, for they show why metaethical constructivism by no means commits one to a simplistic, purely instrumentalist picture of reasons and value. The standards constitutively legislated by our normative judgments are highly complex and varied in structure.

[39] For further discussion, see "A Darwinian Dilemma for Realist Theories of Value."

characterized by no analogous constitutive involvement: one can *desire* to live while having no *desire* whatsoever to undergo the leg amputation.[40] This is so in a perfectly ordinary sense of *desire*—the sense in which "I have no desire to do this" would be a heroic understatement as the doctor (in Civil War times) approached with his saw. While there might be another sense of *desire* in which one *does* "desire" to have one's leg sawed off, this is a broad ("pro-attitude") sense of *desire* that merely encompasses within it the attitude I'm calling *normative judgment* or *taking something to be a reason*—indiscriminately lumping that attitude together with desires in the narrower, more ordinary sense in which one has no desire to have one's leg sawed off, even if one sees that one must. If our goal is to understand how standards of correctness get generated in the normative domain, then it is essential that we zero in on the attitude of judging or taking something to be a reason, as opposed to the attitude of desire in the narrow, ordinary sense—for it's the former attitude, and not the latter, that constitutively involves other attitudes of the same kind in a way that "sets standards" when combined with the non-normative facts.

8. SIX OBSERVATIONS ABOUT METAETHICAL CONSTRUCTIVISM

We may now consider six points about metaethical constructivism that correspond with the six earlier observations regarding restricted constructivism.

First, as is the case with restricted constructivist views, according to metaethical constructivism, when we ask whether the judgment that *X* is a reason to *Y* (for *A*) withstands scrutiny from the standpoint of *A*'s other normative judgments, we are not asking what *A* or anyone else *thinks* withstands scrutiny from that standpoint. Rather, we are asking whether, as determined by the standards set by *A*'s other normative judgments in combination with the non-normative facts, the judgment that *X* is a reason to *Y* (for *A*) *does* withstand scrutiny from that standpoint.

Second, as in the case of restricted constructivist views, this appeal to what *does* withstand scrutiny may prompt the worry that metaethical constructivism is circular as an account of reasons. The question whether a given normative judgment withstands scrutiny from the standpoint of others may sound like a normative question whose answer will presuppose substantive assumptions about reasons, rendering the account unhelpfully

[40] The leg amputation example is Korsgaard's; see n. 37 above.

circular. But this is not so, as the account of ''withstanding scrutiny'' in the previous section indicates. When metaethical constructivism equates facts about *A*'s reasons with facts about what withstands scrutiny from the standpoint of *A*'s other normative judgments, it is not building any substantive normative assumptions into its definiens. It is merely building in observations about what is constitutively involved in making a normative judgment in the first place. To decide whether a given judgment withstands scrutiny from the standpoint of *A*'s other normative judgments, we need not ourselves presuppose any substantive normative judgments; we need only ask what further normative judgments are constitutively entailed by *A*'s actual normative judgments when we take into account the non-normative facts as we know them (and as he may not). Constitutive entailment is *not* rational entailment, it's worth emphasizing again. Being a parent entails having children, but that does not mean that a ''parent'' is making an error if she has no children; she's just not a parent. Metaethical constructivism thus smuggles no substantive normative assumptions into its definiens, and offers a non-circular account of what it is for *X* to be a reason to *Y* for *A*.

Turning to the third observation, concerning determinacy: Just like the restricted constructivist, the metaethical constructivist thinks that in many (though not all) cases, once the relevant grounding set of normative judgments is adequately specified, the question whether a given target normative judgment withstands scrutiny in terms of them has a determinate answer. By now we've seen the basic way in which this is supposed to work. If someone judges that he has reason to do what would promote his long-term health, then this judgment sets a standard according to which the judgment that he has reason to stop smoking is true, and according to which the judgment that he has no reason to exercise is false (assuming the relevant causal connections). But now it's time to get into further complexities. This example assumes something else, namely that the agent's judgment that he has reason to promote his long-term health *itself* stands up to scrutiny in terms of his other normative judgments. For, according to metaethical constructivism, if we are trying to figure out whether it is true, all things considered, that this person has reason to stop smoking, then what is at issue is not just whether this judgment stands up to scrutiny in terms of some *subset* of his other normative judgments, but whether it stands up to scrutiny in terms of *all* of them. And when we turn to this larger question, the potential for conflicts among the person's normative judgments arises, and with it the potential for indeterminacy as to the truth of the judgment that he has reason to stop smoking.

So let us look at this more closely. Suppose Alex endorses not only the judgment that he has reason to do what would promote his long-term health, but also the judgment that he has reason to do things he finds pleasant and relaxing. The former judgment legislates a standard according to which the judgment that he has reason to stop smoking is true, but the latter legislates a standard according to which the judgment that he has reason to *continue* smoking is true. In one sense, so far, so good, for we think it's true that Alex has a pro tanto reason to quit (namely, that doing so would promote his long-term health), and also true that he has a pro tanto reason to continue (namely, that doing so is pleasurable and relaxing). The metaethical constructivist account seems to capture such thoughts quite well. But what about the truth or falsity of the judgment that Alex has reason to quit *all things considered*? How is the truth value of this judgment determined?[41]

Here, as always, the answer is that we look to Alex's other normative judgments. In particular, we look to his normative judgments concerning the proper trade-offs between present pleasures and future health. (Note that, even though we don't always articulate them, we all implicitly endorse innumerable such judgments about the proper trade-offs between different sorts of values.[42]) Alex, if he's anything like the rest of us, accepts (at least implicitly) some normative judgment along the following lines: present

[41] It's worth noting that a lot of the time (indeed perhaps most of the time) we are not concerned with making normative judgments that are genuinely *all things considered* judgments. Our judgments about reasons are often made relative to some more limited set of normative judgments—ones that are merely accepted as working premises and roughly indicated by the context. For instance, if I ask my husband whether we should stop to pick up milk on our way back from seeing *Superman Returns*, I'd be shocked if he answered "no" on the basis that we shouldn't be living our current lives at all but rather should be laboring as relief workers in Sierra Leone. While he might be right about this, I'm not looking for an all-things-considered reply; I'm asking for a reply in the context of a much more limited set of working assumptions about our reasons. This is an important point to keep in mind when assessing metaethical constructivism, for the view does a good job of elucidating the sense in which we have some reasons from some points of view, others from others, and still others again when we are genuinely asking ourselves about what reasons we have all things considered. We need only identify the relevant set of "grounding normative judgments" in each case. Note that the official statement of metaethical constructivism is a statement of what it is for *X* to be a reason to *Y* for agent *A all things considered*.

[42] The examples are endless. To take just a couple: I might think that the interest and enjoyment that would come from a trip to Thailand is reason enough to take the risks to life and limb associated with that trip, but I might think that the interest and enjoyment that would come from a trip to Iraqi Kurdistan is not reason enough to take the risks to life and limb associated with that trip. Such judgments reflect deeper, implicitly held normative judgments about the value of interest and enjoyment versus the disvalue of risks to life and limb. And we make countless similar judgments about values that might seem even less commensurable. For instance, someone might judge that the fact that she needs to grade papers is sufficient reason for missing the semi-final game in her daughter's soccer playoffs, but insufficient reason for missing the final game. This reflects her deeper

pleasure and relaxation, up to a point, provide reason to do something that detracts from one's future health, but only up to a certain point. For example, Alex might judge that many years of pleasant relaxation via smoking are worth, say, a 5 percent increase in his risk of developing lung cancer in his seventies, but not worth a 40 percent increase in his risk of developing lung cancer in his fifties. These and other such judgments reflect an implicit normative judgment by Alex about the appropriate trade-offs between present pleasures and future health, and this judgment sets a standard capable of resolving the conflict under consideration and determining a truth value for the judgment that "All things considered, Alex has reason to quit smoking."

But of course at this point, a similar series of questions arises all over again, for this last statement assumes that Alex's judgment concerning the proper trade-offs between present pleasures and future health *itself* stands up to scrutiny in terms of his other normative judgments. And of course it might not: for instance, it might be that, given the strong reasons he takes himself to have to accomplish certain projects in his seventies, and given the fact that good health is a crucial prerequisite for accomplishing them, he is not placing enough weight (as determined by the standards set by those other normative judgments) on the importance of his future health relative to present pleasures; he may be underestimating how important his future health is to him. At this point, the question of how all this bottoms out arises once more, and this question must be answered before it can be clear how, on a constructivist view, the truth value of a given normative judgment can be determinate not only from the standpoint of some restricted set of an agent's normative judgments, but also from the standpoint of all of them. I return to this "bottoming out" question in the final section.

But first—in addition to the last three of the six observations—there is a related complexity regarding determinacy that needs to be addressed. Suppose some one normative judgment (or set of normative judgments) *J* fails to withstand scrutiny from the standpoint of some other normative judgment (or set of normative judgments) *K*. Since a failure to withstand scrutiny is always mutual, this means that it will also be the case that *K* fails to withstand scrutiny from the standpoint of *J*. When we're asking what reasons a given agent has all things considered—and not just what reasons she has from the standpoint of some (implicitly or explicitly) specified subset of her values—which standpoint gets priority?[43] The answer, roughly, is that the standpoint that determines what reasons she has is whichever

normative judgments about the proper balance between competing values (here, work and family)—judgments which, as this example illustrates, are remarkably fine-grained.

[43] I am indebted to Matt Evans and Elizabeth Harman for helpful discussion of this.

standpoint is most deeply *hers*, where this is a function of how strongly she holds the normative judgments in question and how close to the center of her total web of normative judgments they lie.

In many cases it will be obvious which standpoint this is. Suppose Beth takes herself to have conclusive reason to eat the bowl of chili in front of her, and also takes herself to have conclusive reason to live a long, healthy life. If she has a life-threatening allergy to peanuts, which unbeknownst to her the chili contains, then each of the two normative judgments in question fails to withstand scrutiny from the standpoint of the other. But if Beth is even remotely statistically normal, then there is little doubt which of the two "normative standpoints" in question is more deeply her own. Her judgment that she has reason to live is fervently held, and lies toward the core of her interlocking web of normative judgments, supporting and being supported by countless others. In contrast, her judgment that she has reason to eat this particular bowl of chili is weakly held and lies at the far outer periphery of her web of normative judgments, supported by few others and supporting even fewer.[44] Thus, even though neither judgment withstands scrutiny from the standpoint of the other, it's determinate which standpoint is more deeply hers, and therefore what Beth's reasons are in this case: her judgment that "I have conclusive reason to eat this bowl of chili" is false.[45] In other cases, however, there may well be no fact of the matter

[44] We can imagine a creature for whom things were the opposite—who valued eating this particular bowl of chili with all her soul, while placing little importance on living another day. According to metaethical constructivism (or at least the formalist version I favor—more about this in the final section), this creature would have reason to eat the chili. (Compare this imaginary creature to the creatures in our evolutionary thought experiment. If you burst onto the scene with certain values, those values—whatever they may be—together determine what reasons you have. It's just that most human beings burst onto the scene with values that would never deem eating one particular bowl of chili more important than life itself.)

[45] One might worry that in according priority to those normative judgments which are more strongly held and which lie closer to the core of a person's interlocking web of normative judgments, the account smuggles in a substantive value. My reply is that the priority accorded these normative judgments doesn't reflect a substantive value, but rather reflects the fact that we are asking about *agent A*'s reasons, not someone else's reasons, and agent *A* is, in an important sense, to be identified with her most strongly and centrally held values. It's a commonplace that to know what a person values most is to know a great deal about who he is. Moreover, if asked to sum up yourself as a person, presumably you will not say things like: "I'm someone who values eating this particular bowl of chili." Rather, you'll say things like: "I'm someone who loves life, my family, my friends, other people, the natural world, music, philosophy," and so on. These are the values you hold most strongly and centrally, and they are the ones that define (in large part) who you are. This tie between a person's values and his identity plays an important role in Korsgaard's version of metaethical constructivism, defended in *The Sources of Normativity*. I discuss the differences between Korsgaard's version of metaethical constructivism and my version in §9.

about which of two standpoints is more deeply the agent's own—it might depend on the order in which she thought about it, or the vividness with which the relevant information was presented to her, and so on; if so, then according to the constructivist, there is no fact of the matter about what her reasons are in that case.[46]

To conclude the discussion of determinacy, note that a constructivist has no objection to the idea that in some cases the truth value of a given normative judgment will indeed be indeterminate. This will occur whenever the standards legislated by a person's other normative judgments, coupled with all the relevant non-normative facts about necessary means, etc., are insufficient to yield a result one way or another. Perhaps the judgment in question is true from the standpoint of some of an agent's normative judgments (say, his commitment to the French Resistance, which legislates that he should leave for England to join the free French forces), but false from the standpoint of others (say, his commitment to caring for his mother, which legislates that he should not), and moreover the person endorses no further principle capable of resolving the conflict.[47] In that case, a constructivist would say that while the person has reasons in favor of going to England, and also reasons against going, there is no fact of the matter about what he should do all things considered. This is because, in the absence of any further normative standpoint from which to assess it—that is, in the absence of any further relevant normative judgment accepted by the agent—there simply *is* no standard that determines a fact of the matter.

In reality, such cases may be fairly rare. Given the vast complexity of our sets of normative judgments, it is probably not often the case that absolutely no other accepted principle is available to arbitrate a given conflict. More often than not, the trouble will come not from lack of an applicable endorsed normative principle, but rather from lack of relevant factual information: the trouble, in other words, will be uncertainty rather than indeterminacy. If, for instance, the Frenchman knew exactly how effective he'd be as a member of the Resistance (perhaps he'd never be more than a mediocre fighter at best), and exactly how effective he'd be in executing his filial duties if he stayed at home (he may be a remarkably sensitive, attentive, and comforting son), he might be able to resolve the conflict based on his commitment to the principle "Other things being equal (the values in question being equally worthy, important and so on),

[46] For relevant discussion, see Don Loeb's "Full-Information Theories of Individual Good," *Social Theory and Practice* 21/1 (1995): 1–30.

[47] Jean-Paul Sartre, "Existentialism is a Humanism," reprinted in *Moral Philosophy: Selected Readings*, 2nd edn, ed. George Sher (New York: Harcourt Brace, 1996), 77–86.

in a choice between fundamental roles and values, take the path at which you'll be most effective." Of course it's possible that even in light of all such principles accepted by him and all relevant factual information, there is no clear answer to his dilemma—perhaps he'd be comparably effective at both roles, for example. In that case, it seems both natural and correct to say that there is no fact of the matter about what he should do. He has some reasons in favor of going to England, others against it, and there is no answer to the question of which set of reasons wins out. Metaethical constructivism thus yields an intuitive result.

These considerations lead directly to the fourth observation, concerning radical choice. We saw earlier that the notion of radical choice plays no significant role in restricted versions of constructivism. (In a sense, these views don't push far enough in their account of reasons to get to it; otherwise it would come up.) In contrast, the notion of radical choice does have an important place in metaethical constructivism. First of all, it has a place in the kinds of case we have just been considering, where there is not a single endorsed normative principle available in one's set of normative judgments to settle whether some other normative judgment is correct or not. As I've said, however, in view of the depth and complexity of our sets of normative judgments (and our stock of "other things being equal" kinds of judgments), such cases may be fairly rare in reality.[48]

There is a second, deeper way in which the notion of radical choice is involved in metaethical constructivism. As we have seen, according to metaethical constructivism, a creature has no reasons until the moment it starts taking itself to have reasons, for until then, there are no standards of correctness determining which attributions of reasons to it are true and which are false. With the first making of a normative judgment, however, standards of correctness are legislated into existence: certain other normative judgments are now constitutively entailed, whether the creature realizes it or not, and the creature is now properly said to be making a mistake if it rejects these other judgments. We human beings of course *do* make normative judgments; it is as natural to us as the use of language itself. Every normally developed adult endorses untold numbers of such judgments, thereby legislating wide-ranging, complicated webs of standards

[48] In restricted constructivist views too, there might be cases in which the "grounding set" of normative judgments is insufficient to settle whether some "target" normative judgment is correct. But restricted constructivists needn't think that radical choice is necessary in such a case, for they may think that the case can be settled by appeal to reasons lying *outside* the restricted domain of reasons in question. In any case, when it comes to the contrast between restricted constructivist and metaethical constructivist views, what's more important is the second, deeper role for radical choice that I'm about to mention.

determining the truth or falsity of claims about what reasons she has. But if one accepts the metaethical constructivist idea that a person's having reasons depends on her taking herself to have some reasons or other, then sooner or later the question arises: why take anything at all to be a reason? After all, at least on its face, it seems as though making normative judgments is something that one could *stop*. That is, at least on its face, it seems that one has it in one's power to either go ahead and take something or other to be a reason, or else not take anything whatsoever to be a reason. If this is indeed so, then one has a choice: to value at all or not—to make some normative judgments or other or not.

It's here that the notion of radical choice enters into metaethical constructivism in a second way. For the choice whether to start (or continue) making any normative judgments at all is not a choice one can make *for a reason*.[49] This is because, according to metaethical constructivism, one has no reasons prior to one's making normative judgments, since these judgments set the standards of correctness for attributions of reasons to oneself. Thus, one can either start valuing or not start; one can either continue valuing or not continue. If one does start (or continue), one has reasons, and if one doesn't start (or continue), one has none. In this sense, the choice whether to value at all—which is just the choice whether to be an agent at all and whether to have reasons at all—is necessarily a radical choice according to metaethical constructivism. In order to be a creature with reasons, each of us must simply step into existence as a valuing creature (and stay there)—and this will be due to mere causes, not reasons (at least not one's own). In this respect, we are all a bit like the two creatures in the evolutionary thought experiment, merely arriving on the scene one day with a particular set of values.[50]

The fifth observation concerns the role of reflective equilibrium in metaethical constructivism. As we saw earlier, in Scanlon's version of restricted constructivism, the fact that a normative judgment withstands scrutiny in reflective equilibrium is viewed as a fact of epistemological significance. In metaethical constructivism, in contrast, the fact that a normative judgment withstands scrutiny in reflective equilibrium is understood to be

[49] There are important differences between the question whether to start making normative judgments and the question whether to continue making them once one has started. In particular, if one is already making normative judgments then it is possible to give *question-begging* reasons for continuing to do so, and the nature and availability of such justifications may be important for one's morale. These issues deserve a fuller discussion than is possible here. Thanks to David Velleman for helpful discussion of this.

[50] I am by no means implying that one is saddled with whatever values one comes alive with. One may throw out any one of them, and can be entirely justified in doing so. It's just that if you are justified in throwing out a given value, it will be because doing so is called for by other values that you hold.

not only of epistemological significance but also of constitutive significance; in other words, this fact is understood to be not merely an *indication* that the normative judgment is correct, but what it *is* for that judgment to be correct. This is true, at least, so long as we understand the method of reflective equilibrium to be identical with the method of deciding whether a given normative judgment "withstands scrutiny," as I have been laying it out—a plausible understanding, if not the only one.

In our consideration of restricted constructivist views, the sixth and final observation concerned the extent to which those views constitute positions in metaethics. In contrast to restricted views, metaethical constructivism is a full-fledged metaethical position. It proposes an informative account of the truth conditions of judgments about practical reasons, one that I believe offers compelling answers to all standard metaethical questions. A full defense of this claim is not possible here. But I hope by now it is reasonably clear, at least in outline, how metaethical constructivism addresses metaphysical and epistemological questions about normative reasons; how it reconciles our understanding of reasons with a naturalistic understanding of the world; and how it explains the connections between reasons, judgments about reasons, and motivation.

There remains at least one important objection to the idea that metaethical constructivism as I've presented it is a full-fledged metaethical view.[51] This worry stems from the way in which the view defines reasons in terms of what we *judge* or *take* to be reasons, thereby seeming to invoke in the definiens the very concept the view is meant to explain. According to this objection, in order to understand what it is to *judge* or *take* something to be a reason, one must already understand what it is for something to *be* a reason, so no informative account of the latter can be given in terms of the former.

My reply to this is as follows. There is one sense in which I agree that in order to understand what it is to judge or take something to be a reason, one must already understand what it is for something to be a reason. In particular, there is a sense in which I agree with Scanlon's claim that "Any attempt to explain what it is to be a reason for something ... lead[s] back to the same idea: a consideration that counts in favor of it. 'Counts in favor how?' one might ask. 'By providing a reason for it' seems to be the only answer."[52] Scanlon takes the idea of a reason as primitive, and there is an important sense in which I think he is right that we must do this. The idea of one thing's being a reason for another cannot successfully be reduced to thoroughly non-normative terms. Instead, I would argue, our understanding

[51] I am grateful to Nishi Shah and Sebastian Watzl for helpful discussion of this.

[52] *What We Owe to Each Other*, 17.

of this idea is given by our knowledge of *what it is lik*e to have a certain unreflective experience—in particular, the experience of various things in the world as "counting in favor of" or "calling for" or "demanding" certain responses on our part. I believe it is impossible adequately to characterize this experience except in such primitive evaluative terms, yet I think we all know exactly the type of the experience I am pointing to. We need only think of how we feel when, for example, a tractor trailer swerves toward us on the highway or we see a stranger threatening our child; we all know what it is like to experience (at an unreflective level that we surely share with many other animals) evasive action or a protective response as utterly "demanded" or "called for" by the circumstances. Just as the experience of color cannot adequately be described except by invoking color concepts, so the type of experience in question—what might be called "normative experience"—cannot adequately be described except by invoking normative concepts.[53] In order to understand what it is to *judge* or *take* something to be a reason, then, one must indeed already understand what it is for something to *be* a reason in the sense that one must already be familiar with the kind of conscious experience I am talking about, and thus know what's meant by the idea of one thing's seeming to *demand* or *call for* or *count in favor of* something else. In this sense, we must take the content of "*X* is a reason to *Y*" to be primitive.

But I do not think the story ends there. After explaining that he will take the idea of a reason as primitive, Scanlon goes on to comment: "The idea of a reason does not seem to me to be a problematic one that stands in need of explanation." Here I disagree. Admittedly, there is a sense in which we all understand perfectly well what a reason is—namely, a consideration that counts in favor of something else—just as there is a sense in which we all understand perfectly well what yellow is—namely, the color we see when we look at the petals of sunflowers, and so on. But there is another sense in which—pre-philosophically, anyway—we do *not* understand what a reason is—a sense in which the idea of a reason is clearly problematic and standing in need of explanation. After all, as Ronald Dworkin points out, no one believes in "morons" (or, we might add, "reas-ons") existing out there in the world on par with protons.[54] And in my view, evolutionary considerations, among others, show that to the extent that normative experience attributes

[53] I am indebted to Sharon Hewitt for helpful discussions of this point. For further discussion of what I'm calling "normative experience" and its likely evolutionary origins, see pp. 127–8 and n. 33 of "A Darwinian Dilemma for Realist Theories of Value."

[54] Ronald Dworkin, "Objectivity and Truth: You'd Better Believe It," *Philosophy and Public Affairs* 25 (1996): 87–139.

any property of "counting in favor of" to objects as they are in themselves, utterly independent of us and our attitudes, that experience is in error.[55] So what—if anything naturalistically comprehensible—*are* reasons, how do we know about them, and so on?

It's this kind of question that metaethical constructivism is intended to answer. And its strategy for doing so is to give an account of what a reason *is* in terms of what it is to *judge* or *take* things to be reasons, where our understanding of this attitude is prior to and fully independent of our understanding of what a reason *is* in the relevant sense—that is, a clear, naturalistically comprehensible sense, as opposed to merely the sense of "a consideration that counts in favor of something." In other words, the word *reason* as it appears in the definiens is *not* understood in the same sense as is being defined by the constructivist proposal as a whole; this would make the proposal uninformative and viciously circular.[56] Instead, the constructivist proposal seizes on the primitive notion of a reason as "a consideration that counts in favor of something," understands *judgments about reasons* in terms of that notion, and then proposes a naturalistically acceptable understanding of reasons as "constructed out of" or "legislated by" such judgments. In this way, the proposal operates with an understanding of the attitude of taking or judging something to be a reason that is fully independent of our understanding of what it is to be a reason in the sense we're trying to discover—namely, a naturalistically acceptable sense. Our independent understanding of the attitude of taking something to be a reason is supplied by two main things. It's supplied first of all, as I've indicated, by our understanding of *what it is like* to have a certain unreflective experience. And it's supplied second of all by our recognition of what is constitutively

[55] To what extent *does* normative experience attribute the property of "counting in favor of" to objects as they are in and of themselves, utterly independent of us and our attitudes? This is a philosopher's question, and I am inclined to think we are over-intellectualizing and distorting if we claim to find a determinate answer in the experience itself. Normative experience—which, keep in mind, is an unreflective experience we share with other animals, just as we share color experience with some of them—is I think best characterized simply as the experience of one thing as counting in favor of or demanding another—and *not* as the experience of one thing as counting in favor of another utterly independently of one's evaluative attitudes. Normative experience, in other words, doesn't itself take a position on the realism/antirealism debate. This matter requires further discussion, however, and if the case can be made that normative experience attributes the property of "counting in favor of" to objects as they are in and of themselves, utterly independent of us and our attitudes, then I embrace an error theory about the content of that experience.

[56] For this point as it applies to the case of color, see Paul A. Boghossian and J. David Velleman, "Colour as a Secondary Quality," reprinted in *Readings on Color: The Philosophy of Color*, ed. Alex Byrne and David Hilbert (Cambridge, MA: MIT Press, 1997), 81–103.

involved in the attitude of judging something to be a reason—the kind of purely formal observations sketched in Section 7.[57]

We see then that there is a sense in which metaethical constructivism is a reductionist view, and another sense in which it is not. Metaethical constructivism is *not* reductionist in the sense that it does not try to reduce the notion of one thing's "counting in favor" of another to non-normative terms; it denies that this can be done, and in this sense takes the notion of a reason to be primitive. But metaethical constructivism *is* reductionist in the sense that it reduces facts about reasons to facts about what we *judge* or *take* to be reasons, with the latter understood in a way that is prior to and independent of the former. The result, in my view, is that even though there is one important, unavoidable sense in which the idea of a reason is being taken as primitive, we nevertheless have secured all that is important—namely an account of reasons that is informative, true to our pre-theoretical concept, and naturalistically fully comprehensible.[58] We enter the world having a certain kind of experience, and facts about reasons are best understood as constructions of that experience.

9. CONCLUSION

Before metaethical constructivism has been given a full explanation and defense, further questions remain to be answered. Some of the most important are: (1) Doesn't the proposal fall victim to Moore's "open question" test?[59] (2) Isn't metaethical constructivism *itself* a normative claim, and if so, doesn't it apply to itself and become self-defeating? (3) Doesn't the view involve an unacceptable degree of relativism about reasons? (4) Is it possible coherently to go forward with one's practical reasoning while at the same time believing that one's reasons have their ultimate root in what one takes them to be?[60] (5) Even if metaethical

[57] The account that I gave in §7 was not meant to be an exhaustive account of what is constitutively involved in judging something to be a reason. Indeed, I think there's a great deal more to be said about this. For example, I would argue that there are important constitutive connections between the attitude of valuing and emotions such as fear, hope, regret, sadness, joy, anxiety, frustration and so on.

[58] Or at least, no *less* naturalistically comprehensible than consciousness itself, since the proposed account of reasons makes key appeal to the conscious experience of some things as "calling for" or "demanding" others.

[59] G. E. Moore, *Principia Ethica* (Cambridge: Cambridge University Press, 1903), ch. 1.

[60] David Enoch calls this into question in "An Outline of an Argument for Robust Metanormative Realism," *Oxford Studies in Metaethics*, ii, ed. Russ Shafer-Landau (Oxford: Clarendon Press, 2007), 21–50.

constructivism is a full-fledged metaethical view, is it really a *distinct* metaethical view? That is, isn't it just another form of (choose one) expressivism, sensibility theory, naturalist realism, or even non-naturalist realism?

I believe there are forceful replies to each of these questions—though each requires a fuller discussion than is possible here. Rather than saying anything about them now, I'll return to a question that I've been postponing, namely the question of how investigations into our reasons bottom out, according to metaethical constructivism. This question also needs a more thorough treatment than is possible here, so for now I'll just offer a brief sketch.

As we have seen, according to metaethical constructivism, the correctness of a judgment about one's reasons must be understood as a matter of whether that judgment withstands scrutiny from the standpoint of one's further judgments about reasons. And when we ask about the correctness of those judgments in turn, the same answer will be true of them, and so on all the way down. All this suggests that, if a given judgment about your reasons is correct, then you will eventually be able to trace its support back to a value or values that are part of a set of interlocking, mutually supporting, and mutually consistent basic judgments about reasons held very deeply by you. But arrival at such an interlocking web of basic normative judgments does not end the questions, for of course now you may ask: Why accept this web of normative judgments rather than some other web? What is it for this whole web to be correct or incorrect?

With such questions, we are finally arriving at the ultimate foundations of our reasons, and we are also reaching a key point with respect to which different possible versions of metaethical constructivism diverge. In particular, we may distinguish between two types of metaethical constructivism: *substantive* versions on the one hand, and *formalist* versions on the other. Both of these views agree on the central constructivist point that the standards determining whether a given normative judgment is correct or incorrect are set from within the standpoint of the creature whose reasons are in question, and that this point holds "all the way down," for every judgment about her reasons. Where they part ways is over the question whether there is anything in particular one must value if one values anything at all—that is, whether there are any reasons that *all* agents have, simply in virtue of their being reflective creatures who accept some normative judgments or other.

According to *substantive* versions of metaethical constructivism, there is some thing (or things) in particular that one must value if one values anything at all; in other words, reflection ultimately bottoms out with a certain substantive value or values. In *The Sources of Normativity*, Korsgaard defends such a view. As a metaethical constructivist, Korsgaard holds that the standards determining the correctness or incorrectness of normative judgments are set from within the standpoint of a reflective creature who accepts such judgments. As a *substantive* metaethical constructivist, Korsgaard holds in addition that there are certain normative judgments to which every reflective creature who accepts any normative judgment at all is committed. In particular, she argues, if you take anything at all to be valuable, then you must take humanity to be valuable, both in your own person and in that of others.

According to *formalist* versions of metaethical constructivism, in contrast, there is nothing in particular that one must value if one values anything at all; in other words, the Kantian project of deriving substantive values from a purely formal understanding of the nature of practical reason fails. This is my own view; I am skeptical that there are any particular substantive judgments about reasons to which every agent is committed simply in virtue of valuing anything at all. Instead, on my view, reflection ultimately bottoms out with an understanding of what is constitutively involved in the attitude of taking something to be a reason, and a recognition of the necessary role of contingencies in determining the substantive content of your reasons. If you had entered the world taking entirely different things to be reasons, on my view, you would have *had* entirely different reasons. You might even have had reason to throw yourself under an oncoming boulder, or to die for the sake of eating a particular bowl of chili, or indeed to prefer the destruction of the whole world to the scratching of your finger.[61]

As this last remark brings out, *formalist* metaethical constructivism combines Kantian and Humean ideas. The view is strongly Kantian in that it takes the notion of autonomy, or of giving laws to oneself, to be of the utmost importance in understanding what truth and falsity in the normative domain consist in. It agrees with Kant that standards of correctness for normative judgments are set from within the practical point of view. Yet the view is strongly Humean in that it accepts that practical reason as such commits us to no particular substantive conclusions about our reasons; depending on one's starting set of values, one could in principle have a reason for anything. To put it in Bernard Williams's terminology: Formalist metaethical constructivism is Humean in that it understands each person's reasons ultimately to be a function of his or her "subjective motivational set."[62] But the view is Kantian in that it argues

[61] David Hume, *A Treatise of Human Nature* (1739–40), 2nd edn, ed. L. A. Selby-Bigge and P. H. Nidditch (Oxford: Oxford University Press, 1978), bk. 2, pt. 3, sect. 3.

[62] Bernard Williams, "Internal and External Reasons," in *Moral Luck* (Cambridge: Cambridge University Press, 1981), 101–13.

that the "elements" in that set are most profitably characterized first of all not as *desires*, but rather as *normative judgments* (and unreflective versions thereof). Focusing on desires leaves it obscure exactly how standards of correctness in the normative domain are generated. Focusing on normative judgments, in contrast, makes this clear.

One might be bothered by the idea that had you entered the world taking entirely different things to be reasons, you would have had entirely different reasons. But this thought shouldn't be troubling once coupled with the recognition that had you entered the world taking entirely different things to be reasons, you wouldn't have been *you* at all. Contingencies have shaped what reasons we have in the same way that contingencies have shaped who we are. But now that we're here, we can shed our normative reasons no more easily than we can shed ourselves.

This page intentionally left blank

9

Rawls and Moral Psychology

Thomas Baldwin

In his obituary of John Rawls Ben Rogers remarked that after completing *A Theory of Justice* (*TJ*)[1] Rawls intended to develop his ideas on moral psychology.[2] In the event the debates aroused by *TJ* kept Rawls fully occupied and he never wrote an extended account of the subject. But there are discussions of it throughout his writings and it merits more attention than it has received (for example, the four-volume collection of papers on Rawls edited by Chandran Kukathas contains no papers directly on this theme[3]). My aims here are to elucidate Rawls's conception(s) of moral psychology and then to explore critically some of the complexities and tensions inherent in his uses of it, especially those arising from its roles in his moral and political theories. Some of these questions are focused around the 'problem of stability', the problem of showing how a just society is likely to be stable because it provides a basic framework for the activities of its members which they recognize as congruent with their individual interests; and at the end of this paper I will discuss Rawls's treatment of this issue, both in *TJ* and in his later writings, and, in this context, consider how far the moral psychology Rawls relies on to address this problem has an essential social dimension.

This last question, how far Rawls's moral psychology has a social dimension, may immediately give rise to the reflection that it should be

This is a revised version of a paper presented at Madison in September 2006. I am much indebted to Russ Shafer-Landau and his team for the invitation to present a paper at the conference, and to those present at the conference and a couple of referees for comments which have led me to revise the paper I gave.

[1] Rawls (1971); in 1999 Rawls published a revised edition of the book—Rawls (1999*a*). I give page numbers for both editions in the form '(*TJ* x; y)', where x is the page number in the first edition and y is the number in the second edition. As I explain in n. 7, one of the revisions Rawls makes is relevant to the argument of this paper.

[2] See www.guardian.co.uk/obituaries/story/0,3604,848488,00.html.

[3] Kukathas (2002).

no surprise if his moral psychology turns out to be inherently sociological or political since Rawls was primarily a political philosopher. Yet although Rawls was of course a great political philosopher, his political philosophy was initially founded on a broadly Kantian moral philosophy, and his moral psychology first took shape within this context. It is then a delicate matter, especially in the context of those of his later writings which are primarily contributions to political philosophy, to identify those discussions which are not exclusively directed to questions of political philosophy; but his moral psychology is, I believe, one of the areas less affected by his later emphasis on the political. Hence even for Rawls there is a substantive issue as to how far moral psychology is inherently social or political; and thinking about the way in which this issue plays out for Rawls certainly provides a stimulating way of thinking about the issue itself.

1. WHAT IS MORAL PSYCHOLOGY?

The first issue to be addressed is what it is that Rawls means when he writes of 'moral psychology'. We get an initial answer to this question from chapter 8 of *TJ* where there is a long section (§75) called 'The Principles of Moral Psychology'. Rawls here summarizes and comments on the account of moral development he has presented in the preceding sections, in particular his account of the development of a 'sense of justice', which is a disposition to act in accordance with the principles of justice for their own sake and to feel guilt or shame when one recognizes that one has violated these principles. This gives us one feature of moral psychology, namely that it deals with the development of feelings and judgments whose content is distinctively moral. But a further point is that, according to Rawls, our psychology is itself affected by the moral value of the context in which we grow up and live:

> Perhaps the most striking feature of these laws (or tendencies) is that their formulation refers to an institutional setting as being just, and in the last two as being publicly known to be such. The principles of moral psychology have a place for a conception of justice ... Thus some view of justice enters into the explanation of the corresponding sentiment; hypotheses about this psychological process incorporate moral notions even if these are understood only as part of the psychological theory. (*TJ* 491; 430)

Rawls recognizes that some theorists will regard this as odd: 'No doubt some prefer that social theories avoid the use of moral notions' (*TJ* 491; 430). But, he holds, this is a mistake: 'The justice or injustice of society's arrangements and men's beliefs about these questions profoundly influence the social feelings' (*TJ* 492; 431). Thus in his use here of the term 'moral psychology' Rawls implies that in some respects our psychology is inherently 'moral', not only in respect of its content, but also in respect of its dependence upon the justice, and thus the morality, of our society.

In *TJ* Rawls maintains that our 'natural attitudes' bring with them a 'liability' (*TJ* 489; 426) to moral sentiments, such that in the normal course of human development in a reasonably just society a normal human being develops a sense of justice. Hence, as Rawls puts it 'The moral sentiments are a normal part of human life. One cannot do away with them without at the same time dismantling the natural attitudes as well' (*TJ* 489; 428). This account of the matter suggests that moral psychology, understood as the psychology of the moral sentiments, deals with an aspect of the normal development of human beings, and therefore belongs within a comprehensive account of human psychology. But this position seems to be at variance with that affirmed by Rawls in his later writings. In lecture II of *Political Liberalism* (*PL*),[4] Rawls gives the final section (§8) the somewhat puzzling title 'Moral Psychology: Philosophical not Psychological' (*PL* 86).[5] He then begins the section as follows:

> 1. This completes our sketch of the moral psychology of the person. I stress that it is a moral psychology drawn from the political conception of justice as fairness. It is not a psychology originating in the science of human nature but rather a scheme of concepts and principles for expressing a certain political conception of the person and an ideal of citizenship. (*PL* 86–7)

It seems clear Rawls is now using the expression 'moral psychology' in a rather different way from that in which he had used it in *TJ*, as a way of capturing 'a certain political conception of the person and an ideal of citizenship'. One issue here is the importance of the qualification of this conception of the person as 'political', but I want to set this aside for the moment: one can find largely similar accounts of the conception of the person in Rawls's middle-period lectures 'Kantian Constructivism in Moral Theory' (KC),[6] without any qualification of this conception as political. I should add that in these lectures he does not make much use of the phrase 'moral psychology' to describe this conception of a person, but the phrase does occur at least once with this use (KC 346) and the substance and role here of the conception of a person is much the same as that which it plays in later writings such as *PL* where he routinely describes it as moral psychology.

[4] Rawls (1993).

[5] How, one wants to ask, can a 'psychology' not be 'psychological'?

[6] These lectures were originally published in 1980. They are reprinted in Rawls (1999*b*), and page references are to this edition. See pp. 330–3 for the account there of the conception of a person.

What, then, is the role and substance of this later conception of moral psychology? As the passage above from *PL* indicates, it is intended to capture the conception of a person which is central to moral and political theory. Indeed Rawls recognizes that different moral theories will bring with them different moral psychologies and he especially contrasts the 'sparse' moral psychology implicit in the 'rational intuitionism' of moral realists such as Moore and Ross (*PL* 92) with that which is central to his own neo-Kantian constructivism. Central to this latter psychology is the attribution to persons of two 'moral powers', a capacity for a sense of justice and a capacity for a conception of the good (*PL* 81). The first of these was central to the account of moral psychology in *TJ*, but the second, the capacity for a conception of the good, is not part of that account at all. It comprises both the fact that a person has certain values and final ends, things which they care about or aim at for their own sake and which bring with them a more or less explicit way of thinking about their relationships with others and the world, and also the fact that they have the ability to revise these values and ends in the light of new evidence or other reasons (*PL* 19). Rawls adds that, in addition to having these moral powers, a person should be conceived to have further dispositions which are 'aspects of their being reasonable and having this form of moral sensibility' (*PL* 81). These further dispositions include a 'readiness to propose and abide by fair terms of cooperation' provided one has reasonable assurance that others will also do their part, and a tendency to develop trust and confidence in others as the success of cooperative arrangements is sustained (*PL* 86).

Much of this latter material was in fact present in the account of the development of a sense of justice in *TJ*, so it is primarily the emphasis on the capacity for a conception of the good which marks a substantive addition to the content of his early account of our moral psychology.[7] This difference between *TJ* and his later writings shows that the fundamental difference between the two conceptions of moral psychology lies in their role in Rawls's presentation of his moral philosophy. As I have indicated, its role in his work from 'Kantian Constructivism in Moral Theory' onwards belongs to the description of the capacities and dispositions whose possession by persons is essential to the articulation of a moral theory, be it Rawls's own Kantian constructivism, Ross's rational intuitionism, or Mill's utilitarianism. Rawls

[7] In the revised edition of *TJ* Rawls does signal the importance of this capacity. In § 82 ('The Grounds for the Priority of Liberty') he describes the motivations of the parties in the hypothetical original position and remarks: 'The parties conceive of themselves as free persons who can revise and alter their final ends and who give priority to preserving their liberty in this respect' (*TJ* 475). Like most of § 82, however, this passage was not present in the first edition (as Rawls acknowledges in his 'Preface to the Revised Edition', p. xiii).

does of course also provide descriptions of these capacities and dispositions in *TJ*, especially in the first part of the book (e.g. § 25). But none of this material is here described as moral psychology; instead in *TJ* the 'principles of moral psychology' are the psychological laws which govern the development of our moral sentiments and explain the possession of a sense of justice. These principles are introduced specifically in order to help with the problem of stability (*TJ* 453; 397) because, Rawls thinks, if it is part of normal human psychology that moral sentiments such as a sense of justice develop within the context of life in a just society, then a normal person who is a citizen of a just society should be motivated to fulfil the requirements of justice, with the result that such a society should be reasonably stable. As we shall see below, Rawls thinks that there is more to be said on this matter: but on the face of it this 'stabilizing' role of moral psychology is quite different from its foundational role in *PL*. An easy way to bring out the difference here is to take the case of Rational Intuitionism. According to Rawls the sparse moral psychology implicit in Rational Intuitionism is primarily one which ascribes to persons a capacity for knowledge of moral principles and a capacity for motivation by this knowledge (*PL* 92). It is obvious that this moral psychology does little to show that it is in a person's interest to act in accordance with this motivation; but it was that task which was to be assisted by moral psychology in its stabilizing role.

In the light of this discussion I want to return to Rawls's characterization of moral psychology in *PL* as 'Philosophical not Psychological' (see the passage quoted above, *PL* 86). The contrast he draws here is one between moral psychology conceived as 'a scheme of concepts and principles for expressing a certain political conception of the person and an ideal of citizenship' and 'a psychology originating in the science of human nature'. Plainly the first of these conceptions identifies the foundational role of moral psychology in moral and political philosophy; whereas the second, which is disavowed, concerns an approach to the psychology of the moral sentiments which originates in a 'science of human nature'. This contrast is overtly drawn in terms of origins: philosophical versus scientific. But the question to which this contrast gives rise is whether a further contrast is implied, intentionally or not, between an a priori philosophical moral psychology and an empirical scientific psychology. That would seem to threaten an untenable dualism, reminiscent of Kant's distinction between the noumenal and empirical selves which Rawls would like to think he has discarded (*TJ* 256–7; 226). In fact it is clear that Rawls's contrast between a philosophical moral psychology and a scientific one is not intended to be exclusive: the contrast is fundamentally one of rationale, and Rawls explicitly affirms that it is a condition of any acceptable philosophical moral psychology that it be consistent with our natural capacities: 'Human nature and its natural psychology are permissive: they may limit the viable conceptions of persons and ideals of citizenships and the moral psychologies that may support them, but do not dictate the ones we must adopt' (*PL* 87).

This shows that Rawls is not a metaphysical dualist; however, it still leaves open a question as to the relationship between our 'natural psychology' and the favoured philosophical moral psychology which it 'permits' but does not 'dictate'. The principles of moral psychology propounded in *TJ* are supposed to be psychological laws which concern the development of moral sentiments and capacities such as a sense of justice which belong within Rawls's favoured philosophical moral psychology. So the picture we get there is one of an intimate explanatory relationship between natural and moral psychology. Admittedly, natural psychology does not by itself 'dictate' moral psychology so conceived, since the development of moral sentiments is contingent upon the moral character of the relationships and society in which the individuals concerned grow up and live. Nonetheless, given, to use Rawls's own phrase, 'The Connection between Moral and Natural Attitudes' (the title of § 74 of *TJ*), it follows that a complete understanding of human life, a true 'science of human nature' as one might put it, has to make room for our moral sentiments; for (to repeat a passage quoted earlier) 'The moral sentiments are a normal part of human life. One cannot do away with them without at the same time dismantling the natural attitudes as well' (*TJ* 489; 428). Hence Rawls's early work encourages the prospect of a unified explanatory approach to human psychology which embraces both natural and moral psychology.

On the face of it, this prospect is not sustained in Rawls's later writings, where he seems primarily concerned to put a distance between his philosophical moral psychology, the psychological assumptions inherent in his moral philosophy, and natural psychology, the empirical science of human nature, even if the latter has to 'permit' the former. But the issue is not clear: for Rawls remained concerned to provide a solution to the problem of stability and unless there are some substantive connections between the demands of morality, and thus our moral psychology, on the one hand, and our 'natural' psychology, on the other hand, the problem of stability will remain unsolved. For stability requires that, under normal circumstances, it is in our interest to be moral and our interests are rooted in our natural psychology. One thing that complicates discussion of this issue, however, is Rawls's development and refinement of his constructivist approach to moral and political philosophy. For if one takes it that that anything that merits the description 'natural' is to be discovered, not constructed, whereas moral principles are constructed and not discovered, it is going to be difficult to make substantive connections between the moral and the natural. Hence to take this issue forward, it is necessary to consider Rawls's constructivist approach to moral and political philosophy and the role of moral psychology in this context.

2. CONSTRUCTIVISM AND MORAL PHILOSOPHY

Rawls's approach to metaethics is set out in detail in his 1975 paper 'The Independence of Moral Theory' (IMT).[8] He here develops the brief remarks in *TJ* in which, alluding specifically to Quine, he rejects the application within moral philosophy of a methodology based on the analytic/synthetic distinction which would imply giving priority to questions of definition over substantive issues of principle (*TJ* 51, 578–9; 44–5, 506–7). If anything, Rawls urges, the priority runs in the other direction: just as the advances in logical theory and the theory of meaning due to the work of Frege, Russell, and others have profoundly transformed the philosophy of logic and language, in moral philosophy, he suggests, something similar may occur: 'Once the substantive content of moral conceptions is better understood, a similar transformation may occur. It is possible that convincing answers to questions of the meaning and justification of moral judgments can be found in no other way' (*TJ* 52; 45). So insofar as Rawls has a metaethical perspective in *TJ*, it is a bottom–up rather than a top–down approach that he favours, and this thesis is explicitly affirmed in 'The Independence of Moral Theory':

> A relation of methodological priority does not hold, I believe, between the theory of meaning, epistemology, and the philosophy of mind on the one hand and moral philosophy on the other. To the contrary: a central part of moral philosophy is what I have called moral theory; it consists in the comparative study of moral conceptions, which is, in large part, independent. (IMT 301)

An important element of this bottom–up approach is a willingness to engage with psychology: for psychology lies at the heart of moral theory since such a theory aims to provide

> a deeper understanding of the structure of the moral conceptions and of their connections with human sensibility ... We must not turn away from this task because much of it may appear to belong to psychology or social theory and not to philosophy. For the fact is that others are not prompted by philosophical inclination to pursue moral theory; yet this motivation is essential, for without it the inquiry has the wrong focus. (IMT 302)

[8] Reprinted in Rawls (1999*b*); page references are to this edition.

Thus moral theory has to include a psychological inquiry ('without it the inquiry has the wrong focus'); and so far from the resulting moral psychology being dependent upon philosophy of mind, the dependence runs, if anything, in the other direction. Although by and large abstract philosophical debates about mind and body do not intersect with moral theory, where there are connections, as between natural attitudes and moral sentiments, philosophy of mind has to accommodate itself to moral psychology, to our having a psychology which is not wholly value-free.

Rawls's doctrine of 'Kantian constructivism' in moral theory is to be understood in the light of this bottom–up approach to moral philosophy. The position is not an application of any more general metaphysical or epistemological doctrine concerning truth:

> A constructivist view does not require an idealist or a verificationist, as opposed to a realist, account of truth. Whatever the nature of truth in the case of general beliefs about human nature and how society works, a constructivist moral doctrine requires a distinct procedure of construction to identify the first principles of justice. (KC 351–2)

Instead Rawls's constructivism is grounded in his moral theory. The central claim of this theory is that morality is a way of achieving autonomy, a life which combines respect for individual freedom, especially our status as 'self-authenticating sources of valid claims' (*PL* 72), with recognition of our essential dependence upon others who have equal status, a dependence which is not merely practical but such that we can normally realize our own conception of the good only through co-operative activities with others ('the self is realised in the activities of many selves'– *TJ* 565; 495). This involvement with others necessitates compliance with principles for social cooperation, and these principles count as moral principles only insofar as they can be viewed as principles which we and others would choose to impose upon ourselves because there are reasons for them which respect 'our status as free and equal moral persons' as Rawls frequently puts it (e.g. *PL* 19). Thus by internalizing the fact of our essential dependence upon others we recognize the social requirements of this interdependence as moral principles whose application to us is not a limitation of our autonomy, but a condition of it. So, according to this way of thinking, the substance of morality is not constituted by a set of moral facts which are available to be discovered in the social world; instead it is to be thought of as 'constructed' through agreements made in accordance with an idealized procedure for regulating social cooperation which represents both the equal status of the parties involved and the reasons which favour or oppose different policies. In *TJ* Rawls famously describes this procedure in terms of a hypothetical contract to be agreed under the conditions of an 'original position', but I shall not discuss this matter here. My interest lies in the underlying moral psychology and its connection with his general constructivist approach.

The account above implies that the connection is very close. For, as Rawls puts it, Kantian constructivism is the doctrine that the principles which define the requirements of morality are to be 'viewed as specified by a procedure of construction ... the form and structure of which mirrors both of our two powers of practical reason',[9] that is, our capacity for a conception of the good and our sense of justice. Equally, our essential dependence upon others is not just a matter of the need for practical cooperation under conditions of potential relative scarcity; instead it expresses a deep psychological truth about the conditions for personal self-realization, namely that it is dependent upon mutual interaction and appreciation by others. Thus it is our moral psychology, combined with the recognition that others with the same moral psychology have equal moral status, which sets the constraints within which the construction of morality is conceived. But it is important to note that neither our moral psychology itself nor its status is here thought of as constructed in the same way. Instead our 'two powers of practical reason' owe their status as moral powers to the ways in which they express the requirements of practical reason: the capacity for a conception of the good expresses our rationality, since it is the ability 'to form, to revise, and rationally to pursue a conception of one's rational advantage or good' (*PL* 19); and a sense of justice expresses our willingness to be reasonable, since it is the capacity 'to understand, to apply, and to act from the public conception of justice which characterizes the fair terms of social cooperation' (*PL* 19). So Rawls's constructivism in moral and political theory is founded upon a moral psychology whose status is not constructed at all but is explicated in terms of its role as the expression of practical reason.

In his later writings Rawls often compares his constructivist doctrine with the position of the rational intuitionist who holds that fundamental moral principles are discovered through a capacity for intuitive insight, and it is worth looking at the way in which Rawls makes this comparison. The familiar objections to rational intuitionism, such as Mackie's 'argument from queerness',[10] appeal to general metaphysical and epistemological considerations; in the light of Rawls's metaethical stance, however, it is not surprising that this is not the way in which he argues against the rational intuitionist. Instead his argument is rooted in his moral theory, in the importance of framing a conception of morality whereby the practice of morality can be seen to be a way of achieving autonomy, the expression of one's nature as 'free and equal', and not as a way of fulfilling requirements

[9] Rawls (2000: 237).

[10] Mackie (1977: 38 ff).

which, because they are external to one's own reasons for action, are 'heteronomous'. Rawls puts the point in the following way:

> Yet it suffices for heteronomy that these [first] principles obtain in virtue of relations among objects the nature of which is not affected or determined by the conception of the person. Kant's idea of autonomy requires that there exist no such order of given objects determining the first principles of right and justice among free and equal moral persons. Heteronomy obtains not only when first principles are fixed by the special psychological constitution of human nature, as in Hume, but also when they are fixed by an order of universals or concepts grasped by rational intuition, as in Plato's realm of forms. (KC 345)

Thus Rawls takes it that rational intuitionism is essentially a secular version of a Divine Command theory and that moral demands cannot be in this way altogether independent of us: morality can secure its authority over us only by answering to our nature as free rational beings. This criticism is associated with further points. Although Rawls agrees with the intuitionist that morality aims to be objective, he rejects the intuitionist's inference that moral principles purport to be true since he takes it that truth and facts are inseparable and he denies that there are any moral facts because 'the idea of constructing facts seems incoherent' (*PL* 122). This point is associated with a disagreement concerning the relationship between morality and reason: whereas the intuitionist takes it that moral judgment is an exercise of theoretical reason, for the constructivist it is based upon practical reason, as expressed through our capacities for rationality and reasonableness. As a result, there is an important difference with respect to moral psychology: for the intuitionist, moral theory implies only that our psychology includes a capacity for intuitive moral insight and for motivation by the knowledge thus acquired; but because constructivism relies on the capacities through which practical reason is expressed as a basis for the construction of moral principles it involves the richer moral psychology exemplified by Rawls's account of our two fundamental moral powers, our capacity for a conception of the good and our sense of justice (*PL* 93).

In the passage quoted above Rawls also criticized Hume's moral theory ('Heteronomy obtains not only when first principles are fixed by the special psychological constitution of human nature, as in Hume'), and it is worth setting Rawls's criticisms of Hume alongside his discussion of rational intuitionism, since there is a sense in which he conceives of his Kantian constructivism as a mean between these two positions. As we have just seen, Rawls complains that his position is a kind of heteronomy. What Rawls seems to have in mind here is that Hume treats our moral sentiments as just special cases of more general sentiments whose function and application can be understood without reference to any moral concepts. Rawls sums up his complaint as follows: 'What is distinctive of his view is that it seems to be purely psychological and to lack altogether what some writers think of as the ideas of practical reason and its authority.'[11] I shall not pursue the question of the justice of this verdict; what interests me here is the way in which Rawls damns Hume's moral theory for being too dependent on psychology while equally insisting, as against the Rational Intuitionist, on the importance of moral psychology within his Kantian moral theory. The key issue here seems to be one of reduction. For Rawls what is fundamental to moral philosophy is a conception of us as persons with the two fundamental moral powers, a capacity for a conception of the good and a sense of justice, which are in different ways exercises in practical reason. Moral theory is then supposed to show how putative moral principles are justified by explaining how they belong to the normative framework which would be adopted by a community of persons with these powers. Since the notion of a moral person with these powers is here taken to be fundamental, our moral psychology is here taken not to be derivable from more general psychological capacities and dispositions in the way that, according to Rawls, Hume seeks to achieve. Nonetheless, as we have seen earlier, Rawls himself acknowledges that our moral psychology is connected to our 'natural attitudes' in such a way that, where social conditions are appropriate, natural attitudes develop into moral sentiments. So there is a delicate issue as to how this relationship is understood—how, at the individual level there can be a relation of non-reductive dependence of moral psychology upon natural psychology. I shall come back to this in the next section; but before closing this section of the paper there is one further issue to be addressed.

Suppose a contemporary ethical non-cognitivist were to agree with Rawls in rejecting the thesis that moral attitudes are reducible to non-moral ones: would that commit such a person to being a Rawlsian Kantian? Surely not! For such a non-cognitivist, while agreeing with Rawls that moral judgments express our fundamental and irreducible moral psychology, would deny that there is any objectivity to be constructed on this basis. The disagreement here would be centred on what Rawls referred to, in the passage quoted above, as 'the ideas of practical reason and its authority'. The implication of 'Kantian Constructivism in Moral Theory' is that where the conception of a person includes the exercise of practical reason an objective morality can indeed be constructed on this basis. For Rawls the main part of this construction is, of course, 'justice as fairness'; but at that time, and earlier, Rawls took the view that this was just part of a broader construction of 'rightness as fairness' (see especially *TJ* § 18). The non-cognitivist will, in turn, have

[11] Rawls (2000: 50).

reasons for questioning this construction, and it is then clear enough where the main locus of this disagreement lies—namely on whether, setting aside doubts about the details of Rawls's moral and political constructions, Rawls made it plausible to hold that some such construction can be shown to be capable of objectivity, in the following sense: 'If, on the other hand, such a construction does yield the first principles of a conception of justice that matches more accurately than other views our considered convictions in general and wide reflective equilibrium, then constructivism would seem to provide a suitable basis for objectivity' (KC 354).

In *A Theory of Justice* Rawls appears confident that objectivity in this sense in attainable (*TJ* 517; 453). But in 'Kantian Constructivism' Rawls is more doubtful:

> Of course, this is conjecture, intended only to indicate that constructivism is compatible with there being, in fact, only one most reasonable conception of justice, and therefore that constructivism is compatible with objectivism in this sense. However, constructivism does not presuppose that this is the case, and it may turn out that, for us, there exists no reasonable and workable conception of justice at all. This would mean that the practical task of political philosophy is doomed to failure. (KC 355–6)

On this issue, notoriously, Rawls's doubts continued to grow, at least with respect to the construction of a demonstrably objective system of morality. For he came to think that once one considers dispassionately the variety of ethical systems, including the major religions, one should acknowledge that the resolution of fundamental questions of value is underdetermined by reasonable considerations and accept the 'fact of reasonable pluralism' (*PL* 60–1). Hence in his later writings he accepts that the objectivity of morality in general is doubtful—a matter of faith rather than reasonable belief. Nonetheless Rawls retained the view that within the sphere of public political morality agreement concerning principles of justice among citizens who are both rational and reasonable is attainable; and with this he takes it that objectivity with respect to these principles is defensible (KC 115–16).

3. MORAL DEVELOPMENT

Having explored the foundational role of moral psychology in the context of Rawls's constructivist moral philosophy, I want to return to Rawls's early account of the 'principles of moral psychology', in particular his account of the development of a sense of justice. Rawls himself does not show how these two aspects of his conception of moral psychology fit together; but it is important to consider this matter since, as I mentioned in the previous section, it is not clear how Rawls can maintain that the sense of justice is the product of a normal process of development from natural attitudes without slipping into the Humean 'psychological naturalism' he rejects.

In chapter 8 of *A Theory of Justice* Rawls argues that our sense of justice is the natural outcome of a deep tendency to reciprocity in our psychological constitution. Very briefly, Rawls's account runs as follows: initially, within the family, a young child develops the self-confidence which gives it a capacity for affection and friendship through growing up in an environment in which it feels secure in the love and care of its parents. This then helps the growing child to form friendly relationships outside the family and through these relationships the child develops a capacity for trust and responsibility as it is itself treated in these ways by others. Finally as a young adult it internalizes the requirements of justice as adherence to general principles through the experience of being treated with respect and fairness by others with whom it has no special friendship or relationship. This whole approach suggests that justice itself is a kind of reciprocity; for it is by thinking of justice in this way that one can see how the limited reciprocity of the first two stages becomes a reciprocal disposition which is a 'sense of justice' when it is extended to apply to relationships with just anyone—typically fellow citizens, but in principle strangers as well. Rawls's most famous early paper was 'Justice as Fairness' (1958) and throughout his life he used the phrase 'justice as fairness' to describe his conception of justice. But in his 1971 paper 'Justice as Reciprocity' he in fact distinguishes justice from fairness and argues that their common foundation is reciprocity, which is therefore more fundamental to justice than fairness itself: 'It is this requirement of the possibility of mutual acknowledgment of principles by free and equal persons who have not [*sic*] authority over one another which makes the concept of reciprocity fundamental to both justice and fairness.'[12]

The first issue to be addressed concerning this three-stage development is Rawls's observation, mentioned earlier, that the account 'refers to an institutional setting as being just', so that 'some view of justice enters into the explanation of the corresponding sentiment'—the sense of justice (*TJ* 491; 430). This claim has to be understood in the context of his constructivism, so that the explanatory role here of justice is not that of a distinctive fact which structures the context in which this development takes place. For, according to Rawls, there are no such moral facts. Instead Rawls's position must be that this development is accomplished in the

[12] 'Justice as Reciprocity' as reprinted in Rawls (1999*b*) 209. I am indebted to Patricia Greenspan for directing me to this important paper.

context of relationships informed by ties of personal affection and loyalty through which we come to see the point of fair practices which underpin the relationships from which we benefit, such as growing up within a family, sharing a house with a group of friends, or living as a citizen in a well-ordered society. Hence for Rawls it is by thinking of moral principles as principles setting terms for cooperation which would be agreed by free and equal persons that we should understand the role of social practices and institutions which incorporate these principles as providing the context for the psychological developments which issue in a sense of justice.

In *A Theory of Justice* Rawls suggests that we should think of these developments as based upon transformations in the kinds of desire that we have:

> The three laws describe how our system of desires comes to have new final ends as we acquire affective ties. These changes are to be distinguished from our forming derivative desires … [Instead the laws] characterize transformations of our patterns of final ends that arise from our recognizing the manner in which institutions and the actions of others affect our good. (*TJ* 494; 432)

In his later writings Rawls writes of our capacity for 'principle-dependent' and 'conception-dependent' desires, as opposed to ordinary 'object-dependent' desires (*PL* 82 ff.), and this offers a more detailed way of thinking about his account of moral development. Object-dependent desires are desires whose objects are personal goods whose characterization as such relies on no moral or other normative principle; by contrast specifying the objects of principle-dependent desires such as fidelity involves moral principles, and similarly specifying the objects of conception-dependent desires involves moral ideals such as citizenship. Can one match these three types of desire with the three stages? The match is easy to see at stages two and three: the growing child who becomes trustworthy through shared activities with friends can be thought of as someone who begins to develop principle-dependent desires; and for Rawls the final development of a sense of justice is accomplished as one identifies oneself as a member of a potentially well-ordered society in which one can aspire to the ideal of citizenship. The first stage, however, is not to be thought of as that in which one becomes susceptible to object-dependent desires, since these are all too prevalent anyway. Instead what is important at this first stage is, I think, the development of a capacity to care about others, to make their good one's own good, since this is a prerequisite of the capacity for friendship which enters into Rawls's second stage. Initially, of course, the others in question are those who care about one themselves, most notably the members of one's family. So, at least for the purpose of completing the match between Rawls's three-stage account of moral development and his three-way hierarchy of desires, Rawls's hierarchy of desires needs to be augmented by inserting a category of 'relationship-dependent desires' between the object-dependent and the principle-dependent desires. Object-dependent desires come for free, and do not mark a significant stage of moral development; and the first stage of development requires, not principle-dependent desires, but those which involve concern for others to whom the subject, typically a child, is connected by an affectionate relationship such as the love between members of a family.

In setting things out in this way I have been trying to explore the way in which a Rawlsian moral psychology might be thought to work. The sketch of moral development above does, I think, meet the twin requirements of neither tacitly drawing on rational intuitionism, the intuitive appreciation of moral truths as such, nor turning out to be a form of reductive naturalism which derives moral sentiments from non-moral natural attitudes. For although the account of moral development involves a hierarchy of desires, starting from non-moral object-dependent desires, the progression is achieved though transformations in which the subject's motivational set is thought of as enhanced as a result of including concerns, principles and ideals which the subject recognizes as informing relationships, practices and institutions that are essential to his own good as he grasps a broader sense of his own identity as a member of groups in which his own good is dependent upon that of others and vice-versa. Thus these transformations exemplify Rawls's belief that the moral psychology appropriate to Kantian constructivism is founded upon the exercise of practical reason, understood as a rational appreciation of one's own good and a willingness to cooperate with others on a reasonable basis.

4. THE PROBLEM OF STABILITY

So far, then, so good for this sketch of the way a Rawlsian moral psychology might work. But as I indicated at the start an important further consideration comes from the role that this psychology is supposed to fulfil in Rawls's theory of justice by contributing to a solution to the problem of stability. This is the problem of showing that a state whose political institutions are just will be reasonably stable in the sense that such a state need not rely primarily on coercion to ensure its citizens obey the law and support the state's institutions because, for the most part, these institutions and laws enjoy the support of the citizens anyway. The contribution of moral psychology to the solution of this problem was supposed to be that it would show how the requirements of justice are broadly congruent with the interests of individual citizens, so that, as Rawls put it, 'being a good person is a good thing for that person' (*TJ* 397; 349); and this requires, as one might put it, that the moral psychology which makes a person 'a good person' be congruent with the natural, normal, psychology which identifies what is 'a good thing for that person'.

Rawls's first thought, in his 1963 paper 'The Sense of Justice', was that merely by showing that the development of a sense of justice is the normal psychological outcome of life within a society whose institutions are broadly just one shows that such a society is stable.[13] For there is a virtuous circle whereby institutions and moral sentiments reinforce each other. In *TJ* Rawls starts by repeating this line of thought in chapter 8, arguing that the moral psychology of his conception of justice as fairness is more conducive to stability than, say, the psychology that one would associate with a utilitarian conception of justice (*TJ* § 76 'The Problem of Relative Stability'). But he does not treat this comparison as the end of the matter, since only a few pages later, at the start of chapter 9, he says that only now is he in a position to deal properly with the task of showing that justice and goodness are congruent (*TJ* 513; 450). As he acknowledged later, commenting on the part of *A Theory of Justice* which includes these chapters, it is not clear what is going on here: 'Throughout Part III too many connections are left for the reader to make, so that one may be left in doubt as to the point of much of chs. 8 and 9'.[14] It appears nonetheless that in *A Theory of Justice* Rawls believed that he could not simply rely on the normal development of a sense of justice to vindicate the congruence thesis for the reason that this account of the development of a sense of justice was primarily causal and not normative: it showed how one would expect a sense of justice to be inculcated among those growing up in a just society; but it did not thereby show that it was good for them to have this motivation which might be just a 'neurotic compulsion' (*TJ* 514; 451). And without a demonstration of this, the problem of stability was not fully resolved, since if it remained an open question whether acting justly was in general good for one, one could not reliably expect people to obey the laws of a just state despite their having a sense of justice.

I think that Rawls was right about this; but that Rawls's own favoured account of the congruence of individual good and morality is unpersuasive. He argues that because 'the desire to act justly and the desire to express our nature as free moral persons turn out to specify what is practically speaking the same desire' (*TJ* 572; 501) fulfilment of this desire promotes one's own good. This thesis assumes the dominance of his Kantian moral psychology ('the desire to express our nature as free moral persons') among

[13] 'The Sense of Justice', in Rawls (1999*b*: 105).

[14] 'Justice as Fairness: Political not Metaphysical', in Rawls (1999*b*: 414 n. 33).

the ends which determine one's individual good. Yet the practical doubt which motivates the search for congruence can be readily redirected at the question as to whether the ends intrinsic to this Kantian moral psychology are indeed central for one's own individual good. In chapter 7 of *A Theory of Justice* Rawls had set out a complex account of goodness, which includes an account of a person's good as 'the successful execution of a rational plan for life' (*TJ* 433; 380).[15] So the congruence thesis is the thesis that the successful execution of a rational plan for life normally requires one to act in accordance with the principles of justice. Rawls seems to have thought that the fact that this latter requirement is tantamount to acting in accordance with the motivations specified by his Kantian moral psychology as 'expressing one's nature as a free person' *ipso facto* shows that it contributes to the successful execution of one's rational plan for life. But as it stands this is not persuasive: for why should the aim of 'expressing one's nature as a free person' be given a central position in one's individual plan for life? The introduction here of what in *A Theory of Justice* he calls the 'Kantian interpretation' (§ 40) of moral psychology does not by itself suffice to make a connection between adherence to the principles of justice and individual good. If Rawls were to introduce a Kantian dualism of noumenal and empirical selves, it would be plausible to hold that for the noumenal self individual good and the moral life are inseparable; but Rawls explicitly rejects any such dualism, and, anyway, it would leave the problem completely insoluble as far as empirical selves are concerned, which is where it matters most.

Notoriously, there is a sense in which Rawls himself came to agree about this. For in the 1992 'Introduction' to *Political Liberalism* (esp. pp. xvi–xvii) he explains that he himself came to see that his Kantian argument for the congruence of goodness and justice was unsatisfactory as a general solution to the problem of stability since his Kantian moral theory was but one of several reasonable comprehensive moral theories. As a result, he inferred, no one moral theory can be employed in a political philosophy which aims to provide arguments that will be persuasive for all reasonable citizens, and it had been a mistake to rely on the Kantian theory alone to solve the problem of stability. This last move is not unchallengeable: one might think that insofar as the problem is not solved by explaining how a just society nurtures a sense of justice among its citizens, the problem of stability is essentially theoretical and thus that one cannot expect to avoid drawing on one's moral theory to resolve it—even while recognizing that there are other 'reasonable moral theories' which will promote different solutions.

[15] So Rawls combines a broadly naturalistic account of goodness with his constructivist account of morality.

But Rawls thinks that a different line of thought is available: if one can show that there is a conception of justice which expresses the reasonable political aspirations of adherents of different moral theories who acknowledge the fact of reasonable pluralism, one will thereby be in a position to show that a state which realizes this conception should be stable, since 'it can win its support by addressing each citizen's reason, as explained within its own framework' (*PL* 143).

To follow this line of thought would take me well away from moral psychology. So having noted how the problem of stability continues in Rawls's later writings, I want to return to the problem of stability as he left it in *A Theory of Justice*. For in his discussion there he introduced some suggestive ideas which remain largely unexploited by him but which, I think, can be used not only to provide a better solution to the problem of stability than the approach he favoured but also to enrich his account of moral psychology.

5. SELF-RESPECT AND MUTUAL RESPECT

The line of thought I have in mind is that which Rawls introduces in chapter 9 of *A Theory of Justice*, starting with the conception of a 'social union' which he takes from Humboldt (*TJ* 523–5; 459–60: esp. n. 4). A social union is a collective institution whose members cooperate in a type of joint activity in order to achieve valuable ends which they cannot bring about without such cooperation. Rawls gives the example of an orchestra as a social union of this kind: for it is only within an orchestra which brings together musicians of many different kinds that the individual musicians can take part in performing great orchestral works. In cases of this kind, he writes, 'persons need one another since it is only in active cooperation with others that one's powers reach fruition. Only in a social union is the individual complete' (*TJ* 524–5; 460 n. 4). The existence of social unions shows us something important about the way in which individuals with different abilities need to collaborate with each other in order to achieve valuable ends, and Rawls infers that a just state can itself be regarded as a social union, a 'social union of social unions' (*TJ* 527; 462). This involves more than just the familiar thesis that political cooperation is essential for the achievement of individual goods: instead, if the social union model is to be applicable there has to be some collective good comparable to the performance of a symphony which is not available without the collective participation of the citizens who are members of this social union, the just state, and which is central to the ends of each individual citizen. Rawls suggests that the collective good is just 'the public realization of justice' which meets the requirement of providing a distinctive form of self-fulfilment because 'the collective activity of justice is the pre-eminent form of human flourishing' (*TJ* 529; 463).

This suggestion is not plausible. There are indefinitely many forms of human flourishing, arising from the great variety of individual plans for life, and although 'the collective activity of justice' is an ingredient in many worthwhile ends, this is no reason to give it pre-eminence as an end itself; indeed it is questionable whether it really makes sense to regard 'the collective activity of justice' as a form of human flourishing. What is going wrong here is that the social union model for the state does not really work for Rawls: the Rawlsian liberal state is not comparable to an orchestra, an institution whose members rely on each other's complementary activities to accomplish an essentially collective goal such as the performance of a symphony. Rawls's suggestion that 'the public realization of justice' counts as such an end, for example, is unpersuasive. For although individual citizens are of course required to be just, this is not a collective activity on their part and it is primarily the responsibility of public authorities to maintain justice in general. If Rawls were to hold with the communitarians that the state is a collective association with some dominant goal that supposedly meets the requirement of providing self-fulfilment for all citizens, such as the establishment of a classless society, he could use the social union model for the purposes of his congruence thesis. But, of course, that is exactly not the way in which Rawls conceives of his liberal state.[16]

Yet one should not for this reason dismiss altogether all of the themes that enter into Rawls's discussion of the idea of a social union, in particular the suggestion that

> the members of a community participate in one another's nature: we appreciate what others do as things which we might have done which they do for us, and what we do is similarly done for them. Since the self is realized in the activities of many selves, relations of justice that would be assented to by all are best fitted to express the nature of each. (*TJ* 565; 495)

Rawls's line of thought here is reminiscent of the kind of reciprocity that came up earlier in connection with his account of our moral development, the three-stage development of a sense of justice via the place of love in 'the morality of authority' and that of trust in 'the morality of association'. Rawls never connects this conception of reciprocity that is central to his early moral psychology with his later discussion of congruence; but I want to propose that there are connections to be made here which enable one to fill out both his moral theory and his moral psychology. The place to start

[16] Rawls's rejection of civic humanism is especially notable in this context: see *PL* 206.

is with the good which is for Rawls of primary importance: self-respect. He writes

> It is clear then why self-respect is a primary good. Without it nothing may seem worth doing, or if some things have value for us, we lack the will to strive for them. All desire and activity becomes empty and vain, and we sink into apathy and cynicism. Therefore the parties in the original position would wish to avoid at almost any cost the social conditions that undermine self-respect. (*TJ* 440; 386)

Suppose we now apply to self-respect the developmental approach involving reciprocity from Rawls's moral psychology, so that self-respect is held to be dependent upon respect by others whom one respects oneself. It now becomes easy to argue for the congruence of justice and individual good. The argument starts from Rawls's thesis about the value of self-respect:

(i) Any rational plan for life will acknowledge that self-respect is a primary good.

Add my proposal about the dependence of self-respect on respect by others whom one respects oneself:

(ii) The achievement of self-respect is dependent upon reciprocal relationships of mutual respect.

Now add a Rawlsian thesis about justice as reciprocity:

(iii) The conception of justice as reciprocity is the conception of principles whose institutional realisation would affirm the mutual respect of citizens for each other.

It does now follow that

(iv) Any rational plan for life will bring with it a 'conception-based desire' to living in accordance with justice, at least in a well-ordered society.

The crucial claim of this argument is (ii), that self-respect is dependent upon reciprocal relationships of mutual respect. Rawls himself endorses a thought of this kind when he writes that our sense of our own worth is supported by 'finding our person and deeds appreciated and confirmed by others who are likewise esteemed' (*TJ* 440; 386); and he goes on to infer something comparable to thesis (iii) from this, to the effect that the realization of his principles of justice provides 'background conditions' which ensure that 'in public life citizens respect one another's ends' (*TJ* 442; 388). But although he here (*TJ* 442; 388) intimates that he will return to this thesis in his subsequent discussion of the idea of a social union, in that context he does not in fact make any significant use of it. Instead he advances the idea of the state as a social union of social unions, a proposal which, as I have explained, he cannot adequately substantiate. Yet although the argument from (i) to (iv) is not manifest in *A Theory of Justice* (or elsewhere), it would, I think, be congenial to Rawls.

In thinking about the crucial claim (ii) it is important first to clarify what self-respect amounts to. As the passages cited above indicate, in *A Theory of Justice* Rawls treated self-respect and self-esteem as interchangeable, but this is readily seen not to be correct when one thinks of the difference between behaviour which shows a lack of esteem for someone's work and that which shows a lack of respect for them.[17] To treat someone with a lack of respect is, I take it, to fail to acknowledge their status as (in Rawls's words) a 'self-authenticating source of valid claims', whereas a lack of esteem for someone simply expresses the judgment that their life and work is not especially valuable. In this sense, therefore, self-respect is consciousness of oneself as a self-authenticating source of valid claims, as someone who merits treatment with respect by others; whereas self-esteem is the judgment that one's life includes valuable achievements that are worthy of esteem by others. I take it that both self-respect and self-esteem are important goods. Rawls's description (quoted above) of the situation of someone who lacks self-respect in fact applies best to the case of someone who lacks self-esteem: for someone who lacks self-respect is not so much someone who thinks that nothing is worth doing as someone who thinks that he is worthless, someone whose interests count for nothing. No doubt these two conditions are closely associated: self-esteem, I think, presupposes self-respect, though the converse implication need not obtain (someone who is excessively modest lacks self-esteem but not necessarily self-respect). But it is important for the purposes of the current argument to distinguish them; for (iii) is only plausible when interpreted as a claim about self-respect, properly understood. It is not a requirement of justice that people should esteem each other's life and work.

The key issue, therefore, is whether (ii) is also plausible when self-respect is interpreted as consciousness of oneself as a source of valid claims on others. Where the dependence affirmed by (ii) is understood as a case of normal reciprocal psychological development, of the kind characterized by Rawls in his description in *A Theory of Justice* of the 'principles of moral psychology', (ii) certainly looks to be plausible; indeed it is surely integral to the moral development Rawls describes. So, understood in this way, (ii) is as robust as the rest of Rawls's early moral psychology. But, as we have seen, Rawls hoped that the congruence thesis could be established in a way which did not just rely on the normal course of human psychological

[17] See David Sachs, 'How to Distinguish Self-Respect from Self-Esteem', *Philosophy and Public Affairs* 10/4 (autumn, 1981), 346–60.

development. So the question is whether there is a way of strengthening the form of dependence in (ii).

One can envisage a stronger way of interpreting (ii), as affirming that self-respect constitutively requires mutually interpenetrating attitudes of respect such that one recognizes that one is respected by others whom one respects oneself. To take this view of self-respect would be to model it on Hegel's famous thesis concerning self-consciousness, that self-consciousness is dependent upon the consciousness of one by others of whom one is oneself conscious: 'Self-consciousness exists in and for itself when, and by the fact that, it so exists for another; that is, it exists only in being acknowledged' (*Phenomenology of Spirit* § 178). where this 'acknowledgment' takes the form of mutual 'recognition: 'They *recognize* themselves as *mutually recognizing* each other (*Phenomenology of Spirit* § 184). Just what Hegel's conception of self-consciousness amounts to is notoriously obscure and disputed, and I shall not attempt to elucidate it; what matters for us is whether this thesis holds for self-respect.[18] On the face of it it is vulnerable to counter-examples: a good case to think about is that of Olaudah Equiano, the remarkable slave whose autobiography *The Interesting Narrative of the Life of Olaudah Equiano* clearly shows how he maintained his self-respect in the face of a failure of recognition by others who bought and sold him as a slave. This case shows that a straightforward Hegelian interpretation of (ii) is too strong; and of course without (ii), the route via (iii) to (iv), the congruence thesis, is broken. But there is a way around this, by taking it that self-respect is to be understood precisely in such a way that (ii) is true of it—that is, by taking it to be the kind of publicly affirmed self-respect in which one's sense of one's own worth is confirmed and strengthened through recognition by others whom one respects. Olaudah Equiano did not enjoy this kind of self-respect until he was able to buy his way out of slavery and work with others for the abolition of slavery; but there is every reason to think that this change in his self-consciousness was a change of great value to him. For once his situation had changed his own sense of himself as a 'self-authenticating source of valid claims' was at last confirmed by the recognition of the validity of these claims by others whose similar status he himself recognized. So even though his initial form of private self-respect was of great value to him, the primary social good in this area is the kind of publicly confirmed self-respect which satisfies condition (ii);

[18] In the original version of this paper I tried to use the position presented by Axel Honneth in *The Struggle for Recognition* (Honneth 1995) to develop this line of thought since in some respects his position resembles that advanced by Rawls. But discussion with Carla Bagnoli has persuaded me that it is both unnecessary and confusing to introduce Honneth's position, interesting though it is.

hence proposition (i) holds for this form of self-respect, and since (iii) is plainly also in play, the route to (iv) and the congruence of justice and individual good, is secured.

It may be felt that there is a trick here, in that self-respect has been just defined to be a condition which depends on social recognition. One response to this is to observe that as long as this form of self-respect is agreed to be a primary social good it does not matter that there is another form of self-respect which is not in the same way dependent on recognition. But there is a deeper point here. I have characterized self-respect in Rawlsian terms as consciousness of one's freedom since he takes it that freedom is primarily a matter being a self-authenticating source of valid claims (*PL* 72). For Rawls this freedom depends on one's moral powers, primarily the capacity to be guided by one's conception of the good and to revise this conception in the light of evidence. So freedom is an implication of the conception of a person that is characteristic of Kantian moral psychology, as the case of Olaudah Equiano indicates, since he certainly possessed the relevant moral powers even when he was a slave. The Hegelian move is then to suggest that the consciousness of freedom that comes with self-respect takes us beyond moral psychology because it involves recognition by others. Equiano's case shows that the necessity for this transformation is questionable: in his case self-respect did not, initially, involve recognition by others. But it is important to note that the 'claims' whose self-authenticating validity is affirmed in the attribution of freedom to a person are claims directed at others, with the presumption that their validity is to be recognized by them. So even within Rawls's conception of freedom there is a presumption of recognition by others; and it is this presumption which is then made explicit in the Hegelian account of self-respect as a consciousness of freedom which is dependent upon recognition by others. As we have seen, this suggestion needs qualification; there can be a form of self-respect which is not dependent on actual recognition by others. But since this form of self-respect still makes a claim to recognition and respect by them, there is every reason to think that it is better to enjoy the socially confirmed form of self-respect than the private consciousness of freedom which was Equiano's lot for most of his life. Hence the priority given to the socially confirmed form of self-respect is not a dialectical trick, but is inherent in the conception of self-respect itself.

REFERENCES

Honneth, A. (1995) *The Struggle for Recognition* (Cambridge: Polity).

Kukathas, C. (ed.) (2002) *John Rawls (Critical Assessments)* 4 vols. (London: Routledge).

Mackie, J. L. (1977) *Ethics: Inventing Right and Wrong* (Harmondsworth: Penguin).

Rawls, J. (1971) *A Theory of Justice* (Cambridge, MA: Harvard University Press).

—— (1993) *Political Liberalism* (New York: Columbia University Press).

—— (1999*a*) *A Theory of Justice*, rev. edn (Oxford: Oxford University Press).

—— (1999*b*) *Collected Papers*, ed. S. Freeman (Cambridge, MA: Harvard University Press).

—— (2000) *Lectures on the History of Moral Philosophy*, ed. B. Herman (Cambridge, MA: Harvard University Press).

—— (2001) *Justice as Fairness* (Cambridge, MA: Harvard University Press).

10

Actions, Acting, and Acting Well

Matthew Hanser

Philosophers wishing to understand moral judgments typically focus their attention upon the evaluative or normative predicates that these judgments employ.[1] How is the peculiar force of such predicates to be understood? What are the criteria for their application? I propose to come at matters from another direction. What are the *objects* of evaluation in this or that sort of judgment? To what do our evaluative or normative predicates apply? The guiding idea behind the inquiry is that we cannot properly understand evaluations of a given sort unless we know what they are evaluations of. Modes of evaluation must suit their objects. Of course one might think that it's generally pretty obvious what's being evaluated. In what follows I hope to show that this is not so, and that the question what's being evaluated has significant consequences for moral theory.

1

There are many kinds of moral judgment. In this article I isolate, through a series of steps, the class of judgments that will constitute my target in this article.

First, I shall restrict my attention to judgments directly concerning behavior. I mean this formulation to be somewhat vague both with respect

Earlier versions of this paper were presented at the Third Annual Metaethics Workshop in Madison, Wisconsin; at the University of California, Davis; and at the 2007 Pacific Division meetings of the American Philosophical Association, where the commentator was Andrew Eshleman. I am grateful to Eshleman, and to members of all three audiences, for their questions, comments, and criticisms. I am especially grateful to Mark Schroeder and Adam Sennet for giving me such a hard time.

[1] I follow custom in speaking of moral *judgments*, but my focus will be upon the (English) sentences used to make such judgments, the propositions or Fregean thoughts that they express, and the states of affairs that they describe.

to what counts as behavior and with respect to what it is for a judgment directly to concern behavior. As vague as the formulation is, however, it suffices to eliminate from consideration such judgments as that so and so is a good person or that such and such is a good character trait.

Within the class of judgments directly concerning behavior, we can distinguish between those belonging to what Jonathan Bennett calls first-order morality and those belonging to what he calls second-order morality. First-order morality, Bennett explains, yields such judgments as 'It would be right for me to ϕ' and 'He acted wrongly in ϕing', while second-order morality yields such judgments as 'He is to blame for having ϕd'.[2] The distinction, however we may label it, is familiar, and is often drawn in terms of the judgments' objects: first-order judgments, it is said, attribute moral properties to actions, whereas second-order judgments attribute moral properties to agents. But Bennett argues that "the behavior/person or act/agent way of distinguishing the orders is superficial. What it calls a judgment on behavior is really one kind of judgment on a person: when we say that *what he did* was wrong we mean that *he* acted wrongly."[3] Bennett instead draws the distinction in terms of the judgments' functions: first-order judgments (at least in their prospective form) serve as guides to choice, whereas second-order judgments serve to express reactive attitudes, such as resentment or gratitude.[4] I do not wish to endorse this particular account of the judgments' functions, and I reject the equivalence asserted at the end of the quoted passage, but I agree that we should not treat it as obvious that first- and second-order moral judgments attribute moral properties to objects of different sorts. Here too, then, we must rest content, at least for now, with an unanalyzed, intuitive grasp of the distinction. Whatever its ultimate basis, the distinction seems clear enough to be usable. So to further specify my target: I shall focus upon first-order moral judgments directly concerning behavior.

There is, however, an ambiguity in the notion of *behavior*. Among judgments directly concerning behavior, some concern *things people (might) do,* while others concern what I shall call people's *concrete behavior*. Consider an example. Throwing a baseball is something a person might do. Indeed, it is something that many people have done and that a single person might do on multiple occasions. We may think of the "things people do," then, as act- or behavior- types. A particular person's throwing of a particular baseball on a particular occasion, by contrast, is not an act- or behavior-type. It is a token action, an unrepeatable, particular instantiation of the act-type *throwing a baseball.* It is a piece of what I am calling "concrete behavior". The judgment

[2] Bennett (1995: 46). Bennett credits the 'first-order/second-order' terminology to Alan Donagan.

[3] Ibid.

[4] Ibid.

'It is wrong to throw baseballs' (or equivalently, 'Throwing baseballs is wrong') concerns the act-type *throwing a baseball*; the judgment 'John acted wrongly in throwing that baseball' concerns a bit of John's concrete behavior. (Hence my rejection of the equivalence proposed by Bennett: 'What he did was wrong' concerns some unspecified act-type instantiated by the agent, whereas 'He acted wrongly in doing what he did' concerns the agent's particular instantiation of that act-type. The two judgments are no doubt related, but the exact nature of the relation is far from obvious.[5]) I think that the distinction between judgments concerning behavior-types and those concerning concrete behavior has been insufficiently attended to in moral theorizing. Both are important, but in this paper I shall focus upon the latter. So now to specify my target fully: in what follows I shall focus upon first-order moral judgments directly concerning agents' concrete behavior.

In saying that my target judgments *concern* agents' concrete behavior, I mean to leave it an open question what the *objects of evaluation* are in such judgments. The objects of evaluation are the things to which the judgments' evaluative predicates are applied. Now it might seem obvious—indeed most moral theories have taken it for granted—that the objects of evaluation in such judgments are actions. But I shall argue that often this is not so. (Later in the paper I shall explain why I think this result is significant.) Judgments directly concerning concrete behavior come in a variety of forms. I shall argue that while judgments of some of these forms evaluate actions, judgments of other forms, including certain forms of special interest to moral philosophy, do not. Or perhaps I should say that the judgments in question do not evaluate actions as actions have typically been understood by philosophers of action. Philosophers of action do not speak with one voice, of course, but there are points upon which almost all agree, and in what follows I shall try to rely only upon what is common to all (or almost all) accounts. Most centrally, actions are generally taken to be particular, unrepeatable occurrences that are intentional under some description or other.

2

I begin with judgments of the form 'A ϕd F-ly', where 'F' is a normative or evaluative term. For instance:

(1a) John invested his money prudently.

[5] I have discussed the relationship between 'ϕing is [im]permissible' and 'A acted [im]permissibly in ϕing' elsewhere (2005). Some of what I said there, however, must be revised in light of what I'll argue here.

In such judgments the adverb functions as an adverb of manner. (1a) tells us *how* John invested his money. And I think that the object of evaluation here is an action—which is exactly what the standard Davidsonian analysis would lead us to expect. According to Davidson, action sentences implicitly quantify over events (of which actions are a subclass), and adverbs are predicates of events.[6] If this is right, (1a) should be analyzed along the lines of

(1a′) There was an event which was an investing of his money by John and which was prudent.

There is a complication, however: in the context of (1a′), 'prudent' functions as an attributive adjective.[7] An adjective 'F' is attributive if one cannot infer 'x is an F H' from the conjunction of 'x is an F G' and 'x is an H', or infer 'x is F and x is a G' from 'x is an F G'. A paradigm example of an attributive adjective is 'large'. Suppose that Stuart is a large mouse. He is also a mammal. But it does not follow that he is a large mammal. On the contrary, he is a small mammal. Similarly, from the fact that he is a large mouse we cannot infer that he is both large and a mouse. The adjective 'large' cannot be detached in this way. A thing is not large or small *simpliciter*. Rather, it is (for example) a large or a small K, where the kind K determines the applicable standard of largeness. (We do sometimes say, without qualification, that a thing is large, but in such cases the relevant standard is recoverable from the context.) Likewise with 'prudent'. Suppose that John's investing of his money was also his disappointing of his friend (she had hoped he would use the money to take her to Paris). From the fact that John's act was a prudent investing of his money, we cannot infer that it was a prudent disappointing of his friend. On the contrary, it might have been a very imprudent disappointing of his friend. (The point is even more obvious at the level of (1a)'s surface grammar: John invested his money prudently, but it does not follow that he disappointed his friend prudently. We might thus call 'prudently' an attributive adverb.) Similarly, from the fact that John's act was a prudent investing of his money, we cannot infer that it was both an investing of his money and prudent. Actions are not prudent or imprudent *simpliciter*. Rather, they are prudent or imprudent ϕings. So let us rewrite (1a′) as

(1a″) There was an event which was an investing of his money by John and which was a prudent investing of his money.

[6] See Davidson (1967).

[7] That 'good' is attributive was pointed out long ago by Peter Geach (1956) and has been much emphasized in recent work by Judith Thomson (e.g. 1997, 2006). But the point holds for most, if not all, evaluative terms.

At the level of surface grammar, (1a) appears to attribute the property of having invested his money prudently to *John*, but upon analysis it turns out to attribute the property of having been a prudent investing of his money to John's *action*. Generalizing, let us say that when an adverb specifies an agent's manner of ϕing, it helps to specify a complex property of that action: the judgment that an agent ϕd F-ly says of his ϕing that it was an F ϕing. If this analysis is correct, actions (a subclass of events) are indeed the objects of evaluation in sentences of the form 'A ϕd F-ly', where 'F' is an evaluative term.

Now compare

(1a) John invested his money prudently

with

(1b) Prudently, John invested his money.

While (1b) *can* be used (perhaps poetically, and without the comma) to say exactly what (1a) says, I think it is more naturally interpreted as saying something quite different. As we have observed, (1a) concerns John's manner of investing his money. It says that John's investments were prudent ones—he didn't, for example, put all his money into internet stocks. On the reading that interests me, however, (1b) implies nothing about John's manner of investing his money. According to (1b), what was prudent was that John invested his money at all, when he could, for example, have spent it, or hidden it inside his mattress.[8] (1b) could be true even if John *did* put everything into internet stocks.

In (1b) 'prudently' seems to function as a sentential operator rather than as an adverb of manner. We must be careful, however, how we understand this operator. Strictly speaking, it wasn't the fact that John invested his money that was prudent. Rather, *John* was prudent. But (1b) does not attribute prudence *simpliciter* to John. Perhaps on the whole John was, and still is, a very imprudent person. (1b) attributes a qualified sort of prudence to John—it says that he was prudent *at least insofar as* he invested his money. I thus suggest we understand the adverb in (1b) as a sentential operator indexed to John.[9] That this analysis is on the right track is strongly suggested, I think, by (1b)'s equivalence to

[8] When the adverb occurs immediately after the subject ('John prudently invested his money'), the sentence can be equivalent either to (1a) or to (1b).

[9] Semi-formally, the analysis would be

$\text{Prudent}_{\text{John}}$(that he invested his money).

Alternatively, we might analyze the sentence relationally:

Prudent(John, that he invested his money).

(1c) John was prudent to invest his money.

Here prudence (relative to his having invested his money) is explicitly attributed to John. Admittedly the complement in (1c) is infinitival ('to invest his money') rather than sentential, but there are many contexts in which a bare infinitive stands in for a sentence. 'I hope to win the lottery', for example, should probably be analyzed as 'I hope that I win the lottery'. And indeed (1b) and (1c) both seem equivalent to the somewhat stilted

(1d) It was prudent of John that he invested his money,

which does employ a that-clause. I suggest that (1d) most transparently reflects the structure of the proposition that all three sentences express.

Similarly, compare

(2a) John fought his rival courageously,

with

(2b) Courageously, John fought his rival,

(2c) John was courageous to fight his rival,

and

(2d) It was courageous of John that he fought his rival.

(2a) says something about John's manner of fighting. It is true if (for example) John attacked aggressively when opportunities arose and stood his ground rather than retreating when he in turn was under attack. According to (2a), the actions that constituted John's fighting—or at least enough of them—were episodes of courageous fighting. (2b)–(2d), by contrast, say that John was courageous to fight his rival at all, given that he could (for example) have slipped quietly out of town the night before. It is consistent with these sentences' truth that John's manner of fighting wasn't the least bit courageous. Perhaps he didn't fight that courageously; even so, it was courageous of him *that* he fought.

Generalizing, then, let us say that whereas a-form judgments (e.g. (1a) and (2a)) evaluate *actions*, b–d-form judgments evaluate agents relative to *facts* about their behavior.[10]

[10] I speak of facts rather than propositions because evaluations of the b–d-forms are true only if the relevant propositions about the agents' behavior are *true*. 'John was courageous to fight his rival', for example, is true only if John did in fact fight his rival. But we can also make hypothetical judgments of corresponding forms; and the truth of e.g. 'John would have been courageous to fight his rival' does not presuppose that John actually fought.

So far we've looked only at judgments employing "thick" evaluative terms. Let's turn now to "thin" evaluations. Consider first

(3a) Yvonne sang *La Marseillaise* well.

This sentence says of Yvonne's act that it was a good singing of *La Marseillaise*—her performance was in tune, emotionally stirring, and so on.[11] Interestingly, with this example the b-form sentence seems unavailable. We cannot say

(3b) ?Well, Yvonne sang *La Marseillaise*,

at least not if this is supposed to mean something different from (3a). We can, however, use the c- and d-forms:

(3c) Yvonne was good to sing *La Marseillaise*

and

(3d) It was good of Yvonne that she sang *La Marseillaise*.

As in our other examples, the c- and d-forms say something quite different from the a-form. Suppose that despite her inability to carry a turn, Yvonne had joined Victor Laszlo in singing *La Marseillaise* in Rick's Café, thereby publicly showing her solidarity with those resisting Nazi rule. In that case (3a) would have been false but (3c) and (3d) true. She would not have sung *La Marseillaise* particularly well on that occasion, but it would have been good of her that she sang it. Conversely, it would have been bad of her to join the German officers in singing *Die Wacht am Rhein*, no matter how well she sang it.

Similarly, consider

(4a) Sam played *As Time Goes By* wrong.[12]

This sentence tells us something about Sam's manner of playing *As Time Goes By.* His performance was flawed—perhaps he hit some wrong notes,

[11] As I remarked earlier (see footnote 7), 'good' is an attributive adjective. Likewise, 'well' is an attributive adverb. Suppose that Yvonne's singing of *La Marseillaise* was also her signaling to her confederates. (Perhaps that's how she was supposed to warn them of the guard's approach.) From the fact that her action was a good singing, it does not follow that it was a good signaling. From the fact that Yvonne sang well, it does not follow that she signaled well. (Perhaps she sang so softly that her confederates did not realize she was giving the signal until it was too late.) Likewise, from the fact that Yvonne's action was a good singing, it doesn't follow that it was both good and a singing. The adjective 'good' cannot be detached in this way. (3a) does not say that a certain event, which happened to be a singing, was good *simpliciter*. It says that a certain event was a good instance of the kind *singing*. The act-type *singing* determines the applicable standard of evaluation.

[12] Here 'wrong' is more idiomatic than 'wrongly'. In this context 'wrong' is an adverb, not an adjective.

or played the song in the wrong key, or with the wrong tempo. By contrast,

(4c) Sam was wrong to play *As Time Goes By*

and

(4d) It was wrong of Sam that he played *As Time Goes By*

tell us that Sam erred in playing the song at all.[13] The criticism leveled in (4c) and (4d) concerns the fact that he played the song, not his manner of playing it. The performance itself might have been flawless.

Notice that in examples 3 and 4—the ones employing thin evaluative terms—the a-forms do not express *moral* evaluations. (3a) and (4a) concern the *musical* merits of Yvonne's and Sam's performances: Yvonne sang her song well; Sam played his song wrong. The c- and d-forms, by contrast, *do* express moral evaluations: it was good of Yvonne that she sang *La Marseillaise*; Sam was wrong to play *As Time Goes By*. In order to make moral a-form judgments employing thin evaluative adverbs we must (typically) add the modifier 'morally': instead of saying 'Yvonne sang *La Marseillaise* badly' we must say 'Yvonne sang *La Marseillaise morally* badly'. The latter sentence would be true if there were something morally objectionable about Yvonne's manner of singing—if she sang too loudly, for example, thereby waking the baby, or if she used an offensive, mocking tone of voice. But while such judgments are possible, they are not, I think, the thin evaluations concerning concrete behavior that have been of greatest interested to moral philosophers. Paradigmatically, a thin moral evaluation concerning an agent's concrete behavior concerns the fact that he did such and such, not his manner of doing it.

3

I'd now like to add another form of judgment to our inventory. Compare

(3a) Yvonne sang *La Marseillaise* well

with

(3e) Yvonne acted well in singing *La Marseillaise*.[14]

[13] Opinions differ over the felicity of 'Wrongly, Sam played *As Time Goes By*'.

[14] Some find sentences of the form 'A acted well in ϕing' artificial. I grant that in practice judgments of the form 'A acted F-ly in ϕing' tend to employ thick evaluative terms: there is nothing artificial about 'A acted courageously in ϕing' or 'A acted

These sentences clearly differ in meaning. To return to our earlier scenario: if Yvonne sang *La Marseillaise* with Victor Laszlo, she *acted* well, whether or not she sang well.[15] But how is (3e) to be analyzed? Syntactically, the phrase 'in singing *La Marseillaise*' seems to function as an adjunct. The first three words of the sentence are capable of expressing a complete thought on their own. So let us begin with the shorter sentence,

(S) Yvonne acted well.

(S) has the same surface form as (3a)—both are instances of the schema 'A ϕd F-ly'. And according to our earlier Davidsonian analysis, 'A ϕd F-ly' should be analyzed as 'There was event which was a ϕing performed by A and which was an F ϕing'. Applying this analysis to (S), we get

(S′) There was an event which was an acting performed by Yvonne and which was a good acting.

Now let's restore the adjunct phrase 'in singing *La Marseillaise*'. If (S) should be analyzed as (S′), (3e) should presumably be analyzed as

(3e′) There was an event which was an acting performed by Yvonne, which was a good acting, *and which was a singing of* La Marseillaise.

If this is correct, the chief difference between (3a) and (3e) is that while (3a) says that Yvonne's action was a good singing of *La Marseillaise*, (3e) says simply that it was a good *acting*. In each case the object of evaluation is Yvonne's action. The peculiar form of (3e), then, serves primarily to invoke a special, perhaps moral, standard of evaluation.

But I do not think this analysis of (3e) can be correct. For one thing, (3e′) misses the force of the 'in' linking 'singing La Marseillaise' to 'acted well'. (3e) doesn't just say that some event was both a singing of *La Marseillaise* by Yvonne and a good acting, as if these were independent features of the action. It says that Yvonne acted well *in* singing *La Marseillaise*—it implies a connection between the positive evaluation and the fact that her action was one of singing *La Marseillaise*.

Perhaps a minor alteration to the analysis could solve this problem. A deeper worry is this. The proposed analysis assumes that in e-form judgments the adverb functions just as it does in a-form judgments. But

prudently in ϕing'. Nor, interestingly, is there anything artificial about 'A acted *badly* in ϕing'. It seems to be only the notion of *acting well* that gives some people pause. But if we allow *acting badly*, I see no reason to reject *acting well*.

[15] Since 'A ϕd F-ly' and 'A acted F-ly in ϕing' are not generally equivalent, I was rash to use 'A ϕd [im]permissibly' and 'A acted [im]permissibly in ϕing' interchangeably in an earlier essay (2005).

this assumption is suspect. In a-form judgments the adverb functions as an attributive adverb of manner. 'A ϕd F-ly' says of A's act of ϕing that it was an *F ϕing*, with the act-type *ϕing* determining the applicable standard of F-ness. But *acting* is not just another (perhaps maximally general) act-type, to put along side such determinate act-types as *fighting, singing* and *signaling*; the idea that it too determines standards of F-ness should consequently give us pause. Are there really "manners" of acting, as there are of fighting, singing, and signaling? Consider loudness. The act-type *whispering* determines one standard of loudness, the act-type *singing* another: loud whispering is much quieter than loud singing. But how loud is loud acting? The bare notion of *acting* does not determine a standard of loudness. The situation may seem different when it comes to evaluative properties. We *can* say both that Yvonne sang well and that she *acted* well (in singing). But we should not be quick to assume that 'well' functions as an adverb of manner in the latter case. Good singing is one thing, good cake baking another, but what is good *acting*? Courageous fighting is one thing, courageous reporting of governmental misconduct another, but what is courageous *acting*? It is far from clear that bare notion of *acting* determines criteria for the application of evaluative terms.

A third reason for resisting the proposed analysis is that (3e) is obviously equivalent to (3c) and (3d). More generally, e-form judgments are equivalent to b–d-form judgments. To say that John acted prudently in investing his money is to say that he was prudent to invest it; both judgments are to be contrasted with the a-form judgment that he invested his money prudently. To say that Sam acted wrongly in playing *As Time Goes By* is to say that he was wrong to play it; both judgments are to be contrasted with the a-form judgment that he played the song wrong. Likewise with our other examples. But b–d-form judgments evaluate agents with respect to facts about their behavior; they do not attribute evaluative properties to actions, as a-form judgments do. We should thus reject any analysis of e-form judgments that makes their objects of evaluation out to be actions.

In the next section I shall offer an additional, and I hope decisive, argument against treating e-form judgments as evaluations of actions. First, however, I shall propose my own (tentative) analysis of such judgments, one that respects their affinity with b–d-form judgments. As before, let's start with judgments of the simpler form 'A acted F-ly'. I propose taking such judgments to quantify over facts about behavior, rather than over actions: 'A acted F-ly' should be analyzed, roughly, as 'There was a fact *f* about A's behavior such that it was F of A that *f*', or equivalently, as 'There was a fact about A's behavior such that A was an F agent with respect to that fact'.

(S) Yvonne acted well

should thus be analyzed as

(S$''$) There was a fact about Yvonne's behavior such that she was a good agent with respect to that fact.

We can then analyze 'A acted F-ly *in ϕing*' as 'There was a fact about A's behavior such that A was an F agent with respect to that fact, *and that fact was that she ϕd*'. This means that (3e) should be analyzed as

(3e$''$) There was a fact about Yvonne's behavior such that she was a good agent with respect to that fact, and that fact was that she sang *La Marseillaise*.

According to this analysis, (3a) and (3e) do not attribute two different sorts of goodness to Yvonne's singing of *La Marseillaise*—they do not say, respectively, that her action was a good singing and that it was a good acting. The function of the 'acted well' locution in (3e) is not to signal that Yvonne's action is being evaluated *qua* action, as opposed to *qua* instantiation of this or that more specific act-type. Rather, the locution signals that Yvonne is being evaluated *qua* actor (i.e. *qua* agent)—it signals that she is being evaluated *qua* exerciser of the power to act. (3e) says (roughly) that Yvonne was a good agent insofar as she sang *La Marseillaise*.

This analysis avoids the problems confronting the earlier analysis. It captures the connection between the positive evaluation and the fact that Yvonne's action was one of singing *La Marseillaise*. It does not treat 'well' as an adverb of manner. And it accounts for (3e)'s equivalence to (3c) and (3d): 'Yvonne was good to sing *La Marseillaise*' and 'It was good of Yvonne that she sang *La Marseillaise*' both say, in effect, that Yvonne was a good agent with respect to the fact that she sang *La Marseillaise*.

4

I have proposed an analysis of judgments of the form 'A acted F-ly in ϕing', where 'F' is an evaluative term. I am not wed to the details of this account. What I do want to insist upon, and what is important for my purposes in this paper, is that these judgments do not attribute evaluative properties to actions. The correct analysis must focus upon facts about agents' behavior, not (directly) upon their actions. In this section I offer an independent argument for this claim. (The argument establishes the same thing about the equivalent b–d-form judgments, but in presenting the argument I shall focus upon the e-form.)

The argument is simple: it can't be that 'A acted F-ly in ϕing' makes a claim about an agent's act of ϕing, because not every expression that can yield a truth when substituted for 'ϕing' in this schema characterizes an action. An agent can act F-ly in ϕing (it can be F of him that he ϕs, etc.) even though he performs no *act* of ϕing. I shall discuss examples of two kinds: first, examples in which agents are praised (or criticized) for *not* doing certain things; and second, examples in which they are praised (or criticized) for exhibiting certain *patterns* in their behavior over time.[16] I start with an example of the first kind.

Consider the judgment that John acted well in not speaking rudely to his in-laws during their week-long visit. Here John is praised, not for performing a certain action, but for not performing an action of a certain type. John's non-performance of an action of that type, however, is not itself an action he performs—there is no action here with respect to which John is being said to have acted well. Rather, he is being said to have acted well with respect to the fact that he performed no action of the relevant type during the relevant period. Or so it seems to me.

Suppose someone wanted to insist that, despite appearances, John *is* being praised for performing an action in this example. Here's one way this person might defend his position: he might *identify* John's not speaking rudely to his in-laws with whatever action John performed instead of speaking rudely to them. This move might seem plausible in certain cases. Suppose, for example, that there was a particular moment at which John was sorely tempted to speak rudely to his in-laws, but at which he spoke politely instead. It might seem plausible to identify John's not speaking rudely on that occasion with his act of speaking politely. But in the case I actually described, John was praised for not speaking rudely to his in-laws during their entire week-long visit. There is thus no particular action that can plausibly be identified as the one he performed instead of speaking rudely to them. Perhaps we could identify John's not speaking rudely to his in-laws during their visit with the totality of actions he performed during their visit. Unfortunately, the totality of actions John performed during that week is not itself an action he performed that week. In any case, the judgment that John acted well in not speaking rudely to his in-laws cannot be equivalent to the judgment that he acted well in doing everything that he did do during that week, because his having acted well in not speaking rudely to his in-laws is consistent with his having acted *badly* in doing everything that he did do during that week. His actions that week might

[16] Anselm Müller (2004: 16–17), who also argues that someone's acting well needn't consist in his performance of a particular action, uses similar examples to make his point. But see n. 21 below.

have been thoroughly reprehensible, every one of them. But at least he didn't speak rudely to his in-laws.[17]

The suggestions considered so far have failed because they shift the topic: they interpret the judgment that John acted well in not speaking rudely to his in-laws as concerning something other than his not speaking rudely to his in-laws. But there is another way in which this judgment could be interpreted as being about an action of John's. Perhaps John's not speaking rudely to his in-laws during their visit is an action in its own right, so to speak, distinct from any of the "positive" actions he performed that week. Whether purely negative actions exist is a much disputed question that I cannot hope to settle here. Luckily, I don't have to, since the existence of purely negative actions would not save the thesis that true judgments of the form 'A acted F-ly in not ψing' always concern actions. It cannot plausibly be maintained that *whenever* an agent performs no action of a given type during some period of time, he thereby performs a purely negative action. For it is widely held that every action must be intentional, or at least voluntary, under some description or other.[18] Let us assume that this is right. It follows that agents perform purely negative actions only when they either *intentionally refrain* from acting in certain ways or (perhaps) when they *omit* performing actions of certain types.[19] (An agent omits doing something if he does not do it, but could and should have done it.) Now consider again the judgment that John acted well in not speaking rudely to his in-laws. Speaking rudely to his in-laws was not something that John should have done, so it was not something that he omitted doing. And while we can certainly imagine a version of the example in which John consciously chose not to speak rudely to his in-laws, this needn't be how it happened. Perhaps, despite their many provocations, he never seriously considered speaking rudely to them, and never had to resist an urge to do so. Speaking rudely to his in-laws was consequently not something that he intentionally refrained from doing. Yet it is still true that he acted well in not speaking rudely to them.[20]

[17] Consider also the judgment that John acted badly in not picking up his friend from the airport. And suppose that John was asleep when he was supposed to be picking up his friend. In that case we cannot identify John's not picking up his friend with an action, because John did not perform any actions at all during the relevant time interval. It could still be true, however, that he acted badly in not picking up his friend.

[18] See Davidson (1971).

[19] There is a tradition of regarding omissions, even when not intentional, as voluntary. See e.g. St Thomas Aquinas (1265–73: 1a–2ae q.6 a.3).

[20] There is another possible view: perhaps 'A acted F-ly in not ψing' asserts something, not about a purely negative action, but about a purely negative *event*. The idea would be that a purely negative event *does* occur whenever something of a certain type

Let us turn now to the second sort of example. Consider the (true) judgment that John acted well in calling his depressed friend once a day for a week. During the week John performed seven acts of the type *calling his depressed friend*, but did he also perform a single action of the type *calling his depressed friend once a day for a week*? I think not. Or rather, he needn't have done. Many philosophers believe that the "fusion" of any two events is itself an event. But actions cannot be so promiscuously fused. Many (perhaps all) actions are composed of other actions, but smaller actions must be related to one another in a special way if they are jointly to constitute a larger action. Sam's baking of a loaf of bread is an action, and it includes as parts such sub-actions as his mixing of the ingredients, his kneading of the dough, his turning on of the oven, and so on. But these smaller actions compose a single, larger action only because they are united by Sam's intentions: Sam performs the smaller actions in order thereby to further his goal of producing a loaf of bread. Intentions, or (perhaps) more broadly, plans, are what bind smaller actions together into larger ones. Now let us return to John and his depressed friend. We can certainly imagine a version of the example in which John, concerned about his friend, formed the intention to call him daily until the worst had passed. Given this scenario, John's seven acts of calling his friend could perhaps plausibly be thought of as parts of a single larger action of calling his friend every day for a week. But that needn't be how it happened. Perhaps the thought "I should give my friend a call to find out how he's doing" simply struck John anew each day. Perhaps he never formulated any longer-term policy or plan on the subject. Even on this scenario, however, I think it could be true that John acted well in calling his friend once a day for a week. Here John is praised not for performing a certain action, but for exhibiting a certain pattern in his behavior over time. He is praised with respect to the fact that he instantiated this pattern.[21]

doesn't happen. (Unlike actions, events needn't be intentional or voluntary under any description.) Although I cannot argue the point here, I believe that this understanding of negative events is plausible only on a view that effectively erases the distinction between events and facts (or states of affairs).

[21] In arguing that an agent's acting well needn't consist in his performance of some particular action, Müller (2004: 17) offers the example of an agent who acts well in twice warning his stubborn friend against investing all his money in shares. The case is under-described, but I think it is natural to regard this agent's two warnings as together constituting a single, larger act of warning his friend off a bad investment. Compare: a man kills his enemy by shooting him several times, repeating as necessary. Perhaps he expected the first shot to do the trick, but he was prepared to revise his plan as the situation developed. Here the multiple gun firings are parts of a larger action that wasn't over with until the agent had achieved his goal of killing his enemy. Müller's agent had the goal of dissuading his friend from making a bad investment. He might have

It is admittedly possible to interpret 'A acted well in ψing n times' distributively. The judgment that John acted well in calling his friend once a day for a week, for example, could be interpreted as meaning that every day for a week, John acted well in calling his friend—that is, that he acted well on each of seven occasions. But this analysis does not capture the meaning of our original judgment. It may be true that John acted well each time he called, but in saying that he acted well in calling once a day for a week we are saying more than this. We are saying that he acted well in calling so often and so regularly. This can perhaps be brought out more vividly if we alter the example a little. Suppose that given his friend's condition, calling just once a day for a week wasn't nearly enough. It would still be true that John acted well each time he called, but it would not be true that he acted well in calling his friend once a day for a week. He would not have acted well in instantiating that particular pattern in his behavior.[22]

The example I've been discussing concerns John's pattern of response, over time, to a single, ongoing situation, namely his friend's depression. We can also evaluate agents with respect to their patterns of response, over time, to recurring *types* of situation. One agent may be evaluated positively for having devoted, over the years, so much of his time and money to helping people in need; another may be evaluated negatively for having done so little for others over the same period. Sometimes, of course, an agent has a settled long-term policy regarding his charitable activity. In such cases the agent's individual acts of charity might plausibly be thought of as parts of, or episodes in, a single, on-going charitable act. But an agent needn't have adopted such a policy in order to have acted well (or badly) in giving so much (or so little) to others over the years. Nor, as we saw above, can such pattern-judgments be reduced to sets of judgments about how well (or badly) the agent acted on various individual occasions. The duty to give to charity is imperfect; the agent's failure to fulfill it may show up only in his pattern of charitable activity over time.

I have pointed out that in the e-form schema 'A acted F-ly in ϕing', 'ϕing' can be replaced with expressions of the form 'not ψing' or 'ψing n times'. Note, however, that in the a-form schema 'A ϕd F-ly', 'ϕd' *cannot* be replaced with expressions of the form 'didn't ψ' or ' ψd n times'. One cannot say, for example, that John didn't speak rudely to his in-laws well.

expected one warning to suffice, but when it didn't he repeated the warning until his friend changed his mind.

[22] Note that an agent who doesn't ψ at all during a certain period also exemplifies a pattern in his ψings over time. (We can think of this as a limiting case of a pattern.) Examples in which agents are praised or condemned for not acting in certain ways can thus be subsumed under the present category.

Or rather, if one did say this, one could only mean that it's not the case that John spoke rudely to them well (whatever speaking rudely to them well might amount to). Similarly, in the case where John had no plan or intention to call his friend seven times, one cannot say that he called his friend seven times well. Or rather, if one did say this, the judgment could only be understood distributively, as saying that seven times, John called his friend well (whatever calling someone well might amount to). The range of expressions that can be substituted for 'ϕd' in the a-form schema 'A ϕd F-ly' is restricted because a-form judgments quantify over, and attribute evaluative properties to, actions (a subclass of events). The substituted expressions must consequently designate act-types—they must be convertible into predicates of actions. There is no such restriction with e-form judgments because these judgments do not attribute evaluative properties to actions. They concern facts about agents' behavior, and the relevant facts needn't be to the effect that a given agent performed an action of a certain type.

This asymmetry also supports my contention that 'A acted F-ly' cannot be analyzed as other instances of the general schema 'A ϕd F-ly' are analyzed. We can always add an 'in' clause to 'A acted F-ly', thereby expanding it into the e-form 'A acted F-ly in ϕing'. Conversely, we can always truncate the e-form by dropping the 'in' clause. But I have argued that when it comes to the e-form, there need be no action available to serve as an object of evaluation. The same must consequently be true of judgments of the shorter form.

5

Of the forms canvassed so far, only the a-form attributes evaluative properties to actions. Judgments of the other forms evaluate agents with respect to facts about their behavior. And typically, a-form judgments employing thin evaluative terms (e.g. 'Yvonne sang *La Marseillaise* badly') express non-moral evaluations. (When they do express moral evaluations, they express what we might call "moral manner" evaluations.) So far, then, the thin evaluations of central importance to moral philosophy are the ones that don't have actions as their objects of evaluation.

I turn now to evaluations of the form 'A's ϕing was F'. Here the subject expression appears to denote an action. Judgments of this form thus appear to attribute evaluative properties to actions. And 'Sam's playing *As Time Goes By* was wrong' surely expresses a *moral* evaluation—and not a mere "moral manner" evaluation, either. Likewise for 'Yvonne's singing *La Marseillaise* was good'. It thus appears that there *are* thin evaluations of central importance to moral philosophy that attribute moral properties to actions. Appearances, however, are deceiving.

The schematic gerundial nominal 'A's ϕing' is ambiguous, and on only one reading does it denote an action. Consider the sentence 'Yvonne sang *La Marseillaise*'. From it we can derive either the *perfect* gerundial nominal 'Yvonne's singing of *La Marseillaise*' or the *imperfect* gerundial nominal 'Yvonne's singing *La Marseillaise*'.[23] Both can be used in the subject position of a sentence—'Yvonne's singing of *La Marseillaise* surprised me' and 'Yvonne's singing *La Marseillaise* surprised me' are equally acceptable—but they differ grammatically. The perfect nominal, but not the imperfect, can be pluralized: 'Yvonne's singings of *La Marseillaise* surprised me' is acceptable, but 'Yvonne's singings *La Marseillaise* surprised me' is not. Similarly, the possessive noun 'Yvonne's' can be replaced with a definite or indefinite article in the perfect nominal, but not in the imperfect one: 'The singing of *La Marseillaise* surprised me' is fine, but 'The singing *La Marseillaise* surprised me' is not. Perfect nominals take attributive adjectives ('Yvonne's loud singing of *La Marseillaise* surprised me'), whereas imperfect nominals take adverbs ('Yvonne's singing *La Marseillaise* loudly surprised me'). And the gerunds in imperfect nominals can be negated, tensed, or modified with auxiliaries—'Yvonne's not singing *La Marseillaise* surprised me' and 'Yvonne's having sung *La Marseillaise* surprised me' are both acceptable—but we cannot perform these operations on the gerunds in perfect nominals. In short, while the gerunds in perfect nominals behave entirely like nouns (that's what makes them "perfect" nominals), the gerunds in imperfect nominals retain many of the characteristic features of verbs.

When the sentence from which a gerundial nominal is derived has a direct object (e.g. 'Yvonne sang *La Marseillaise*'), the presence or absence of the word 'of' reveals whether the nominal is perfect or imperfect ('Yvonne's singing *of La Marseillaise*' is perfect, 'Yvonne's singing *La Marseillaise*' is imperfect). From the sentence 'Yvonne sang', however, we can derive only 'Yvonne's singing'. Is this nominal perfect or imperfect? It could be either. Which it is on a given occasion of use depends upon how the speaker would be willing to modify it. If he'd use an attributive adjective ('Yvonne's loud singing surprised me'), then it's perfect. If he'd use an adverb ('Yvonne's singing loudly surprised me'), it's imperfect. If he'd negate the gerund, or modify it with an auxiliary ('Yvonne's not singing surprised me', 'Yvonne's having sung surprised me'), then the nominal is imperfect. And so on.

[23] My treatment of this distinction follows that of Bennett (1988: 4–6), who in turn acknowledges the influence of Zeno Vendler (1967).

We can now see why the schematic gerundial nominal 'A's ϕing' is ambiguous: instances of it can be either perfect or imperfect. And the difference matters. For as Jonathan Bennett persuasively argues, while perfect gerundial nominals derived from action sentences denote actions, imperfect gerundial nominals denote facts.[24] 'Yvonne's loud singing of *La Marseillaise* surprised me' attributes the property of having surprised me to an action of Yvonne's. But 'Yvonne's singing *La Marseillaise* loudly surprised me' is equivalent to 'That Yvonne sang *La Marseillaise* loudly surprised me', and the nominal employed in the latter sentence undeniably denotes the fact that Yvonne sang *La Marseillaise* loudly. (Similarly, 'Yvonne's not singing *La Marseillaise* surprised me' is equivalent to 'That Yvonne did not sing *La Marseillaise* surprised me'. In this case it is clearly not an action that surprised me.)

The claim that imperfect gerundial nominals denote facts, while perfect gerundial nominals derived from action sentences denote actions, coheres well with what I argued earlier about the difference between b–e-form judgments on the one hand and a-form judgments on the other. For I think it is clear that evaluations employing imperfect nominals are equivalent to b–e-form evaluations, while those employing perfect nominals are (roughly) equivalent to a-form evaluations.

(2f) John's fighting his rival was courageous,

for example, is equivalent to

(2d′) That John fought his rival was courageous (of him),

which is just a variant of

(2d) It was courageous of John that he fought his rival.
(2g) John's fighting of his rival was courageous,

by contrast, tells us something about John's *manner* of fighting. It is (roughly) equivalent to

(2a) John fought his rival courageously.

(I say the two are *roughly* equivalent because although both attribute a property to an action performed by John, the former employs a nominal expression denoting John's action, while the latter, according to the Davidsonian analysis, employs an existential quantifier.) Likewise,

(1f) John's investing his money was prudent

[24] See Bennett (1988: 6–9), who again follows Vendler (1967). Only perfect gerundial nominals derived from action sentences denote actions. Other perfect gerundial nominals denote events ('the lightning's illuminating of the house') or states ('the house's possessing of a fireplace').

is equivalent to

(1d′) That John invested his money was prudent (of him),

which is just a variant of

(1d) It was prudent of John that he invested his money.
(1g) John's investing of his money was prudent,

however, tells us about John's manner of investing. It is roughly equivalent

(1a) John invested his money prudently.

Now let's consider an example involving a thin evaluative adjective.

(3f) Yvonne's singing *La Marseillaise* was good

expresses a moral evaluation. Indeed, there are two different moral evaluations it might express. On one reading, the sentence is equivalent to 'That Yvonne sang *La Marseillaise* was good'; on the other, it is equivalent to 'That Yvonne sang *La Marseillaise* was good *of her*'.[25] On the former reading, the sentence says simply that it was a good thing, morally speaking, that Yvonne sang *La Marseillaise*. On the latter reading, the sentence is equivalent to (3c)–(3e). (Recall that with this example the b-form is unavailable.) But either way, the evaluation concerns the *fact* that Yvonne sang *La Marseillaise*.

(3g) Yvonne's singing of *La Marseillaise* was good,

by contrast, is most naturally interpreted as expressing a non-moral evaluation of Yvonne's action: it says of her action that it was a good singing of *La Marseillaise*. In other words, (3g) is (roughly) equivalent to

(3a) Yvonne sang *La Marseillaise* well.

Some people have difficulty hearing the difference between (3f) and (3g). The difference is perhaps easier to detect when the perfect nominal is non-gerundial in form. Some verbs allow for the derivation of both gerundial and non-gerundial perfect nominals. From the sentence 'Yvonne performed *La Marseillaise*', for example, we can derive both the perfect gerundial nominal 'Yvonne's *performing* of *La Marseillaise*' and the non-gerundial nominal 'Yvonne's *performance* of *La Marseillaise*'. ('Sang', by contrast, permits only the derivation of gerundial nominals.) Now compare 'Yvonne's performing *La Marseillaise* was good' with 'Yvonne's performance of *La Marseillaise*

[25] 'John's fighting his rival was courageous' is not ambiguous in this way. It makes no sense to attribute courage to the fact that John fought his rival. The sentence can only mean that it was courageous *of John* that he fought his rival.

was good.' I think it is obvious that only the former sentence expresses a moral evaluation. The latter says of Yvonne's action that it was a good *performance* of *La Marseillaise*. (In order to express a thin moral evaluation with a perfect nominal in subject position, we must make it explicit that we mean to attribute a moral property to the action in question: we must say, for example, 'Yvonne's performance of *La Marseillaise* was morally good'. I take this to be equivalent to 'Yvonne performed *La Marseillaise* morally well'.) Likewise with deontic terms. 'Sam's performing *As Time Goes By* was wrong' expresses a moral judgment. 'Sam's performance of *As Time Goes By* was wrong', by contrast, seems roughly equivalent to 'Sam performed *As Time Goes By* wrong'.

To sum up, 'A's ϕing was F' can concern either A's act of ϕing or the fact that A ϕd, depending upon whether the gerundial nominal 'A's ϕing' is perfect or imperfect. And when 'F' is a thin evaluative term, 'A's ϕing was F' (typically) expresses a *moral* evaluation only when 'A's ϕing' is imperfect—in which case the sentence concerns a fact, not an action. The difference between the g- and f-forms thus perfectly parallels that between the a- and b–e-forms.

6

I have argued that although the a- and g-forms attribute evaluative properties to actions, the b–f-forms evaluate agents with respect to facts about their behavior. Sometimes the relevant fact is that the agent performed an action of a certain kind; sometimes it is that he didn't perform an action of a certain kind, or that he instantiated a certain pattern in his behavior over time. But why should this conclusion be thought significant for moral theory? It's hardly news that we evaluate people for things other than their actions. We evaluate them for their character traits, for their talents, even for their looks. What have I done but add a few more items to the list? We also evaluate people for what they don't do and for their patterns of behavior over time. So what?

There is a difference. When we pass from the judgment that John acted well in driving his neighbor to the hospital (or that it was good of him to do this, etc.) to the judgment that John is a nice person, or a good driver, or handsome, there's a noticeable shift of evaluative gears. We move from one type of evaluation to another, and as we do so we bring different criteria of evaluation to bear. But there is no shift of evaluative gears as we pass from the judgment that John acted well in driving his neighbor to the hospital to the judgment that he acted well in not speaking rudely to his in-laws, or in calling his depressed friend once a day for a week, or in giving so much to charity over the years.[26] Our evaluative *focus* shifts as we move from one judgment to another, encompassing now this aspect of John's behavior, now that aspect, now more of his behavior, now less of it, but the *mode* of evaluation employed remains the same throughout. It would seem, then, that for each evaluative term 'F', judgments of forms b–f should be understood uniformly. This places a constraint on how we understand the criteria of evaluation brought to bear by such judgments: the criteria must be applicable with respect to the whole range of behavioral facts that the judgments can concern. The criteria provided by the standard normative theories, however, do not satisfy this constraint. They are tailor-made for actions and cannot readily be extended to cover the full range of admissible behavioral facts.[27]

Let us begin with Kant's categorical imperative. In the spirit of its first formulation, we might say that an agent acts well in ϕing if and only if his maxim in ϕing is one that he could at the same time will to be a universal law.[28] The problem with this proposal is that it applies only to cases in which agents have maxims for acting as they do. Roughly speaking, a maxim is a principle that encapsulates an agent's reason for doing what he does.[29] Now when an agent intentionally refrains from doing something, he refrains for a reason, and so he can perhaps be seen as acting on a maxim ("For reason R, I will refrain from ψing"). Likewise, when an agent has a plan or policy encompassing a (possibly open-ended) series of actions, he can perhaps be seen as acting on a single maxim in performing that series of actions ("For reason R, I will ψ once a day for the next week"). But as we have seen, cases in which agents act well or badly needn't be like this. An agent can act well (or badly) in not ψing even though the possibility of ψing never occurs to him. He can act well (or badly) in instantiating a pattern in his behavior over time even though his practical thoughts never stray beyond the actions available to him at the present moment. To return

[26] For simplicity I shall focus primarily upon e-form judgments for the remainder of this paper.

[27] I shall not here discuss whether the standard normative theories provide adequate criteria for judgments that attribute moral properties to act-types (e.g. 'Killing is *pro tanto* wrong'). I am interested only in how the theories fare with respect to judgments concerning agents' concrete behavior.

[28] In this section I focus upon possible accounts of acting *well*. The difficulties raised should also apply to analogous accounts of acting *rightly*. Let me stress that in raising difficulties for this version of Kant's formula of universal law, I do not mean to imply that no theory Kantian in spirit can provide an adequate account of judgments of the form 'A acted well [or badly] in ϕing'.

[29] According to Kant (1785: Ak. 421/1959: 38), "a maxim is the subjective principle of acting ... [It] contains the practical rule which reason determines according to the conditions of the subject (often its ignorance or inclinations) and is thus the principle according to which the subject acts."

to our familiar examples, John needn't have made it his maxim to refrain from speaking rudely to his in-laws, or to call his depressed friend once a day for a week. The Kantian criterion of evaluation consequently cannot explain how it is that John acted *well* in not speaking rudely to his in-laws, or in calling his friend once a day for a week.[30] It is striking that when Kant applies his formula of universal law to a case involving the imperfect duty to help others, he imagines an agent who has explicitly made it his maxim not to help others (presumably so that he can devote all of his resources to advancing his own happiness).[31] But what of an agent who fails to fulfill this imperfect duty without ever having adopted a maxim on the subject? It's not that he's made it his policy to not help others; it's just that when opportunities to help arise, he usually thinks he has better things to do. Yet surely he acts badly in helping others so seldom.

It might be suggested that these agents are guided by *implicit* maxims or policies. I grant that a maxim can be implicit: an agent can ϕ for reason R without explicitly thinking to himself "In circumstances such as these, I shall ϕ for reason R". Perhaps an agent can even have an implicit general policy. But implicit maxims and policies must still capture agents' actual reasons for acting as they do; and I see no reason to think that an agent who has acted well (or badly) in giving so much (or so little) to charities over the years must have acted, at least implicitly, on the basis of some general policy regarding charitable activity. Nor must an agent have had at least an implicit maxim of not ϕing in order to have acted well (or badly) in not ϕing.

Perhaps we can explain the relevant facts about agents' behavior by appealing to their character traits, rather than to the idea that they were guided by implicit maxims. Perhaps a certain character trait explains John's pattern of charitable activity over the years; perhaps another trait explains why he didn't speak rudely to his in-laws during their visit (the trait might explain this by explaining why it never even occurred to him to speak rudely). I doubt that an adequate criterion for judging whether agents have acted well or badly can be fashioned from an appeal to character traits, but I shall not pursue that question here. Such a criterion would in any case be a large step away from Kant's formula of universal law. Character traits may help explain agents' patterns of sensitivity to (putative) reasons, but they are not themselves "subjective principles of acting".

[30] The same problem undermines a proposal of Anselm Müller's. Müller (2004: 19) writes that "simplifying a little, we may say: *In Φ-ing (i.e. in doing or not doing anything) you act well as long as you Φ for good reasons*" (italics in the original). He goes on to explain why he regards this as a simplification, but the complications he discusses do not address the central difficulty: an agent can act well in ϕing even though he does not ϕ for a reason.

[31] Kant (1785: Ak. 423/1959: 41).

Let us turn next to consequentialism. If judgments of the form 'A acted well [or badly] in ϕing' always concerned particular acts of ϕing, there would be no special obstacle to understanding the applicable criterion of evaluation in consequentialist terms. We could simply say that an agent acts well in ϕing if and only if there was no alternative action available to him such that the outcome would have been better had he performed that action instead. But I have argued that the judgment that an agent acted well in ϕing evaluates the agent with respect to the *fact* that he ϕd. Can we adapt the consequentialist criterion to cover evaluations of this sort? Can we say that an agent acts well in ϕing if and only if there is no alternative state of affairs regarding his behavior such that the outcome would have been better had that state of affairs obtained instead?

The first problem with this proposal is that it is not at all clear what it is for two states of affairs regarding an agent's behavior to be alternatives to one another.[32] The relevant alternatives cannot include every possible state of affairs regarding his behavior that the agent had it in his power to make obtain. If an agent's ϕing and his ψing are completely independent of one another, it should not count against his acting well in ϕing that the outcome of his ψing would have been even better. Perhaps he could have done both. The range of relevant alternatives must thus be narrowed down. Should we say that two possible states of affairs regarding an agent's behavior count as alternatives to one another only if they are incompatible? No, this is too restrictive. Consider the judgment that John acted well in ϕing more than five times. And suppose that what he really needed to do, given the circumstances, was ϕ at least *ten* times. In that case he didn't act well in ϕing more than five times. In this evaluative context, ϕing at least ten times is a relevant alternative to ϕing more than five times. But the two states of affairs are compatible. Perhaps John ϕd exactly twelve times.[33] Likewise, it will not do to say that two possible states of affairs regarding an agent's behavior are alternatives only if they concern exactly the same time interval. In order to determine whether an agent acted well in ϕing exactly once during the week, we must consider not just alternative scenarios in which he did something else when he actually ϕd, but also scenarios in which he e.g. ϕd twice instead of just once. This last scenario, however, concerns, in part, the agent's behavior at times other than that of his actual ϕing. Should we say, then, that two possible states of affairs regarding an agent's behavior are alternatives only if they concern partially overlapping

[32] I do not think that the idea of an alternative *action* is as clear as it is generally thought to be either.

[33] Another problem, which I shall not pursue: how does one determine the "outcome" of e.g. the fact that John ϕd more than five times?

time intervals? This might indeed be a necessary condition, but it's still not a sufficient one—it doesn't narrow down the range of relevant alternatives far enough. It should not automatically count against an agent's having acted well in ϕing at a certain time that the overall outcome would have been even better had he both ϕd at that time and ψd at another time. John acted well in driving his neighbor to the hospital. This is true despite the fact that the outcome would have been even better if he had both driven his neighbor to the hospital *and* donated $1,000 to Oxfam the next day.[34]

I think this is a serious problem. But even if it can be overcome, there is a second problem. If the proposed criterion for acting well is applicable at all, it is applicable with respect to *every* fact about an agent's behavior. That is, for any fact regarding an agent's behavior, the criterion will yield an answer to the question whether the agent acted well with respect to that fact. It should thus make sense to ask whether an agent acted well in e.g. ϕing at t_1, ψing at t_2, and χing at t_3, for arbitrarily chosen act-types and times. It should make sense, for example, to ask whether an agent acted well in going to the store on Monday, brushing his teeth on Tuesday, and reading a book on Wednesday. And there must be answers to such questions: either the outcome would have been better had some alternative state of affairs regarding the agent's behavior obtained, or the outcome would not have been better. But can we really make sense of the claim that an agent acted well in going to the store on Monday, brushing his teeth on Tuesday, and reading a book on Wednesday? In the absence of some special story, I can make sense of this only on a distributive reading: the agent acted well on each of these three occasions. I cannot understand the claim that in addition to acting well in doing each of these things, the agent acted well in instantiating this larger pattern in his behavior. This simply isn't the sort of pattern with respect to which an agent can sensibly be said to have acted well or badly. So we have here a second constraint that must be satisfied by any adequate account of the evaluative criterion invoked by judgments of the form 'A acted well [or badly] in ϕing'. In addition to explaining how agents can act well or badly with respect to a wide range of facts about their behavior (that was the first constraint), the account must also explain why

[34] Instead of saying that an agent acts well in ϕing if and only if there is no possible alternative state of affairs regarding his behavior such that the outcome would have been better had that state of affairs obtained instead, we might say that an agent acts well in ϕing if and only if the outcome would have been no better had he *not ϕd*. According to this suggestion, in order to determine whether an agent acts well in ϕing we need consider only *one* alternative scenario: that in which he doesn't ϕ. But this would often yield the wrong result. An agent might act badly in ϕing because, given the circumstances, he should have ψd instead. But there is no guarantee that had he not ϕd, he would have ψd instead. It might thus be the case that although the outcome would have been better had the agent ψd instead of ϕing, it would not have been better had he simply not ϕd.

agents cannot sensibly be said to act well or badly with respect to just *any* facts about their behavior.[35]

Of course the consequentialist can side-step these difficulties by insisting that he is not interested in the moral predicates that we actually use. He is introducing a new (thin) moral predicate, one that is by stipulation a predicate of actions. Nothing I've said here shows this approach to be illegitimate. I would point out only that when it is formulated as an account of what makes concrete actions good or bad (or right or wrong), consequentialism is a far more revisionary doctrine than is generally appreciated.

7

I do not propose to offer here a substantive account of the evaluative criterion invoked by judgments of the form 'A acted well in ϕing'. I will close, however, by briefly describing the general form of an account that I think might satisfy the constraints I have identified. Let's begin with the second constraint. Reflection on examples suggests that the patterns of behavior with respect to which agents can sensibly be said to have acted well or badly are those that constitute patterns of response to particular (types of) reasons. John's seven calls to his depressed friend, for example, were all responses to his friend's depression. As long as that condition persisted, John had a persisting conditional reason to call his friend if they had not recently spoken. John's calling his friend once a day for a week thus constitutes a pattern of response, over time, to this persisting reason. Similarly, every time John helped someone in need over the years, he was responding to a particular type of reason. His pattern of helping people over the years is a pattern of response to this type of reason. But now consider the pattern of behavior consisting in John's going to the store on Monday, brushing his teeth on Tuesday, and reading a book on Wednesday. These actions were not responses to a single, persisting reason, nor were they responses to a single type of reason. His having performed them consequently does not constitute a pattern of response to some (type of) reason. I think this explains why John cannot be evaluated with respect to his having done these three things. He can be evaluated only with respect to his having done each of them, taken singly.

[35] It has been suggested to me that an agent does act either well or badly with respect to every fact about his behavior, and that in most cases it is simply pragmatically inappropriate to say so. In order to be properly assessed this proposal would have to be supplemented with an account of *why* it is usually pragmatically inappropriate to express such judgments.

If the patterns of behavior with respect to which agents can be said to have acted well or badly are those that constitute patterns of response to (kinds of) reasons, then it is plausible to conclude that judgments about how well or badly agents have acted concern how well or badly they have responded to reasons. This same result can be reached by reflecting upon the notion of *acting* that such judgments employ. The power to act is the power to do things for reasons. To judge that someone acted well or badly is to evaluate him *qua* exerciser of this power. It is plausible to conclude that an agent acts well in ϕing if and only if he responds well to reasons in ϕing.[36] (This is a formal, and not a substantive, account of the operative evaluative criterion because it tells us nothing about what it is to respond well to reasons.)

The account satisfies the second criterion of adequacy, but does it satisfy the first? Does it cover the full range of cases in which agents can sensibly be said to have acted well or badly in ϕing? We have just seen how it can cover cases in which an agent's ϕing consists in his behavior's instantiating some appropriate pattern over time. The account can obviously also cover cases in which an agent's ϕing consists in his performance of a single action. When an agent performs an action, he does something for a reason. His action may thus be thought of as his response to that reason. Indeed, it may be thought of as his response to the totality of the reasons that he confronted in his deliberations. But what of cases in which an agent's ϕing consists in his *not* performing an action of a certain type? Consider again the case of John and his in-laws. Many people in John's situation would have spoken rudely. John did not. Perhaps there was no good reason for him to speak rudely to his in-laws, and his responding well to reasons consisted in his assigning no weight, either in thought or in practice, to considerations that others would have taken to support speaking rudely. Or perhaps John had an outweighed *pro tanto* reason to speak rudely to his in-laws, and his responding well to reasons consisted in his not letting that particular reason unduly influence his deliberations or his practice. Either way, no *act* of responding to reasons, no decision to refrain from speaking rudely, is presupposed by the claim that John responded well to reasons in not speaking rudely to his in-laws.

In this section I have focused upon a special class of my target judgments: those concerning whether agents acted well or badly in ϕing. But I think that the approach can be extended to judgments employing other adverbs. If so, perhaps all judgments of the form 'A acted F-ly in ϕing' concern

[36] I thus agree with Müller (2004) that acting well is a matter of responding well to reasons. But I do not agree that an agent's responding well (or badly) to reasons must always consist in his *doing something* for good (or bad) reasons. (See n. 30, above.)

agents' ways of responding to reasons. The difference is that judgments employing 'well' or 'badly' provide overall evaluations, whereas judgments employing other adverbs provide more specialized evaluations. Whether an agent acted courageously in ϕing, for example, might turn (roughly) on whether the agent, in ϕing, resisted the lure of the *pro tanto* reason to avoid danger. Whether an agent acted altruistically in ϕing might concern (roughly) whether the agent, in ϕing, sided with reasons to benefit others when those reasons conflicted with reasons to benefit himself. Whether an agent acted wrongly in ϕing might concern (roughly) whether there were decisive moral reasons for the agent to act in some other way. And so on.

REFERENCES

Aquinas, St Thomas (1265–73) *Summa Theologiae*. Tr. by the Dominican Fathers as *Summa Theologica* (New York: Benziger Brothers, 1948).

Bennett, Jonathan (1988) *Events and their Names* (Indianapolis: Hackett Publishing Company).

—— (1995) *The Act Itself* (Oxford: Clarendon Press).

Davidson, Donald (1967) 'The Logical Form of Action Sentences' in N. Rescher (ed.), *The Logic of Decision and Action* (Pittsburgh: University of Pittsburgh Press); reprinted in D. Davidson, *Essays on Actions and Event*s (Oxford: Clarendon Press, 1980).

—— (1971) 'Agency' in R. Binkley, R. Bronaugh, and A. Marras (eds.), *Agent, Action, and Reason* (Toronto: University of Toronto Press); reprinted in D. Davidson, *Essays on Actions and Events* (Oxford: Clarendon Press, 1980).

Geach, Peter (1956) 'Good and Evil' *Analysis* 17/2: 33–42; reprinted in P. Foot (ed.), *Theories of Ethics* (Oxford: Oxford University Press, 1967).

Hanser, Matthew (2005) 'Permissibility and Practical Inference' *Ethics* 115/3: 443–70.

Kant, Immanuel (1785) *Grundlegung zur Metaphysik der Sitten*. Tr. by L. W. Beck as *Foundations of the Metaphysics of Morals* (Indianapolis: Bobbs-Merrill, 1959).

Müller, Anselm Winfried (2004) 'Acting Well' in A. O'Hear (ed.), *Modern Moral Philosophy* (Cambridge: Cambridge University Press).

Thomson, Judith Jarvis (1997) 'The Right and the Good' *Journal of Philosophy* 94/6: 273–98.

—— (2006) 'The Legacy of *Principia*' in T. Horgan and M. Timmons (eds.), *Metaethics after Moore* (Oxford: Clarendon Press).

Vendler, Zeno (1967) 'Facts and Events' in his *Linguistics in Philosophy* (Ithaca, NY: Cornell University Press).

This page intentionally left blank

11

Hume on Practical Morality and Inert Reason

Geoffrey Sayre-McCord

> That's good, but right now I'm not interested in what's good; I'm a bad fellow.
>
> Cal Trask (James Dean), *East of Eden*

INTRODUCTION

David Hume's dramatic conclusions concerning the role of reason in practical life are well known. According to him, "reason is, and ought only to be the slave of the passions, and can never pretend to any other office than to serve and obey them" (T. 415).[1]

In serving and obeying the passions, Hume recognizes, reason can of course influence our behavior by changing our view of the world. It might inform us that four dollars are more than two, or that one course of action will have certain effects while another will not, or that what appears to be a glass of wine is one of water. If we are concerned to have more money rather than less, or concerned to bring about certain effects, or concerned to have wine rather than water, reason's conclusions will make a difference to what we do. Yet, when it comes to our concerns, to setting ends and adjudicating among them, reason not only takes

Thanks are due to Don Garrett, Mike Ridge, John Corvino, Adam Cureton, and Sophie Botros, for very helpful comments on earlier versions of this paper, and to audiences at NYU's Conference *Reason and Emotion* and at the Ohio State/Maribor/Rijeka Conference *Reason and Action* in Dubrovnik, as well as to the members of the Philosophy Department at Bowling Green State University, the Chapel Hill Metaethics Group, and the Wisconsin Metaethics Conference, for probing feedback.

[1] Quotations from Hume's *A Treatise of Human Nature* (1739–40) are indicated with a "T." followed by the page number.

a back seat to the passions, it remains utterly silent. Indeed, Hume maintains,

> ’Tis not contrary to reason to prefer the destruction of the whole world to the scratching of my finger. ’Tis not contrary to reason for me to chuse my total ruin, to prevent the least uneasiness of an Indian or person wholly unknown to me. ’Tis as little contrary to reason to prefer even my own acknowledg’d lesser good to my greater ... (T. 416)

Much as these preferences and choices might offend morality or prudence, they are not contrary to reason, as Hume understands reason. In fact, according to Hume, no preferences, choices, or actions can be contrary to reason. Nor, he claims, can reason have any influence upon the will without the cooperation of the passions, over which it has no say.

In contrast, Hume holds that preferences, choices, and actions can be contrary to (or conform with) morality and he holds as well that morality can, by opposing or approving of the preferences, choices, or actions, have an influence upon the will.

This contrast, Hume argues, shows that moral distinctions between right and wrong, good and bad, virtuous and vicious, cannot themselves be derived from reason (alone). ‘‘Reason is wholly inactive,’’ he writes, ‘‘and can never be the source of so active a principle as conscience, or a sense of morals’’ (T. 458).[2]

Hume’s argument has been incredibly influential. It has also been the source of a great deal of controversy. By and large, though, I believe Hume’s grounds for thinking that reason alone is inert have been misunderstood. That misunderstanding has been complemented by another, I think, concerning the way in which morality is supposed, by Hume, to be practical. These misunderstandings have gone hand in hand with seeing Hume as embracing two views about moral judgment: (i) *motivational internalism*, according to which moral judgments, sincerely made, are intrinsically motivating (to some degree), and (ii) *non-cognitivism*, according to which moral judgments are properly seen not as expressions of belief that might be true or false but as expressions of certain non-cognitive attitudes, that is, passions.

There is, I believe, good reason to think Hume was neither a motivational internalist nor a non-cognitivist. And, I will argue, there is good reason too to think that the arguments Hume actually offered do not commit him otherwise. As a result, Hume’s reliance on the arguments he offers causes

[2] While Hume does not dwell on the point, the same observations, considerations, and arguments, hold with respect to prudence as hold with respect to morality, and he thinks that the requirements of prudence, no less than those of morality, cannot be derived from reason.

no problem for the coherence of his position. But my purpose here is not so much to defend the over-all coherence of Hume's view (that would take going in to his positive account of moral judgment[3]) as to uncover what I think are his compelling arguments against the rationalism he was attacking.

THE (NOW) STANDARD READING

Hume's arguments, and the now Standard Reading of them, are pretty familiar in outline. Without trying to do the arguments justice just yet, let me recall the main line of thought.

A good place to start is with Hume's claim that reason is inert. Hume is clear that, when using the term "reason" strictly, he is referring to the capacity to (and/or the faculty by which we) determine truth and falsity. And the determinations of reason are those beliefs (or judgments, or opinions) of ours that are arrived at through reasoning. According to Hume, such beliefs (judgments, or opinions) emerge in one of two ways, either as a result of the comparison of ideas (when the reasoning is demonstrative) or as a result of inferences from matters of fact discovered by experience (when the reasoning is probable). In making claims about reason, then, Hume is referring to beliefs (at least those arrived at as a result of inference) and the processes by which we arrive at them.[4]

To the extent the beliefs are arrived at through demonstrative reasoning, Hume argues, they will concern only the realm of ideas. Yet, since "the will always places us in that of realities, demonstration and volition seem, upon that account, to be totally remov'd, from each other" (T. 413). Of course, the realm of ideas can quickly become relevant to volition, but only when the demonstrations have implications for things that are of concern to an agent.[5] Similarly, probable reasoning will be relevant to volition, but only when its conclusions are related to things that are

[3] I offer an account of Hume's theory of the nature and role of moral judgment in Sayre-McCord (1994). See also Garrett (1997).

[4] There is some reason to think that Hume distinguished among beliefs as between those that are the product of reason (i.e. some form of inference) and those that are caused by, but not inferred from, experience. This can make a difference to whether one thinks that in arguing that moral judgments are not a product of reason Hume is thereby arguing that they are not beliefs or only arguing that they are not inferred.

[5] The crucial steps in the argument are: "Abstract or demonstrative reasoning ... never influences any of our actions, but only as it directs our judgment concerning causes and effect." Yet "It can never in the least concern us to know, that such objects are causes, and such others effects, if both the causes and the effects be indifferent to us" (T. 414).

of concern to an agent. Whether one believes something as a result of demonstration or of probable reasoning, coming to believe it will have no influence on action if the agent is wholly indifferent to what is discovered.

Yet, for an agent to be other than indifferent is for her to be concerned with, or engaged by, the matter in question, and that is itself for her to have (in Hume's broad sense) a passion, to which it is related. Remove all such passions and the discovery of truths or the uncovering of falsehoods will influence the agent's actions not at all. In every case, Hume claims, reason's impact turns upon the presence of an appropriate passion. Beliefs cause action only if the agent also cares about what the beliefs are about. Reason alone, Hume concludes, is inert, since reason's influence depends on the passions. It is in this sense that reason is inevitably a slave to the passions.

Hume first offers this argument in Book II of the *Treatise*. There his aim is to establish the essential role of the passions in determining the will: no action, he argues, in the absence of the passions. When he refers back to the argument, in Book III, Hume's aim is to show that moral distinctions cannot be founded exclusively on reason.

In the Book III discussion, Hume contrasts reason with morality, arguing that "Morals excite passions, and produce or prevent actions. Reason of itself is utterly impotent in this particular. The rules of morality, therefore, are not conclusions of our reason" (T. 457).

In offering this argument (which I will refer to as the Motivation Argument), Hume simply adds to his Book II conclusion—that reason alone is inert—the observation that morality (presumably alone) is not inert. Morality alone, he suggests, provides a motive to action. If that is right, then morality cannot itself be (merely) a conclusion of reason, since (we have seen) the latter never, alone, provides such a motive, and an "active principle can never be founded on an inactive" principle (T. 457). Thus, while morality's rules may depend in part on reason, they must be in part the products of aversions or propensities—otherwise they could not themselves motivate action.

Rendering this as a valid argument requires some work and additions. Most commonly, people recast it along these lines: Moral judgments, alone, motivate. No judgments based on reason, alone, motivate. Therefore, moral judgments are not based on reason, alone. Put this way, Hume's claims about reason and morality become, specifically, claims about *judgments* (moral and otherwise). Thought of in this way, the argument relies on motivational internalism about moral judgment as a premise. And it has non-cognitivism about moral judgment as a pretty direct implication, since moral judgments could (on Hume's view) intrinsically motivate only if they were in some way expressions of aversions or propensities, and not merely beliefs.[6]

Few people who find these arguments in Hume think that, as they stand, they are fully compelling. For instance, a number of people accuse Hume of artificially restricting the reach of reason (by limiting it to the discovery of truth and falsehood and so ruling out, from the start, practical reason). Others accuse Hume of begging the question against those who hold that moral judgments themselves provide a compelling counter-example to his claim that beliefs alone never motivate. And still others accuse Hume of begging the question (in the other direction, so to speak) by assuming that moral judgments do, by themselves, provide a motive for action.

In addition, few people who find these arguments in Hume think that they are compatible with all that Hume himself seems to believe. For instance, Hume appears explicitly to reject motivational internalism about moral judgments when he notes that "'Tis one thing to know virtue, and another to conform the will to it" (T. 465) and he in any case acknowledges that people can recognize what is good or right and still, as the sensible knave does, utterly fail to feel its pull. Hume does of course want to explain morality's influence—its capacity to motivate. That is central to his project. But he does not assume it motivates everyone, nor does he assume that if it motivates someone sometimes, it motivates that person always. Quite the contrary.[7]

Moreover, later in the *Treatise,* Hume develops carefully a standard of moral judgment that parallels closely his account of our judgments of primary and secondary qualities. In each case, he maintains, our capacity to make the relevant judgments depends on our having experiences of certain kinds, but the making of the judgments is not to be identified with having the experiences. It is one thing to have the experience of something as being red, it is another (Hume recognizes) to judge that it is red. Similarly, Hume holds, it is one thing to have the experience of moral approval of something, it is another to judge that it is approvable. Along with judgment in these areas comes both the possibility of distinguishing how things appear from how they are, and the possibility, it seems, of having corresponding

[6] For influential interpretations along these lines see Harrison (1976), Stroud (1977), and Mackie (1980).

[7] Hume does advance—as an undoubted maxim—the claim that "no action can be virtuous, or morally good, unless there be in human nature some motive to produce it distinct from the sense of its morality" (T. 479). But whatever that distinct motive is, it will not be one provided *by morality*, let alone by morality alone, nor will it be more specifically the result of moral *judgment*. See Cohon (1997) for an interpretation of the motivation argument that avoids a commitment to non-cognitivism.

beliefs. Hume's careful and elaborate account of moral judgment thus suggests that he thinks that moral judgments, no less than judgments of shape and of color, express beliefs (even as these judgments depend on very different kinds of experience), which is incompatible with him advancing non-cognitivism.[8]

Finally, given Hume's official view that "Any thing may produce any thing" (T. 173), his apparently *a priori* determinations that reason alone cannot cause passions, volitions, or actions, and that passions must always be present seem quite dubious, to say the least.[9]

THE ARGUMENT AGAIN, WITH MORE DETAIL

These concerns about the force of Hume's arguments and the possibility of his advancing them consistently, given his other commitments, recommends revisiting them with more care.

It is important, first, to note that Hume's motivation argument plays out against the background of one argument—the only argument offered in full in both Book II and Book III—for thinking reason alone is inert (which is crucial to his argument for thinking that, since morality is practical, its standards are not derived from reason alone), which I will call the Representation Argument. The Representation Argument receives a bit less attention than the others Hume offers, but it is nonetheless central to understanding Hume's position.

The Representation Argument addresses a worry that can be put this way: Suppose beliefs can produce action only with the cooperation of passion. If reason can nonetheless produce passions or volitions, as well as beliefs, then reason alone—by producing both beliefs and passions or volitions—would be able, after all, to produce action.

But putting the worry this way continues a common mistake embedded in my initial description of the motivation argument. That mistake needs to be cleared up. It consists in thinking that Hume assumes—or is in some way committed to thinking—that reason alone cannot cause action, period. This is a more sweeping claim than he accepts and than his argument requires (though he does sometimes write as if he accepts this sweeping claim). And it is more sweeping than he can legitimately claim given his proper acknowledgment that causal relations can only

[8] See Sayre-McCord (1994) for an interpretation of Hume's account of the standard of moral (and other) judgment.

[9] For these and other worries about Hume's argument, see Harrison (1976), Stroud (1977), and Mackie (1980), as well as Botros (2006).

be established empirically. What Hume actually assumes—and needs—is that reason cannot cause action "*by contradicting or approving of it*" (T. 458).[10]

Here is how the Representation Argument goes: passions, as well as volitions and actions, are 'original existences' and contain no representative quality of the sort that would render them "a copy of any other existence or modification."[11] In other words, they do not represent things, either relations of ideas or matters of fact, as being a certain way. Truth and falsity, though, turn specifically on whether such representations conform or not to how things actually are. "Truth or falsehood consists in an agreement or disagreement either to *real* relations of ideas, or to *real* existence and matter of fact" (T. 458). As a result, passions, volitions, or actions are simply not the sorts of thing that can, themselves, be either true or false. And that means that it is impossible for them to be "oppos'd by, or be contradictory to" reason (T. 415; see also T. 458). This entails in turn, as Hume notes, that "reason can never immediately prevent or produce any action *by contradicting or approving of it.*"

Claiming this is, of course, perfectly compatible with holding that the process or the products of reasoning might immediately cause all sorts of things—headaches, or various pleasures, or particular passions, or the urge to move. And these might combine with various beliefs to prompt action.

[10] Hume does repeatedly, both in Book II and Book III, summarize his claim without limiting its scope. Of course, this *could* be because he actually believes the more sweeping claim. But there are multiple reasons for thinking that he does not rely on the sweeping claim. One is that his arguments do not establish it. Another is that he clearly holds that reasoning does cause some things (e.g. beliefs), so he cannot think that reason is utterly inert. Still another is that he offers the Representation Argument explicitly to "confirm" the conclusion of his initial argument in Book II, and, in Book III, he offers the Representation Argument alone to support the claim, as he puts it (unrestrictedly) that "reason is perfectly inert, and can never either prevent or produce any action or affection" (T. 458). Since there is no doubt that the conclusion of the Representation Argument is the more restricted claim that reason cannot cause action "by contradicting or approving of it," it would be uncharitable, to say the least, to think Hume saw that argument as confirming or establishing a conclusion that was unrestricted. These considerations are not decisive. But they do provide substantial grounds for thinking Hume is not actually relying on the unrestricted claim, especially if his position does not require it (which is what I argue in the rest of this paper).

[11] Some have objected to this claim, highlighting that our passions have intensional objects that mean they do have some representative quality. Annette Baier (1991) dismissed as silly (and unnecessary to the argument) Hume's view that passions have no representative quality. While she is right that that view is silly and Hume does not need it, Hume does need the claim that, whatever representative quality passions might have, it is not such as to render them copies that might then be true or false.

What Hume rejects is the idea that reason could have these results by *contradicting or approving* the pains or pleasures or urges.[12]

Taking this into account, and moving back to the core argument that the Representation Argument is supposed to support, it should change how we understand the putative contrast between reason and morality.

According to Hume, the power reason lacks is the power to influence action specifically by contradicting or approving of actions (or the passions and volitions that give rise to actions, in conjunction with an agent's beliefs). Morality, in contrast, does (according to Hume), have the power to influence actions by contradicting or approving of them (as well as the passions and volitions that give rise to them, in conjunction with an agent's beliefs).

If indeed morality can contradict or approve of things not because they are false or true, but on some other grounds, an important part of the contrast Hume needs will be in place. Morality, of course, can. Its terms of appraisal are not 'true' and 'false' but 'good' and 'bad', 'right' and 'wrong', 'virtuous' and 'vicious' and those appraisals can and do apply to things that cannot be either true or false. The fact that passions, volitions, and actions are not representational, in the way required for them to be true or false, poses no obstacle to morality approving or disapproving them. Moreover, the moral standing of various actions can, according to Hume, make a difference to what people do. "The merit and demerit of actions," Hume notes, "frequently contradict, and sometimes controul our natural propensities." Yet "reason," Hume observes, "has no such influence" (T. 458). And that is true whether or not reason sometimes causes actions: reason never does so by contradicting or approving the action, because, in principle, it cannot.

According to Hume, for a passion, volition, or action to be contradicted by morality is for it to run afoul of the standards of virtue and vice that would secure approval from the general point of view. And according to him, sometimes, the fact that morality disapproves an action does have an impact on behavior. "[M]en are often govern'd by their duties, and are deter'd from some actions by the opinion of injustice, and impell'd to others by that of obligation" (T. 457). Often, but not always.

Here, as elsewhere, Hume avoids claiming that morality always has an influence. What he does claim, and needs to claim, is that morality can contradict or approve of actions and can, by contradicting or approving of

[12] The limit on reason's causal powers, such as it is, is no offense against Hume's general doctrine that particular causal relations are all established only by experience. But it is a limitation discovered a priori and depends on Hume being right both that reason contradicts and approves only what is capable of being true or false and that passions, volitions, and actions can never be either.

them, have an influence on the will. That is enough for his purposes, given that reason cannot contradict our behavior at all. For that is enough to show that the standards of morality, which enable it to approve or contradict passions, volitions, and actions on grounds other than truth and falsity, must go beyond what reason alone can provide.

NO ACTION ABSENT PASSION

While I will not go through the arguments here, it is worth noting that Hume considers various ways one might try to show that, ultimately, the morality of an action can actually be traced to truth or falsity, so that moral distinctions might after all prove to be in the ken of reason. He looks both to the causes of actions and to their effects, as well as to the relations between actions and the circumstances in which they might be performed. In each case, he argues, moral distinctions between right and wrong, virtue and vice, do not coincide with truths that reason might discover. He bothers to do this because the success of any of these proposals would establish, contrary to his Representation Argument, that reason might, after all, contradict or approve of actions after all, albeit by discovering of their causes, or effects, or relations something true or false.

In addition to various specific replies to proposals, Hume offers a general argument that moral distinctions cannot be a matter simply of truth and falsehood. If they were, he points out, there would be no moral difference "whether the question be concerning an apple or a kingdom, or whether the error be avoidable or unavoidable" (T. 460). Moreover, since truth and falsehood do not admit of degrees, such a view would provide no grounds for distinguishing among the virtues or among the vices as between those that are better or worse.[13] Significantly, Hume is not here (or elsewhere) arguing that moral judgments might not be true or false or that reason, in light of experience, could not have a role in discovering which are true and which false.[14] Hume's crucial point is that what makes the true judgments

[13] Hume's strategy, when it comes to the proposal that moral distinctions might be derived from reason because they are constituted by relations, is to argue that the distinction between virtue and vice does not line up with whether or not any particular relation holds. He actually goes further and argues that unless there is some heretofore unidentified relation that holds only between the passions, beliefs, volitions, and actions of sentient agents, on the one hand, and the agent's circumstances on the other, there is no hope for founding moral distinctions on relations.

[14] Hume does, of course, famously argue that we cannot infer moral conclusions from a set of exclusively non-moral premises. So reason's role in discovering the truth of moral judgments is limited by the need for moral input that it cannot itself provide. See T. 469.

true cannot be a matter of the actions judged conforming (or not) to reason.

Hume does not stop the argument there, however. Suppose that there is a way around these worries and one or another of the proposals were to work. In that case, the morality of an action would have been found to coincide with the truth or falsehood of an action, or its causes, or its effects, or with the presence or absence of a relation discoverable by reason alone.

Even then, Hume argues, the discovered coincidence of some truths and falsehoods with what is moral and immoral would leave us without an account of what makes the particular falsehoods immoral. What is still required, Hume maintains, is a "plausible reason, why such a falsehood is immoral. If you consider rightly of the matter you will find yourself in the same difficulty as at the beginning" (T. 462 n.). We would still require an account of why some falsehoods are immoral, while others are not (which means we would effectively be in the same situation we were in when asking why some actions are immoral, while others are not).

But suppose in addition some such account was given. In that case, moral distinctions might, Hume grants, be derived from reason. Yet would reason's right to rule the passions then have been established as well? Only if, Hume holds, the relevant truths, when discovered, could successfully govern behavior. And this is because, in practical matters, part of the proof is in the performance. If reason's distinction between the moral and the immoral were of no concern, and no further operations of reason could work to make them so, then they would have no dominion over the passions.[15]

Might reason alone, though, without the aid and cooperation of the passions, somehow secure its right to rule? Hume is skeptical, to say the least. But why? In this context, his claim that reason cannot contradict or approve of action and so cannot cause action by contradicting or approving of it, is not relevant. We are granting, for the sake of the discussion, that actions can be contradicted or approved of by reason thanks to their causes or their effects being true or false, or thanks to their standing in certain relations that are discoverable by reason. The question is whether somehow reason alone might, under these circumstances, work to govern and sometimes control action in opposition to the passions.

Hume pretty clearly thinks not. While he is prepared to allow that, in the circumstances, reason might influence the will by contradicting and approving of actions, its effectiveness will depend upon, and not be wholly

[15] See esp. T. 465.

in opposition to, the passions. But why does Hume think reason is even here powerless in the absence of passion?

Although Hume does not offer this argument, one might defend his position in the following way: However alike two people might be in what they believe and how they came to believe as they do, they might respond differently to the situation they take themselves to be in. The differences in response cannot, by hypothesis, be explained by differences in the operation of reason. But they can and indeed should be explained by appealing to differences in their aversions and propensities, that is, in their passions. What one is led to do by one's beliefs thus depends upon one's passions. If one were altogether passionless, one would not be led to do anything. Something along these lines seems to be behind Hume's willingness to postulate passions so calm their presence is imperceptible.[16]

Alternatively, one might defend Hume's position by distinguishing between, on the one hand, actions and the considerations that motivate them, and, on the other hand, mere behavior and the things that might cause them. Hume neither articulates nor defends this distinction. Nonetheless, the distinction is consistent with Hume's views, I think, and can be used to defend the idea that in every case where something serves as a motive to action (rather than merely as a cause of behavior) a passion is in play.

What the distinction does is mark a difference between something being an influencing motive of the will and it being a mere cause of behavior. While all cases of motivated action are cases in which something causes behavior, not all instances of the latter are instances of motivated action. Saying what exactly makes the difference is, of course, a terribly tricky business. But for our purposes the key point is that, when actions (and not mere behavior) are at issue, the considerations that cause the behavior, and serve as the agent's motives, must be related so as to render her behavior intelligible as a case of the agent pursuing her aims or goals or, more loosely, as her behaving as she is concerned to do.

[16] There is an issue here, though, about how seriously to take Hume's claim that the passions are themselves *perceptions* in the mind and not, say, *dispositions* of the person. The argument goes much more smoothly on the latter view than on the former. The mere fact that there must be a difference that explains the difference (between those moved by the beliefs and those not), might well show more or less trivially that there is a difference in dispositions to respond to the beliefs. But it wouldn't show that those dispositions are themselves perceptions in the mind as opposed to tendencies to respond to perceptions. Fortunately, by and large, Hume's characterizations of passions, in the context of explaining actions, seem to suggest the dispositional rather than the mental entity view. Thus, he notes that certain desires and tendencies "are more known by their effects than by the immediate feeling or sensation" (T. 417).

With the distinction in mind, we can say that any consideration that causes behavior, but is of no concern to the agent, is not a motive of hers, and the behavior it causes will not count as an action performed by the agent. We can also say that if a consideration does motivate an agent to perform an action, then it was not a matter of indifference to her. "It can never in the least concern us to know, that such objects are causes, and such others effects," Hume thinks, "if both the causes and the effect be indifferent to us" (T. 414) and such knowledge, whatever its effects might be, will not serve as a motive of the will. Exactly the same line of thought applies where the knowledge is of eternal relations rather than causal connections: it can never in the least concern us to know a relation obtains if both the relation and the things that stand in the relation are indifferent to us. Thus, whether the knowledge is of a matter of fact or of a relation among ideas, it will be a motive for the agent only if it is of concern to her, and for it to be of concern is for it to be the object of a passion.

While Hume does not himself press this distinction between a mere cause of behavior and a motive for action, he seems careful to respect it in his discussion of the influencing motives of the will. Most notable there is Hume's focus on the relevant influences on the will being passions that have objects, and that motivate action, by influencing the will, only in light of discoveries about those objects. The particular discoveries in turn motivate as well, it must be said, but only as the discoveries relate to objects that are of concern to the agent. So a full articulation of the influencing motives of the will has to appeal both to the agent's beliefs and to her concerns.

The underlying idea is that an agent's beliefs work as they do *to direct action* only as the beliefs are related to what is of concern to the agent. And for something to be of concern to an agent—for her not to be indifferent to it—is for it to be the object of one of her passions. To imagine an agent utterly indifferent to the world is just to imagine an agent with no concerns, that is, with no passions. Whatever such an agent might discover, she will remain unmoved unless and until she loses her indifference and acquires a relevant concern.

For something to be of concern to us involves our having an aversion or propensity of some sort with regard to it, so that we are disposed to act one way or another in light of discoveries concerning it. Propensities and aversions of this sort simply are (as Hume uses the term) passions. Sometimes they are calm, so calm as to go unnoticed, at other times they are violent and impossible to miss. But at all times, if an agent is motivated to act, they are present.

While this claim may sound like a substantial empirical hypothesis, the reasoning that leads to the conclusion is utterly insubstantial and treats as perfectly trivial an inference from the fact that an agent performed an *action* to her having had, in the sense Hume requires, a relevant propensity or aversion (i.e. passion). In effect, the necessity of passion, when it comes to identifying an agent's motives for action, is treated as an analytic truth. It leaves open discovering in particular cases that no relevant passion was present, in which case the behavior won't count as an action, and it leaves open discovering that what motivates some people is radically different from that what motivates others. What it closes off is the thought that a consideration might motivate action in the absence of a passion, since that would be for the consideration to motivate in the absence of motivation.

This means that, even if moral distinctions were, in some way, discoverable by reason, knowledge of them would still serve as a motive of someone's actions only in light of her passions, whether these are "certain instincts originally implanted in our natures, such as benevolence and resentment, the love of life, and kindness to children; or the general appetite to good, and aversion to evil, consider'd merely as such" or still some other propensity or aversion (T. 417). The required passions may be, he acknowledges, so calm as to be indistinguishable from reason in their operation, and they may be "more known by their effects than by the immediate sensation," yet if the effects (on action) are there, so too must be they.[17]

DOES MORALITY ALONE MOTIVATE?

But if, as Hume seems to hold, there is no action without (at least calm) passion, how is it that morality *alone* might be an influencing motive of the will? Won't morality's impact on action depend on the presence of an independent (albeit, perhaps calm) passion?

One answer would involve holding that morality has an impact because our moral opinions themselves are, or at least involve having, certain concerns. On this view, such opinions are not (or not merely) beliefs that might be true or false, but are instead (or in addition) motivating states of the agent who holds them. This suggestion fits reasonably well with one natural reading of Hume's famous claim that if you examine any vicious act,

[17] This argument, if it works, establishes that in every case where a person performs an action, she must have had a relevant passion. It does not establish that every consideration that works to motivate action does so by answering to a *pre-existing* passion. Hume does believe there are some dispositions implanted by nature, but his argument for thinking there is no action absent passion is not an argument for such dispositions. For all the argument shows, the required passions might come new on the scene with the recognition of the conditions in which one finds oneself.

> you find only certain passions, motives, volitions, and thoughts. There is no other matter of fact in the case. 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, towards this action.

While there are complications in treating this passage as a defense of the idea that opinions of injustice and obligations are feelings, not beliefs, none of the complications seem insuperable.

Nonetheless, this non-cognitivist answer fits poorly, as I mentioned above, with a number of things in the *Treatise*. For instance, it is hard to reconcile with Hume's recognition that people can intelligibly acknowledge what morality requires and yet remain unmoved, either because of weakness of will or because of doubts about morality's authority. And it is also hard to reconcile with Hume's careful and detailed account of a standard of moral judgment that is so directly modeled on the account he offers for other judgments that he clearly thinks express beliefs.

If, as I am inclined to think, Hume thought that moral judgments expressed genuine beliefs (albeit beliefs the having of which depended upon the capacity to feel approbation and disapprobation), then it seems they, like all other beliefs, will succeed in motivating an agent only if the agent has certain propensities or aversions. In what sense, then, could morality count, any more than reason, as able alone to influence action? Will not its effect always depend on the presence of an independent passion?

I think not, but to explain why it is necessary to shift attention for a moment to Hume's account of the operations of reason. Reason alone doesn't influence the will, but it does influence beliefs. Yet when *reasoning*, alone, influences belief, it is not *belief* alone that has that effect—a belief's effect depends on the operation of certain dispositions—certain habits of mind—the having of which is partially constitutive of reason. Thus while Hume sees reason alone as unable to serve as an influencing motive of the will, he sees the activity of reasoning—from cause to effect or concerning the relations of ideas—as wholly a matter of reason's operations, even as he also recognizes that these inferences are explained necessarily by appeal to dispositions of the mind that are not themselves inferences or conclusions of reason.[18]

Reason can contradict or approve certain conclusions given certain ideas or present impressions and it can, by contradicting or approving those conclusions, sometimes influence belief. But, Hume is clear, even in these cases reason's influence on belief depends upon the mind being,

[18] See Book I, Part III, Section VIII, "Of the causes of belief."

in the relevant way, "well-disposed." What allows this requirement to be compatible with thinking that reason *alone* can influence belief is that the dispositions upon which the inferences depend are the dispositions the having of which constitute one as having a rational mind. In noting that they are required we are not thereby appealing to something that is not a part of reason.

In the same way, I think, Hume considers certain propensities and aversions (specifically, certain passions) that combine with beliefs to motivate behavior, as dispositions the having of which constitute one as a moral agent. These passions are, in the relevant sense, not independent of morality, even as they are not moral beliefs. To count as having a well-disposed mind, from the point of view of morality, one must be concerned with, and so moved by, certain kinds of considerations. Thus, for instance, to be benevolent is to be moved by a recognition that others are in need, and to be just is to be restrained by the thought that something belongs to another. Many of these dispositions (all the dispositions that constitute the natural virtues) are available prior to convention. However, some (those that constitute the artificial virtues) require the existence of conventions. And some of these last—for instance the disposition to be moved by the thought that so acting is one's duty—require specifically the conventions that make possible the thought that something is one's duty. Hume's acknowledgment of the essential role of the passions is compatible with thinking that morality *alone* can influence action precisely because the dispositions upon which the actions depend are dispositions the having of which constitute one as being a moral person. In noting that these are required we are not thereby appealing to something that is not a part of morality.[19]

The contrast with reason is therefore still in place. The dispositions that are required by, and partially constitutive of, reason are dispositions to reach various conclusions in light of experience or reflection on ideas. They are not dispositions to act in light of the conclusions one reaches. Whereas, the dispositions that are required by, and partially constitutive of, morality are dispositions to act in various ways in light of certain considerations. The former cannot explain action (as such) without appealing to passions that are not required by reason, whereas the latter can, sometimes, explain action by appealing only to dispositions required by morality.

[19] It is worth emphasizing that, on this account, in saying that morality alone motivates one is not saying that moral beliefs (or judgments) alone motivate. The capacity of moral beliefs (or judgments) to motivate still depends upon the presence of a relevant passion. The important point is that at least sometimes the requisite passion is itself properly regarded as something the having of which is a part of what it is to be moral.

THE WORRIES PROMPTED BY THE STANDARD READING RECONSIDERED

Early on in the paper I mentioned a number of complaints people (rightly) have against Hume's core argument, as it is standardly understood. I want now to go through those complaints, with the alternative understanding on the table. I will go back through them in reverse order, starting with the concerns that focused on apparent inconsistencies in Hume's own view.

Illegitimate A Priorism

The first such concern focused on Hume's explicit commitment to thinking causal connections are discoverable only a posteriori. As he writes, "there is no connexion of cause and effect ... which is discoverable otherwise than by experience, and of which we can pretend to have any security by the simple consideration of the objects" (T. 466). Yet, on the standard understanding of Hume's argument he seems to be declaring a priori both that reason can never alone cause action and, thanks to his apparent internalism, that morality alone always does.

If, as I maintain, Hume's only a priori claim here is the *negative* causal claim that reason cannot cause actions by contradicting or approving of them, he is on safe ground. Hume's support for this is, on the one hand, that reason can approve or contradict something only by finding it either true or false, and on the other hand, that actions are not the sort of thing that can be true or false. This is an argument that turns not on the evidence provided by experience, but instead solely on the comparison of our ideas of 'reason', 'truth and falsehood', and 'action', and they are such that we can pretend to have some security concerning them. Hume may of course be wrong about our idea of 'reason', or of 'truth' and 'falsehood', or of 'action'. But if he is not, reason's inability to cause an action by contradicting or approving of it is secure and consistent with his claims about what is required to establish positive causal claims.

The distinction between negative and positive causal claims is no help, though, in defending Hume against an inconsistent *a priorism* if he is embracing internalism a priori. Even here, the inconsistency is not inevitable. Hume could in effect simply stipulate that a judgment does not count as a moral judgment unless the person making it has some motive to act accordingly. And he might defend this on the grounds that, given the distinction between the speculative and the practical, "morality is always comprehended under the latter division." But what is agreed to on all sides, when morality is counted as practical, is clearly not that moral judgments always provide some motivation. So in embracing internalism stipulatively, Hume would almost certainly be relegating his argument to the sidelines or begging the central question. Alternatively, Hume might be seen as holding his internalism as an empirical thesis. But then the evidence he has would be woefully weak, especially in light of his own recognition that in a lot of cases no motive seems apparent (and if it is present it is only because the passions can be so calm as to be imperceptible).

Fortunately, if I am right, no part of Hume's argument requires internalism, a priori or otherwise. Hume does think—and this is granted on all sides—that moral distinctions can and commonly do make a difference to how people act. And he recognizes (in a way not everyone does) that this imposes an important constraint on accounts of those distinctions: the accounts must be able to make sense of how and why the distinctions make the difference they do to how people act. What any account of morality needs to do is explain how it is that the connection between morality and the will "is so necessary, that in *every well-disposed mind*, it must take place and have its influence" (T. 465, my italics).

Hume's Apparently Cognitivist Account of Moral Judgment

The second concern was that Hume offers a positive account of moral judgment that puts it on all fours with what are indisputably beliefs. Moral judgments, are, he clearly argues, bound up with our sentimental constitution and our capacity to feel approbation and disapprobation, in much the way that our judgments of color are bound up with our capacity to have color experiences. In both cases, though, Hume is at pains to distinguish the sentiments and experiences that are required from the beliefs we might make in light of them.

But if, as the standard reading of the core argument would have it, Hume thought that moral judgments necessarily motivate, while beliefs only contingently motivate, he would be committed to saying moral judgments are not, after all, beliefs.

If I am right, though, his argument does not depend on holding that moral judgments necessarily motivate. So while he is committed to thinking beliefs motivate only contingently, he can hold the same view of moral judgments, without undermining his argument for thinking moral distinctions are not derived from reason.

Begging the Question against Externalists

Once the idea that Hume is committed to internalism is put to one side, accusations that he begs the question against externalists simply lose their grip. Hume does hold that the connection between morality and the will must be explained, but he does not think that explanation requires that our moral opinions are intrinsically motivating, nor does he think that the explanation will do without an appeal to the aversions and propensities of those who count as having a "well-disposed mind." It is worth emphasizing, though, that Hume does think an acceptable theory of moral judgment will need to account for the intimate connection there is between those judgments and our motivations. While that connection is not so tight as to guarantee the presence of a motivation whenever someone forms a moral opinion, it is nonetheless tight. Specifically, Hume thinks in the normal case people are motivated to act as their judgments would endorse and that fact is not, Hume thinks, a mere coincidence. Indeed, if no such connection existed, the judgments could not, he believes, make out their claim to allegiance. While this is not the place to go into Hume's positive theory of moral judgment, it is worth mentioning that he thinks the capacity to feel moral sentiments (which are not themselves judgments) is as crucial to moral judgment as the capacity to have visual experiences is crucial to visual judgments. Morality's ability to motivate action is bound up with the role the moral sentiments have both in making moral judgments and in constituting the standard by which those judgments are to be counted as correct.[20] The judgments might, in particular cases, be made without consulting, or even having, moral sentiments. But if there were no such sentiments at all, Hume holds, the judgments would not be possible. And while the standard for the judgments is not set by the sentiments people actually feel, it is a standard set by the sentiments they would feel were they to correct their view in appropriate ways.

Begging the Questions against Cognitivist Internalism

What about those who hold that moral judgments do necessarily motivate and that, precisely because they are beliefs, they serve as counter-examples

[20] To say that morality's ability to motivate is bound up with the role of moral sentiments is not to say that those sentiments themselves necessarily motivate. There is at least some reason to think that Hume sees the moral sentiments of approbation and disapprobation, which are particular kinds of pleasure (on his view), as having no specific motivational implications. Certainly he thought we might approve of certain characters without having any particular motivation. At the same time, though, the prospect of acting in ways that we ourselves would approve or disapprove of would have the kind of implications for action that Hume thinks the prospect of pleasure or pain regularly has in humans.

to Hume's claim that reason alone can never motivate? This seems directly to undercut Hume's claim that reason and its products, alone, cannot cause action in the absence of a passion.

It is worth noting that the claim that beliefs, in the absence of a relevant passion, are inert, is compatible with holding that some genuine beliefs actually always motivate thanks to the inevitable presence of a relevant passion. This is Hume's explicit view concerning beliefs one has about the prospect of pleasure or pain for oneself. "'Tis obvious," he says, "that when we have the prospect of pain or pleasure from any object, we feel a consequent emotion of aversion or propensity, and are carry'd to avoid or embrace what will give us this uneasiness or satisfaction" (T. 414). According to Hume, the prospects of pleasure and pain do always motivate. But this is because, and only because, we have a propensity to pleasure and an aversion to pain.[21] In a parallel way, Hume could acknowledge that, as a matter of fact, moral judgments always motivate. Of course this falls short of holding that moral judgments necessarily motivate regardless of one's concerns. But, if I am right, Hume has a fairly compelling a priori argument against that view.[22] What that argument leaves open is a view according to which the conditions for making moral judgments guaranteed the presence of a relevant passion. According to such a view, moral judgments would necessarily motivate, but they would not do so regardless of one's concerns.

In any case, while Hume neither defends nor assumes internalism about moral judgment, were there good arguments for it, nothing would preclude his accepting it and accepting as well the idea that the judgments were themselves beliefs that are the product of reasons. He could and would still hold, though, that the impact of those beliefs depend on the presence of a relevant passion.

Artificially Restricting the Reach of Reason

There is no denying that Hume draws a fairly sharp and clear line around what he will count as reason. He does, I've suggested, include within its scope not only the faculty of reason and its operations (inferences) and its products (beliefs, doubts, conclusions ...) but also the habits and dispositions that make it possible for reason to work as it does. Nonetheless, from the start he seems to have excluded out of hand just what many have wanted to defend: the idea of a truly practical reason that moves one from various premises to action.

[21] As Hume notes, "there is implanted in the human mind a perception of pain and pleasure, as the chief spring and moving principle of all its actions" (T. 118).

[22] The argument is only fairly compelling, not decisive, since one might insist that two people with the same moral beliefs must equally have the same motivations.

But that may not be fair to Hume. Hume is perfectly willing to talk of standards of morality that countenance or condemn acting on the basis of certain considerations. And, though he says little about it, Hume seems equally willing to talk of a standard of prudence that countenances or condemns acting on the basis of certain considerations. In both cases, Hume evidently has no difficulty with the idea that such standards might exist and be properly influential in our thinking and acting. What he would reject, with respect to these standards, or any others that might be advanced, is that their credentials might be established independently of their authority, or that their authority might be established without showing how they might successfully govern. To do this last, one must show that the truths on offer are such that adopting them as a standard for action will solve well the practical problems that give rise to the need for a standard in the first place. Doing this is inevitably a matter of showing how those standards might actually have a grip on all who are "well-disposed."[23]

With that in mind, consider the common suggestion that it is irrational to will an end and not will the necessary means to achieving that end, or irrational not to will what one believes to be the necessary means, or irrational not to will the most efficient means (or what one takes to be the most efficient means), or irrational not to will the best means (or what one takes to be the best means) to one's ends.[24]

Whichever of these one might accept, it could be offered (as Kant offers his version) as an analytic truth or merely as the correct substantive standard of practical rationality. Either way, actions will count as rational or not (in *this* sense) not in virtue of being true or false but rather in virtue of being appropriately related to some such standard. What is relevantly true or false are claims concerning what the correct substantive standard is. There are two points Hume would make about any such standard, however it is defended. One is that an agent could satisfy the standard (i.e. take the appropriate course of action in light of her ends) only thanks to having the appropriate aversions and propensities. Knowing the relevant truths would not be enough.[25] The other is that one can reasonably ask, of any such standard, whether it matters whether one is rational in *that* sense. No

[23] The thought here is that Hume's account of the authority of practical standards will, like his account of political legitimacy, make actual effectiveness a necessary condition. Just as a ruler's claim to legitimacy depends on his capacity to govern effectively, so too will a standard's claim to authority.

[24] Alternatively, one could consider the less often defended but often acted upon "That'll teach 'em" principle, according to which the appropriate response to frustration is to lash out.

[25] Kant clearly appeals to reverence for the law to account for how it is that real agents are moved by recognition of the categorical imperative. Does he recognize the need for something similar to account for the effectiveness of the hypothetical imperative? If the

answer that stops short of engaging the concerns of agents will work, and thus an answer that works does so by engaging not merely the intellect but the heart.

CONCLUSION

If I am right, the arguments Hume offered against rationalism do not depend on motivational internalism nor do they entail non-cognitivism. And while they do depend on some analytic truths (e.g. "what cannot disapprove or approve of actions cannot cause actions by disapproving or approving of them"), this is not a matter of Hume mobilizing a *priori* constraints on causation of a sort that he could not countenance. Moreover, they are fully compatible with Hume taking seriously the possibility that some people might be unmoved by moral considerations even as they recognize them and compatible too with Hume developing a substantive standard of morality in light of which some moral opinions are true.

What the arguments preclude is thinking that the truth of such opinions is independent of what might motivate those who are subject to moral demands. As a result, the arguments imply that not everyone who fails to be moral is properly criticized as having been irrational (as opposed to immoral) in Hume's sense. One might, of course, expand the notion of rationality so as to be able to decry all immorality as irrationality. But then one would have expanded the notion of rationality to the point where it makes sense to wonder whether people have reason to be rational. That this would make sense is, though, no objection in itself. The problems would come only if it turned out that, by the very standard of rationality on offer, there was no reason to be rational. In that case, we would have discovered reason to be concerned not with what is rational (in this extended sense) but with something else.

REFERENCES

Baier, Annette (1991) *A Progress of Sentiments* (Cambridge, MA: Harvard University Press).

Botros, Sophie (2006) *Hume, Reason and Morality* (London: Routledge).

Cohon, Rachel (1997) 'Is Hume a Noncognitivist in the Motivation Argument?' *Philosophical Studies* 85: 251–66.

position I have attributed to Hume regarding the motives (as opposed to mere causes) of action (as opposed to mere behavior) is right, he would need to.

Garrett, Don (1997) *Cognition and Commitment in Hume's Philosophy* (Oxford: Oxford University Press).

Harrison, Jonathan (1976) *Hume's Moral Epistemology* (Oxford: Oxford University Press).

Hume, David (1739–40) *A Treatise of Human Nature*, ed. L. A. Selby-Bigge and P. H. Nidditch (Oxford: Oxford University Press).

Mackie, J. L. (1980) *Hume's Moral Theory* (London: Routledge).

Sayre-McCord, Geoffrey (1994) "On Why Hume's 'General Point of View' Isn't Ideal—and Shouldn't Be" *Social Philosophy and Policy*, 11/1: 202–28.

Stroud, Barry (1977) *Hume* (London: Routledge & Kegan Paul).

Index