OXFORD

Oxford Studies in Metaethics

Volume 2

OXFORD STUDIES IN METAETHICS

This page intentionally left blank

Oxford Studies in Metaethics

VOLUME 2

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 2007

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 978–0–19–921807–3
ISBN 978–0–19–921806–6 (Pbk)

1 3 5 7 9 10 8 6 4 2

Contents

Notes on Contributors

Stephen Darwall is John Dewey Collegiate Professor of Philosophy, University of Michigan

David Enoch is Senior Lecturer in Philosophy, and Cardinal Cody Lecturer in Canon Law, the Hebrew University in Jerusalem

Stephen Finlay is Assistant Professor of Philosophy, University of Southern California

Joshua Gert is Associate Professor of Philosophy, Florida State University

Patricia Greenspan is Professor of Philosophy, University of Maryland

Robert N. Johnson is Associate Professor of Philosophy, University of Missouri-Columbia

Mark N. Lance is Professor of Philosophy, Georgetown University

Margaret Olivia Little is Associate Professor of Philosophy, Georgetown University

Michael Ridge is Reader, University of Edinburgh

T. M. Scanlon is Alford Professor of Natural Religion, Moral Philosophy and Civil Polity, Harvard University

Mark Schroeder is Assistant Professor of Philosophy, University of Southern California

Judith Jarvis Thomson is Professor of Philosophy, Massachusetts Institute of Technology

Introduction

Russ Shafer-Landau

Oxford Studies in Metaethics is devoted to providing an annual selection of some of the most exciting new work in the foundations of ethics. I am pleased that this aim has been so successfully met in this second volume of the series.

The entries begin with an essay by T. M. Scanlon, in which he offers his latest thoughts on a central metaethical topic: the relationship between wrongness and practical reasons. Scanlon's work has been very influential in this area. Here he offers a reappraisal of that work, situated in a context that seeks to account for the shift from motivational concerns to those centrally to do with practical reason, within the literature devoted to metaethics.

David Enoch next offers a general argument in favor of non-naturalistic normative realism—the idea that there are non-natural, irreducibly normative truths, objective and universal in nature. The argument is modeled on those in many areas of philosophy that seek to vindicate something's existence by displaying its explanatory indispensability. Enoch, however, modifies this form of argument with an eye to showing that robust normative truths are *deliberatively* indispensable—that our practices of practical deliberation require the assumption that there are such truths.

Michael Ridge sees things quite differently, approaching the central metaethical questions from the opposite end of the spectrum from Enoch's normative realism. Ridge offers a defense of what he calls 'ecumenical expressivism,' which is the thesis that normative sentences are conventionally used to express both beliefs and desires. He confines his work here to developing this brand of expressivism, and arguing that it is superior to traditional expressivist accounts, which limit moral judgment to the expression of some essentially non-representational, practical attitude.

Joshua Gert's contribution picks up themes discussed in both Enoch's and Ridge's articles. Gert is concerned with the breadth of disagreement on normative matters. He finds it fruitful to begin the investigation with

the case of color terms, and claims that cognitivist analyses of such simple notions are most appealing for cases in which there is great agreement about their extension. Once we move along a spectrum of response, to the point at which there is very great disagreement, an expressivist analysis becomes more appealing. So too, claims Gert, for matters involving the application of non-complex normative notions; some will be best construed as cognitivists would do, while others are best understood as expressivists would recommend.

Stephen Darwall next presents a new basis for understanding the essence of morality's reason-giving power. Darwall proposes that traditional efforts to argue for morality's ability to provide categorical, overriding reasons for action invariably omit a crucial element: the second-person perspective. Darwall argues that understanding moral responsibility requires that we take this perspective quite seriously. He then argues that there is a conceptual connection between moral responsibility and moral obligation that enables us to appreciate the availability of a new kind of argument, one from the second-person perspective, for the categorical nature of moral obligations and reasons.

Darwall's work here, as elsewhere, is undertaken from within a broadly Kantian framework. Robert Johnson next explores some of the nuances of this outlook in his consideration of a foundational question for Kantians, and for metaethicists generally: is the normative authority of moral obligations grounded in what has unconditional value, or does it have some other source? Recently a number of scholars have argued that the authority of moral obligation must derive in some way from the value of humanity, or the good will. Johnson seeks to resist that line of thought, and to argue instead that a traditional view of Kant's project, one that underwrites normative authority by invoking our capacities for autonomous agency, is correct.

One of the most exciting areas in the intersection of normative ethics and metaethics these days lies in research being done on the merits of ethical particularism. Mark Lance and Margaret Little here present a piece of their growing body of collaborative work in this area. They agree with particularists that the moral generalizations we are apt to rely on are replete with exceptions. But they reject the lesson that particularists seek to draw from this, namely, that moral rules either are non-existent, or are practically useless. Rather, Lance and Little seek to vindicate the existence of moral rules whose importance lies, at least in good part, in the explanatory work that they are able to do. What they resist is the idea that such work can be done only by exceptionless rules or laws. In order to substantiate such a view, it is crucial to explain the notion of a defeasible generalization,

and defend its philosophical importance. This is the project they have set themselves in the article on offer here.

The remaining articles focus on the themes of normativity and practical reasons that have taken center stage in so much recent metaethical theorizing. Patricia Greenspan first tackles one of the deepest metaethical concerns: how moral obligations might provide categorical reasons for action. Standardly, defenders of such a claim have sought to show that those who deliberately flout their acknowledged moral obligations are in some deep way irrational. Greenspan argues that this is a mistake: we can, if she is right, defend the existence of categorical moral oughts and reasons while allowing that one can rationally fail to be motivated by such acknowledged considerations.

Mark Schroeder next defends the so-called Humean theory of reasons, according to which all of one's reasons are explained by reference to a psychological state (such as a desire) of the agent for whom they are reasons. Schroeder seeks to account for a claim that has struck so many as extremely plausible, namely, that some reasons apply to every agent, while others, such as those that direct us to pursue stamp collecting or marathon racing, apply only to some. Rather than developing the common argument for the theory, according to which reasons must be capable of motivating, and motivation must stem from an agent's desires, Schroeder here offers a novel argument for the Humean theory, based on quite general philosophical methodological principles.

Stephen Finlay shares Schroeder's enthusiasm for the Humean theory, and offers his own defense of the specific variation on it that claims that all practical reasons must stem from our desires. Indeed, if Finlay is correct, the normative authority of not only reason, but value and obligation, can be comprehensively explained by reference to our desires. He too offers a novel argument, which he calls the *Argument from Voluntary Response*, that seeks to lay the foundations of a new vindication of the Humean theory.

This volume closes with Judith Jarvis Thomson's latest thoughts on the nature of normativity. Thomson considers another perennial issue in ethical theory: what the connection might be between evaluative claims (to the effect that things are good or bad, well made or defective, delightful or awful, etc.) and so-called directives (claims about what one ought to do). She agrees that there must be some essential connection, but rejects the standard account—that offered by consequentialists—of what it might be. Thomson thinks that the content and status of directives requires explanation—we cannot rest content with a mere assertion that we are bound to do something, no matter how uncontroversial that something may be. She believes that such explanations can be provided by evaluative claims. On her account, however, directives are made true by facts about

relevant defects, rather than, as consequentialists have urged, facts about goodness.

The contents of this volume represent polished versions of papers originally given at the Second Annual Metaethics Workshop, held in Madison, Wisconsin, in September of 2005. I'd like to extend my appreciation to the following fine philosophers, who served on the program committee for that event, and so as de facto referees for this volume: David Brink, David Copp, Jamie Dreier, David Sobel, Nick Sturgeon, and Mark Timmons. I'd also like to offer my thanks to Simon Kirchin, and another OUP referee who prefers anonymity, for offering such helpful comments to each of the contributors on the penultimate versions of their articles. Finally, I'd like to record my gratitude to Peter Momtchiloff, philosophy editor at Oxford University Press, for his excellent stewardship of the series.

1

Wrongness and Reasons: A Re-examination

T. M. Scanlon

In the beginning, metaethics focused on morality and on motivation. When Hume famously observed that 'morals … have an influence on the actions and affections,' the 'influence' in question was a matter of motivation,[1] and much of metaethics has been focused on the problem of explaining this influence. This emphasis on the problem of 'moral motivation' has not been confined to neo-Humeans. Thomas Nagel's landmark book, *The Possibility of Altruism*, was largely devoted to an argument against the Humean position. But he nonetheless presented this argument as an inquiry into the motivational basis of prudence and altruism.

Today, for at least many of us, the subject has changed in two ways, First, the questions we are concerned with are not just about morality but also about practical reason more generally. Second, our concern is with reasons and rationality rather than with motivation. I have things to say about both of these shifts. But I will concentrate in this paper on what might be called the interaction between them: that is to say, on questions that emerge when the question about morality is shifted from a question about motivation to a question about reasons.

Put in terms of motivation, the influence of morals on action that needs to be explained might be put as follows:

MM: The fact that a person accepts the judgment that it would be wrong to do *X* can explain the fact that he does not do *X* (despite the advantages to him of doing it), and it would be odd for someone to accept this judgment yet feel no reluctance to do *X*.

I am grateful to participants in the conference for their comments, and especially to Derek Parfit for his extensive comments on several drafts.

[1] David Hume (1746: Book III, Part I, Section I).

I should add here a note on how I understand the semantics of 'wrong.' I take it that the most general meaning of 'wrong' is something like 'open to serious (decisive) objection.' Many things can be wrong in this sense, including answers to problems in arithmetic, chess moves, and career choices, and different objections are relevant in different cases. This does not mean that the word 'wrong' is ambiguous, but only that there are many ways in which something can be wrong in this univocal sense. When I first wrote about this subject, I was inclined to say that the phrase, 'morally wrong' identified one particular way in which an action can be wrong. It now seems to me, for reasons that I set out in more detail in *What We Owe to Each Other*, that a number of quite different kinds of objections can plausibly be called moral. There is thus a family of different ways in which things can be wrong, and I was offering an account of only one of these.[2] What I was concerned with in that book, and what I am concerned with here, is one particular way in which an action can be morally wrong, the way that involves wronging someone or, as I say there, violating 'what we owe to others.' When I talk about wrongness in the rest of this paper, it is this way of being wrong that I have in mind.

The problem of explaining MM is the problem of moral motivation. Putting the matter in terms of reasons, we might say instead:

MR: If it would be wrong for a person to do *X* in certain circumstances, then he or she has strong (normally conclusive) reason not to do so.

Several questions now arise: From the truth of MR it does not follow that the fact that it would be wrong to *X* itself constitutes a reason not to *X*. The conclusion that an act would be wrong might just be (or entail) that there are other reasons that count decisively against it. In this case wrongness might be what I have called a buck-passing notion, indicating the presence of other reason-providing considerations, rather than a reason-providing notion. So the first question is which of these is correct: is wrongness a reason-providing property or a buck-passing one? Second, if wrongness is a reason-providing property, how is this reason to be understood? Third, how is this reason related to the reasons provided directly by properties such as being harmful or dangerous, which can make an action wrong?

Several lines of reasoning seem to support the idea that wrongness is a buck-passing notion. The first is that a moral person who avoids a wrongful action usually avoids it because it is likely to harm someone, or because it would involve breaking a promise, or some similar specific reason. These specific reasons seem sufficient in themselves, and it may seem that they make 'it would be wrong' redundant as a reason-provider.

[2] Scanlon (1998: ch. 4).

But this is not so clear. Doing *A* would involve 'breaking a promise' in the sense that arguably constitutes a decisive reason against it only if there is no adequate justification for failing to fulfill this promise. So the conclusion that doing *A* would involve breaking a promise is in part a conclusion that there is no such justification: that is to say, the conclusion that it would be wrong to do *A*. The fact that *A* would be likely to harm or kill an innocent person may seem to be more independent of the idea of wrongness. But consider the following examples.

If I could easily prevent someone standing nearby from being injured, then I should do so. It would be wrong to just stand there and do nothing. The fact that this other person will be injured if I just stand there is a good reason not to do that. Any sensible moral view will tell me not to just stand there, but to offer help. And any such view will say that this is so *because* of the injury that would otherwise result: the person's need for help provides the obvious morally relevant reason for helping him. In this case there may seem to be no work for wrongness to do as a reason-providing property.

Now consider another case: I have been hired as a guard, by someone who has good reason to believe he is likely to be attacked. While standing guard, I see someone else about to be injured by a thug. I could run from my post and prevent this, but I would be leaving my client exposed to attack. So it might not be wrong for me to refuse to go to this person's aid, despite the fact that he will be injured if I do not.

In this case, the idea of wrongness seems to be doing more work. But this work is not primarily that of providing a new direct reason for a certain course of action. Rather, it lies in shaping the way I should think about the decision I face, and in determining which other considerations I should take to be reasons. The fact that I have undertaken to guard my client makes it the case that injury to the other person is no longer a conclusive reason for action in the way that it was in the previous example. This suggests that ideas of moral right and wrong were playing a role in that case too, but an unnoticed one of ratifying the status of the person's possible injury as a conclusive reason given the absence of other considerations that would have affected this status.

In my book, *What We Owe to Each Other,* I observed that this 'shaping' function is the way that the idea of wrongness most commonly influences action.[3] The idea that moral principles are imperatives which command actions is therefore somewhat misleading insofar as it suggests that these principles must be backed up by strong reasons (analogous to sanctions) for obeying them. In fact, when a moral person 'does the right thing' this is most often explained by the fact that she sees the considerations that might

[3] Scanlon (1998: 155–8).

tempt someone to act wrongly in such circumstances as not providing eligible reasons for action, rather than by the fact that she sees these reasons as outweighed by some powerful 'reason to be moral' that is triggered by the fact that the action would be wrong.

Although I recognized that this was so, I nonetheless continued, in *What We Owe to Each Other,* to regard wrongness as a reason-providing property. This was in part because I saw that, in addition to the 'shaping' function I have just described, wrongness plays what I called a 'backstop' role. I wrote:

> When one has reached the conclusion that a course of action would be wrong but is tempted to pursue it nonetheless, the considerations one finds tempting are ones that have been excluded or overridden at an earlier stage—that is, they have been ruled out as reasons insofar as one is going to govern oneself in a way that others could not reasonably object to. What one is asking in such a case is how much one should care about living up to this ideal, and this question presents itself in the form: How much weight should I give to the fact that doing this would be wrong?[4]

I recognized that this was not the only way in which the concept of wrongness affects our reasoning about what to do, or even the most important. But I concentrated on it because it seemed to me that focusing on cases in which wrongness played a clear reason-providing role (more specifically, focusing attention on the experience of feeling its reason-providing force) would shed light on the content of wrongness. I wanted to ask, 'When wrongness presents us with a reason for not acting a certain way, what kind of reason are we aware of? And what does wrongness have to be like to provide *that* reason?'

Reflection of this kind seemed to me, for example, to count against utilitarianism as an account of right and wrong. The value of happiness alone, I wrote, 'does not seem to account for the motivation we feel to do what is right and avoid what is wrong. When, for example, I first read Peter Singer's famous article on famine and felt the condemning force of his arguments, what I was moved by was not just the sense of how bad it was that people were starving in Bangladesh. What I felt, overwhelmingly, was the quite different sense that it was *wrong* for me not to aid them given how easily I could do so. It is the particular reason-giving force of this idea of moral wrongness that we need to account for.'[5]

The strategy of my argument was thus based on what might be called the *remorse test*: that is, the idea that an account of wrongness and its normative significance ought to fit with our sense of the kind of self-reproach that is occasioned by having done something wrong. In order to see this test as relevant one need not hold the view that it is part of the content of the

[4] Scanlon (1998: 157). [5] Scanlon (1998: 152).

judgment that an act is morally wrong that it would be appropriate for the agent to feel remorse.[6] I do not myself hold such a view. It is enough to support the remorse test if remorse involves a belief that what one has done is open to objection of the sort that makes it morally wrong. If this is so, then one can hope to gain insight into the nature of these objections by considering what remorse is like.

I will have more to say later about this test and about the relation between the reason-providing nature of wrongness and the remorse that is appropriate when we realize that we have acted wrongly. First I want to consider the relation between the backstop role of wrongness and the shaping role I have described. These two roles may seem very different. In backstop cases, wrongness is called upon to provide a reason, whereas in its 'shaping' role it may seem not to be doing this. But the two roles are more similar than might at first appear.

I said earlier that a moral person will generally not to need to appeal to a reason provided by the fact that an action would be wrong, because lower-order reasons, such as the fact that it would injure someone, or the fact that one promised not to do it, will strike such a person as primary and conclusive. That is to say, these reasons will seem conclusive *to a person who is thinking about what to do in the right way* (the way that morality requires.) A person who is thinking in this way will see these reasons as conclusive and other considerations (such as how advantageous to her it would be to break the promise or to cause the injury) as irrelevant. But one can ask, 'Why think about what to do in *that* way?' and this question needs an answer.

This question might seem not to need an answer if moral wrongness were identified with what we ought not to do in the all-encompassing sense of 'ought' which just expresses what is supported by the balance of all relevant reasons. It would make no sense to ask, 'Why decide what to do by considering what the balance of all the relevant reasons dictates?' But even on this view there would be questions to be asked. Given any particular claim about what the balance of relevant reasons dictates, one can ask 'Why think that *those* are the relevant reasons?' or 'Why think that the reasons balance out in *that* way?' Why, for example, should these reasons include the fact that one made a promise but exclude the fun of breaking it?

Moreover, it does not seem that the 'ought' of moral wrongness can be identified with this 'all-encompassing' ought. As the 'backstop' cases show, one can ask intelligibly why one should not, all things considered; do an

[6] As for example Mill's view that 'we do not call anything wrong, unless we mean to imply that a person ought to be punished in some way or other for doing it; if not by law, by the opinion of his fellow creatures; if not by opinion by the reproaches of his own conscience' (Ryan 1978: 321).

action that one admits would be wrong. The question one would be asking in such a case is, 'Why take the results of thinking about what to do in the way morality prescribes as authoritative and conclusive?' This is just the question that, I pointed out, needs to be answered in regard to the role of wrongness in 'shaping' our thinking about what to do. So the question, in response to which wrongness needs to provide reasons (or to invoke them) is same in the two cases.

In the most common kind of case in which wrongness plays a shaping role, two kinds of reasons are in play, corresponding to two kinds of 'why?' questions. There are first-order reasons such as 'it would hurt someone' or 'you promised,' which explain why a certain action would be wrong. In addition, there are higher-order reasons, which might be offered in response to the question, 'Why care about wrongness?' or 'Why accept *that* as the way to think about what to do?' The conclusion that an act would be wrong claims that considerations of the first of these kinds count decisively against it. So, considered in this role, wrongness is not itself a reason-providing property.[7]

Is it reason-providing in response to the higher-order 'why' question? That is to say: Is there a property shared by actions that are decisively ruled out by these first-order moral considerations that is *itself* reason-providing? And is this property properly called the property of moral wrongness. Note that the higher-order reason or reasons provided by this property need not be reasons for action. As my remarks about the 'shaping' role of moral wrongness indicate, they could instead be reasons for thinking about what to do in a certain way (a way that involves taking a certain view of which other considerations count as first-order reasons for action, and how these considerations are to be weighed).

One thing that seems clear is that the *concept* of moral wrongness is not tied to any particular answer to the question of why one should take conclusions about right and wrong to be authoritative guides to action. It is not clear exactly how this concept is best understood. It might be something like, 'open to serious criticism because it violates standards of conduct that everyone has good reason to regard as authoritative.' But even this may be too specific: someone might employ the concept of moral wrongness without referring to standards or principles of conduct. So perhaps wrongness should be understood more minimally as just the idea that something 'mustn't be done.' But however this concept is best understood, it is clear that it cannot involve a commitment to any specific higher-order reasons. People who use

[7] In chapter 2 of *What We Owe to Each Other*, I referred to goodness as a 'buck-passing' notion, because it provided not reasons on its own. Since the idea of wrongness plays a role in shaping and supporting these other reasons, however, it should perhaps be called a 'reason-referring' property.

the expression 'morally wrong' without linguistic oddity can disagree not only about which things 'mustn't be done' but also about why this is so.

When people call something morally wrong they must, I think, believe that one has serious reasons not to behave that way, but they may have no very clear idea what these reasons are. Alternatively, they may have one or another specific view about this. They may be utilitarians, or contractualists, or believe that the only thing that could possibly provide the kind of normative backing that moral standards need is the authority of a benevolent God. Since the concept of moral wrongness must allow for all of these views, an account that takes wrongness to be a reason-providing property and sets out to identify the reason that it provides cannot be an analysis of this concept.

In *What We Owe to Each Other*, I did not offer my version of contractualism as an account of the *concept* of wrongness or the meaning of the English expression 'morally wrong.' How, then, should I describe what I was doing in following the strategy set out in the remorse test? In one sense it is perfectly clear what my project was. To claim that an action is wrong is to claim that it violates standards that we have good reason to take very seriously. What I was doing was trying, in very general terms, to describe certain standards in a way that also identified what I claimed was a good reason for taking them seriously as ultimate guides to conduct (namely, the justifiability of our conduct to others). My thesis was that these standards and this reason provide the best way of understanding a large and central class of cases of moral wrongness.

This thesis was partly interpretative and partly reformist. I offered it as a way of making sense of what many of us believe when we say, in these cases, that an act is morally wrong, but also as an account of moral wrongness that people might endorse on reflection, even if they had previously accepted some other understanding of the standards underlying their use of 'morally wrong' and of the reasons supporting these standards. I was thus making a substantive normative claim about moral wrongness (about what standards of conduct we should take seriously). I acknowledged that when some people claim that an action is morally wrong they may have in mind standards other than the ones I was describing, which they take to be authoritative for reasons other than the one I described, and that in some cases these reasons may be worthy of respect.

This ground level description of my project seems to me entirely correct, and I stand by it (both as a description and as a project). Given that my contractualist formula was a substantive claim about wrongness, I might have described it as an account of what makes acts wrong.[8] I resisted this

[8] Derek Parfit suggested this to me at the time.

description for two reasons. First, that phrase seemed to me more properly used to describe first-order properties such as harmfulness, in virtue of which actions violate moral standards. Second, insofar as moral wrongness is taken to *provide* a reason for action, and my contractualism aims to explain what this reason is, this view seemed to be a thesis about what it is to be wrong, not just about what it takes to give a particular action that status.

So, drawing on an analogy with natural kind terms, I presented my contractualism as an account of the property of moral wrongness: as an account of the normative property that is shared by many of the actions we call morally wrong and explains their observed normative features, just as an account of gold aims to identify the physical property that is shared by observed instances of gold and explains their observed features. I pointed out immediately that this analogy is imperfect.[9] In the case of natural kinds, the property in question is unique (except in twin-earth type cases). But this need not be so in the case of wrongness. When different people call actions morally wrong some of them may have in mind different standards, and different reasons that they take to support them.

Given that this is so, I should have avoided describing my version of contractualism as an account of the *property* of moral wrongness. The error involved in doing so, it might be suggested, is similar to (although not the same as) the one that Moore called the naturalistic fallacy.[10] Moore's leading example was (a certain interpretation of) a utilitarian analysis of 'good.' But the point can also be put in terms of a utilitarian account of right and wrong. Bentham wrote:

> Of an action that is conformable to the principle of utility one may always say either that it is one that ought to be done, or at least that it is not one that ought not to be done. One may say also, that it is right that it should be done: that it is a right action; at least that it is not a wrong action. When thus interpreted, the words *ought*, and *right* and *wrong*, and others of that stamp, have a meaning: when otherwise, they have none.[11]

Bentham might be interpreted here as making a claim about the meaning of 'right', 'wrong', and 'ought'. So interpreted, he would be open to the objection that 'right' does not *mean* 'compatible with the promotion of the greatest happiness.' There is, however, a more charitable interpretation of what he may have had in mind. Putting things in the manner I have above, one could say that what he believed was that the general happiness was the only consideration capable of giving a standard of conduct the authoritative

[9] Scanlon (1998: 13). [10] Moore (1903: ch. 1).
[11] Bentham (1799), in (Ryan 1978: 67).

status invoked in the concept of wrongness. When he said that when 'right' and 'wrong' are used in some other way they lack meaning, what he was saying, in an overheated way, was that if they are understood to refer to standards backed by some consideration other than the greatest happiness then the judgments they express cannot have the authority that the words 'right' and 'wrong' normally convey. So these judgments are pretending to an authority that they do not have.

Similarly, my version of contractualism seemed to me, employing the remorse test, to be a plausible interpretation of what at least many of us have in mind when we think about right and wrong. And it also seemed, on reflection, to be normatively defensible—that is, capable of accounting for the priority and importance that our ideas of right and wrong claim for themselves. Taking a more moderate line than Bentham, I did not denounce all other accounts of the normative basis of right and wrong as meaningless or inadequate. But my claim was otherwise a claim of the same kind that he was making (on the more charitable reading I have suggested).

I was careful to say that I was not offering an analysis of the concept of wrongness, or the meaning of 'morally wrong.' So I was not vulnerable to an open question objection. But insofar as I claimed that I was giving an account of the property of wrongness, I was open to a related objection, which might be called the 'talking past each other' objection. If my version of contractualism is correct as an account of the property of wrongness, this has the odd consequence that when a teleological utilitarian or a divine command theorist says that an action is wrong, and a contractualist denies this, their disagreement does not consist in the fact that one side is affirming that the action has a certain property, and the other denying this. The property that one is claiming to apply is not the same as the one that the other is denying.[12]

It might be that the parties to such a disagreement are using the words 'morally wrong' to express different concepts. If this is so, then they are simply 'talking past one another' when one says 'This action is wrong' and the other says 'No, it is not.' But if they are using the words 'morally wrong' to express the same concept, such as 'must not be done' or 'violates standards we all have good reason to treat as authoritative' then there can still be a disagreement between them. For one thing, they may disagree about what standards we have most reason to take as ultimate standards for action. More fundamentally, they may have conflicting views about which reasons suffice to justify ultimate standards of conduct.

[12] Even if both parties affirm that the action is wrong, they will still be talking past one another in an important sense, since they will not be ascribing the same property to the action.

Applied to a case like the one just mentioned, this account describes the two sides as disagreeing about the applicability of the concept of wrongness to the action in question, but not in the properties they are claiming this action to have. They agree that the action is contrary to God's commandments, that it does not maximize happiness, and that it would be permitted by some principles that could not reasonably be rejected, but some say it is wrong, others that it is not.

Put this way, in terms of properties and concepts, this controversy has a distinctly academic character. But the underlying issue bears on the question of how we should understand the controversies about morality that are such a prominent part of our current political discourse. Morality is regularly invoked in political speeches, newspaper editorials, and letters to the editor. But it is sometimes unclear what the people who invoke it have in mind, and whether they are all talking about the same thing.

There is, of course, a range of cases on which everyone, or almost everyone, seems to agree: that it is wrong to kill children, for example, or even to kill one's business rival. (A few years ago I would have included torture on the list of things that are universally agreed to be wrong, but now I am not so sure.) There are also areas of first-order disagreement: over abortion (not surprisingly, since it is a difficult question), over assisted suicide and euthanasia, and, it seems, especially over homosexuality and other issues of sexual conduct.

The nature of this disagreement suggests, however, that the participants are not disagreeing only about how best to interpret a common set of standards. There are, of course, some sharp first-order disagreements, such as over the permissibility of abortion. But in addition to these first-order disagreements there are what appear to be extreme differences in emphasis. For some people, the main moral issues facing us are such things as the alleviation of suffering due to poverty, and the prevention of the harms that will be caused by global warming. Others seldom mention these things as *moral* issues. For them primary examples of moral issues are questions concerning sex, such as homosexuality, pre-marital sex, and even masturbation.[13] For many people in the first group, however, these are not *in themselves* moral issues at all, or if they are moral issues they are ones of lesser importance.

When people in these differing groups say that something is morally wrong, what are they claiming? I will take it that they are all using the

[13] This description of the matter is oversimplified in several respects. It overlooks the fact of manipulation by political leaders that leads people to focus on only some of the moral views they hold. It seems to me very unlikely that the people described as focusing on sexual morality do not also, *at some level*, share many of the general moral views held by their opponents. But I will not explore these matters here.

same concepts and thus that, they all mean, at a minimum, that these things 'mustn't be done,' or perhaps that they are forbidden by standards of conduct that we all have good reason to treat as authoritative. But the great divergence in the emphasis that people in these two groups place on different kinds of conduct suggests to me that the participants in these debates are not disagreeing over the best interpretation of a common set of standards. It seems, rather, that their ideas of morality involve different (albeit overlapping) standards which they hold to be authoritative. And at least in some cases it seems that they take these differing standards to be authoritative guides to conduct because they have different ideas about the reasons that could give any such standards the requisite authority.

It seems to me that these disagreements are well described in the terms I have been using to describe the sense in which proponents of different moral theories could be talking past one another. Participants in the debates I have just been describing *are* talking past one another in one important respect, although there is another way (or ways) in which they are making claims that genuinely conflict.

When they make claims about 'morality' they have some particular set of (vaguely described) standards in mind, and, perhaps some (even vaguer) idea of the reasons for taking these standards seriously. But their views of these matters are different: Those for whom sexual conduct is a preeminent moral issue are thinking about what might be called sin. Those for whom human rights, poverty, and global warming are paramount moral issues may be moved by something more like the justifiability of their actions and institutions to others. Perhaps others are moved by other ideals. When these people speak of 'morality' it is primarily these particular ideals or vaguely described sets of standards that they have in mind. Insofar as they are each employing the concept of moral wrongness, however, they are implicitly claiming that there are in fact good reasons to take these standards as authoritative guides to conduct. Perhaps it is more accurate to say that their claims *presuppose* that this is the case (since, not being philosophers, they do not make these commitments explicit). Insofar as their more specific claims about what is right and wrong have different, incompatible presuppositions, they are in a way talking past one another. But since these presuppositions make incompatible claims to authoritativeness, those who hold them are in genuine disagreement at the level of ultimate justification.[14]

[14] If these conflicting presuppositions are part of the meaning of 'morally wrong' as these people use that expression, then they are not making conflicting claims about the concepts that apply to the action in question. But even if they are talking past one another in this way, the disagreement I have just described would remain. They would still be disagreeing about what, ultimately, we have good reason to be guided by.

There is also a further way in which they may be disagreeing. In many cases people not only believe that the principles that they think of as the requirements of morality are well justified (by reasons of the kind that they may vaguely or not so vaguely have in mind). They may also believe that these standards and the reasons they take to support them are the best, and perhaps the only, way of making sense of what 'everyone' (or at least every morally serious person) intends when they speak of moral right and wrong. If so, then the people who hold these views are making what I called above conflicting interpretative claims about our ordinary morality.

Earlier, I described myself as making such a claim in what I called the 'ground level description' of my project in *What We Owe to Each Other*. My thesis, I said, was partly interpretive and partly reformist. I offered my version of contractualism as a way of making sense of what many of us intend to be claiming when we say that an act is morally wrong. But I was also making a substantive normative claim about moral wrongness (about what standards of conduct we have good reason to take seriously). I was thus offering an account of moral wrongness that people might endorse, on reflection, even if they had previously accepted some other understanding of the standards underlying their use of 'morally wrong' and of the reasons supporting these standards. But I acknowledged that when some people claim that an action is morally wrong they may have in mind standards other than the ones I was describing, which they take to be authoritative for reasons other than the one I described.

I stand by this description of the project, which seems to me to provide a good framework for understanding moral disagreement. Difficulties arose for it only when I claimed to be providing an account of the property of moral wrongness. This claim can be dropped from my account without affecting the other claims I make for contractualism. One possibility would be to accept a version of Parfit's proposal, and describe my thesis as an account of 'the single highest level property that makes actions wrong.' One of my objections to taking my contractualist formula as describing a property that 'makes acts wrong' would be met as long as it is understood that having this property makes an act wrong in a different way than, say, being harmful does.

It now seems to me, however (here referring back to my earlier remarks about the semantics of 'wrong'), that the best thing for me to say is that I am describing one *way* of being wrong. My contractualist formula describes a property (being allowed only by principles that could reasonably be rejected) that is shared by an important subclass of the actions that are morally wrong (that is, actions to which there are conclusive objections of the kind we call moral). This property is reason-providing: we have reason to care about whether our actions could be justifiable to others on grounds they could not

reasonably reject. But this reason is mainly a higher-order reason of the kind I described earlier. It is in the first instance a reason to think in a particular way about what to do and to accept as reasons the first-order considerations that this mode of thinking directs us to. Thus understood, my version of contractualism describes a property that is reason-providing, but not the property of being morally wrong (being something that 'mustn't be done'). Rather, it is one way in which actions can have that property.

To say that what my version of contractualism describes is only 'one way' of being morally wrong may sound rather weak and permissive. If I said instead that I was describing 'the single highest-level property that makes actions wrong' I would be making the more ambitious (perhaps more aggressive) claim that anyone who claims that some standards of conduct are worthy of the kind of status we give to moral standards for reasons other than those my version of contractualism describes is mistaken: there is no morality outside of contractualism. As I said above in discussing Bentham, my intention in my book was to take a softer line, and to allow for the possibility that some actions are wrong—open to serious criticism of the general kind we call moral—for non-contractualist reasons.

So I need to say something about how permissive I mean to be—about the kind of pluralism that I mean to leave open as a possibility. Pluralism of the kind I am now considering goes beyond the kind of interpretive claim that I have discussed above and allows for the possibility that there are multiple properties which provide reasons of a sort that makes them count as ways of being morally wrong. Two kinds of plurality should be distinguished. The first allows for the possibility that some conduct may be open to moral criticism on grounds other than the way it affects individuals—for example, because it fails to respect certain values, such as the value of natural objects, or of great human creations. Such ways of being morally wrong are quite distinct from the one that my version of contractualism describes. If something is wrong for one of these reasons it may also be unjustifiable to others, but if this is so this unjustifiability would be a mere consequence of an independent objection that, by itself, made the action wrong. (Some people, of course, believe that this is always true—that unjustifiability is always an unnecessary shuffle, which adds nothing. I of course do not think that this is so *in general*, but I agree that it is so in cases in which the objection to an action is rooted in some impersonal value, such as the value of nature.)

Whether the appeal to unjustifiability is otiose in a given case may be indicated by what I called above the remorse test. If something I did was wrong because I injured someone as a result of my failing to take the risk to her sufficiently into account in governing my actions, that injury and the reasons to avoid it are central to my self-reproach. But the character of my

remorse is also affected by the awareness of the unjustifiability *to her* of my lack of due care. I failed to give her interests the weight she could reasonably demand, and my relation with her is altered as a result. By contrast, if I have acted contrary to some impersonal value, my action may be unjustifiable to others, but this unjustifiability does not play a similar role in my remorse (unless I have also injured them by depriving them of the opportunity to experience something of value).

So one kind of pluralism is that which allows for the possibility of moral values other than what we owe to each other. I want to allow for pluralism of this kind, even though I maintain that we have reason to give the moral claims of what we owe to each other priority over these other values.[15]

A second form of pluralism would allow for the possibility of other ways of being morally wrong that are in more direct competition with the one that my version of contractualism describes. These ways of being wrong would be in more direct competition because they offer rival accounts of the standards governing our conduct toward one another. Consider, for example, a morality based on a code of honor. The content of such a code might not differ greatly from the morality we normally accept. It might, for example, forbid unprovoked violence and require the keeping of agreements. But even if it did not differ from contractualism in the duties it required (by, for example, requiring retaliation for injury) it would offer a very different basis for these duties. They would be based not on the value of justifiability to others but rather on a perfectionist ideal of the person: they would express the kind of self-discipline, strength, and dignity required to be a person of a certain kind, held to be valuable.

Some forms of religious morality would differ from contractualism in an analogous way. They might require concern for others, together with a kind of purity in one's personal life, not because these things are, at the most basic level, *owed* to others, but because they constitute the kind of life that God wants us to lead, and that love of God helps us to attain.

One question to ask about such moral views, and about versions of contractualism as well, is the interpretative question that I mentioned above: Which of them comes the closest to capturing the content and apparent basis of our ordinary moral thinking? Setting this question aside, however, the question is not which of these views 'gets it right' by describing the content of morality correctly. What we should ask instead are two other questions. The first is whether the values on which they are based are in fact worthy of respect—are they ones that we, or the people who hold them, have good reason to be guided by? Second, since such views provide very different bases for the standards governing our conduct toward one another,

[15] I defend it in the section on priority in chapter 4 of *What We Owe to Each Other*.

they put the relations between us on very different footings. They would do this even if the standards they support have much the same content: it is one thing to be able to rely on people because they are concerned about the justifiability of their actions to *us*, and quite another if their concern is at the most basic level not with justifiability to us but with their relation to God, or with an ideal of excellence. So one question about such views is how we should understand our relations with those who hold them. It might be suggested that an answer to the latter question (of interpersonal significance) provides an answer to the former (the question of ultimate justifiability): that we should reject at least some views of this kind because we have reason to object to its implications about our relations with each other. But to take this line in general about the question of justifiability would be to bias things in the direction of some form of contractualism, the identifying mark of which lies in the importance it places on justifiability to others as compared with other values. To avoid this bias, we need to treat the two questions as separate, at least in principle.

The relation between my version of contractualism, on the one hand, and various forms of utilitarianism or consequentialism, on the other, can be understood in a similar way. One route to moral views that are utilitarian, or consequentialist, in their *content* begins from a form of contractualism. Like my version of contractualism, this line of thought takes the idea of justifiability to others as basic, but it takes a broader view of the kind of justification that is relevant, dropping the individualist restriction to what I have called 'personal' reasons, and allowing appeal to aggregative interests and, in its consequentialist variant, impersonal values. Genuinely teleological versions of utilitarianism or consequentialism are quite different. They begin from the idea that what matters, ultimately, is not just our relations with one another, but the overall value of the states of affairs that we produce.

Morally serious individuals can hold moral views of any of these three kinds. When they make conflicting claims about right and wrong there is a sense in which they are 'talking past one other.' What they are really disagreeing about, however, is how standards of conduct can ultimately be justified: about the importance of our relations with each other, as compared with other values, and about the kind of relations (the kind of justifiability) that is most worth striving for.

REFERENCES

Bentham, Jeremy (1799) *Introduction to the Principles of Morals and Legislation*, in Ryan, 1978.

Hume, David (1746) *A Treatise of Human Nature*, ed. L. A. Selby-Bigge, rev. P. H. Nidditch (Oxford: Oxford University Press, 1978).

Mill, John Stuart (1861) *Utilitarianism*, in Ryan, 1978.
Moore, G. E. (1903) *Principia Ethica* (Cambridge: Cambridge University Press).
Nagel, Thomas (1970) *The Possibility of Altruism* (Oxford: Clarendon Press).
Ryan, Alan (ed.) (1978) *Utilitarianism and Other Essays* (London: Penguin).
Scanlon, T. M. (1998) *What We Owe to Each Other* (Cambridge, Mass.: Belknap Press).

2

An Outline of an Argument for Robust Metanormative Realism

David Enoch

1. INTRODUCTION

Robust Metanormative Realism is the view, somewhat roughly, that there are response-independent, non-natural, irreducibly normative truths, perfectly universal and objective ones, that when successful in our normative inquiries we discover rather than create or construct.[1] Normative truths include—but are not limited to—the truths of morality, so Robust Metanormative Realism is the natural generalization of Robust Metaethical Realism.

Robust Realism—in either its metaethical or more general metanormative form—is out of philosophical fashion today,[2] and is often criticized, but more often ridiculed or ignored, by supporters of such -isms as Non-cognitivism, Expressivism and Quasi-Realism, Ethical Naturalism (either in its old-fashioned, a priori, version, or in its more recent, a posteriori, not-reductive-in-a-strict-sense-of-this-term, version), Dispositionalism,

For comments and discussions on this paper and the larger project on which it is based, I want to thank Stephanie Beardsman, Hagit Benbaji, Thérèse Björkholm, Paul Boghossian, Terence Cuneo, Stephen Darwall, Cian Dorr, Hartry Field, Ernesto Garcia, Pete Graham, Alon Harel, Ulrike Heuer, David Heyd, Peter Kung, Andrei Marmor, Tom Nagel, Derek Parfit, John Richardson, Josh Schechter, Mark Schroeder, Russ Shafer-Landau, Nishi Shah, Assaf Sharon, Brad Skow, Sigrún Svavarsdóttir, Kevin Toh, Pekka Väyrynen, Crispin Wright, Masahiro Yamada, and two anonymous readers for *Oxford Studies in Metaethics*.

[1] Unlike Oddie (2005), in my mouth 'Robust Realism' does not include a commitment to the normative truths and facts being a part of the causal network.

[2] Though perhaps not as much so as some ten or twenty years ago. See Bloomfield (2001), Stratton-Lake (2002), Shafer-Landau (2003), Oddie (2005), and Dancy (2005). And note also the change in tone with regard to such a view from Gibbard (1990) to Gibbard (2003).

Constructivism, Relativism, Subjectivism (sensible or otherwise), and Error Theories of different shapes and forms (this list of the non-robust-realist metanormative options is neither exhaustive nor exclusive). In this paper I embark on the project of defending Robust Metanormative Realism.

A full defense of Robust Realism would consist of two main parts: A positive argument with Robust Realism as its conclusion, and detailed replies to common objections to Robust Realism. Because Robust Realists—perhaps like realists more generally—typically put more effort into the latter,[3] I will here focus on the former. I want, then, to offer a positive argument for Robust Realism, an argument from the deliberative indispensability of irreducibly normative truths.

My argument in support of Robust Realism is modeled after arguments from explanatory indispensability common in the philosophy of science and the philosophy of mathematics. I argue that irreducibly normative truths, though not explanatorily indispensable, are nevertheless deliberatively indispensable—they are, in other words, indispensable for the project of deliberating and deciding what to do—and that this kind of indispensability is just as respectable as the more familiar explanatory kind. Deliberative indispensability, I argue, justifies belief in normative facts, just like the explanatory indispensability of theoretical entities like electrons justifies belief in electrons.

My discussion starts with an antirealist challenge—the one I call Harman's Challenge—that claims that moral truths or facts are explanatorily redundant, and that we therefore have no reason to believe they exist. Having presented the challenge (in section 2), I proceed to reject the explanatory requirement on which it is based. I then show—in section 3—how doing so is compatible with a rather strict requirement of ontological parsimony. In section 4, I argue that indispensability—the kind of indispensability that purportedly justifies ontological commitment—need not be explanatory, and that deliberative indispensability may be just as respectable as explanatory indispensability. In section 5, I say more about what indispensability is, dividing the discussion into an account of what I call instrumental and intrinsic indispensability. In the following two sections—6 and 7—I characterize the phenomenology of deliberation, arguing that it satisfies the desiderata needed for my argument for Robust Realism to go through. In section 8, I briefly address a general epistemological worry about the

[3] See Korsgaard's (1996: 31) accusation (referring to Clarke and Price). For realists' admission of guilt, see Nagel (1986: 143–4) and Parfit (2006: section 2). Much of Shafer-Landau's recent defense of moral realism (2003) is also of this nature. And I too embark on the project of replying to common objections elsewhere. See Enoch (unpublished manuscript) and Enoch (work in progress).

move from indispensability to belief, and in section 9, I hint at why it is unlikely that any other metanormative view can supply all that is needed for deliberation. Many of the argumentative moves here call for further elaboration, which I supply elsewhere.[4] For this reason, the discussion here at most establishes a fairly tentative conclusion, and I state it in section 10. Despite this incompleteness, though, by the time I reach the tentative conclusion enough will have been said to make the outline of the argument clear, and also—so I hope—to emphasize its strengths and to frame further discussion.

The argument here developed aims not just at soundness but also at sincerity in the following sense: It is supposed to be an argument for Robust Realism that is sensitive to why it matters whether Robust Realism is true. Often when reading philosophy one gets the feeling that the writer cares more deeply about his or her conclusion than about the argument, so that if the argument can be shown to fail, the philosopher whose argument it is will simply proceed to look for other arguments rather than take back his or her commitment to the conclusion. And there need be nothing wrong with arguing in this way. But it nevertheless seems to me that there is something to be said for an argument in which the underlying concerns are put in clear view. And the argument I develop here is, if I am successful, of this kind. If it can be rejected—if, in other words, normative truths robustly-realistically understood are not after all indispensable for deliberation[5]—then I no longer care whether Robust Realism is true, and am then happy to reject my argument's conclusion rather than look for other arguments that can better support it.

2. HARMAN'S CHALLENGE

Seeing a vapor trail in a cloud chamber, a physicist thinks to herself: 'There goes a proton'. That she makes the observation that she does is at least some evidence for there having been a proton in the cloud chamber, Harman argues plausibly, because the best explanation of her observation involves the fact that there really was a proton in the cloud chamber at the relevant time. If the physicist's observation is best explained by an alternative explanation, one that does not involve the proton in an appropriate way, her observation gives no reason to believe that there was a proton in the cloud chamber.

[4] Enoch (2003), and a book in preparation.

[5] Actually, there is another job I need such normative truths to perform, so that the conditional in the text is not the whole story. But this other job—a political one—is closely related to the deliberative one, and will in any case not be my topic here.

Seeing a few children set a cat on fire you think to yourself 'That's wrong'. How is the fact that you immediately make this judgment best explained? Harman argues that it is best explained by psychological, sociological, historical, cultural, and other such facts about you. Whether or not what the children are doing really *is* wrong is not at all relevant, Harman says, for the best explanation of your immediate moral judgment. Seeing that the relevant observation or judgment is best explained without assuming the existence of the relevant purported (irreducibly) moral fact, Harman concludes, we have no reason to believe there are such (irreducibly) moral facts. Realism refuted.[6]

The general thought seems clear enough: Moral facts do not play an appropriate explanatory role (the No Explanatory Role Thesis), and, given that playing such a role is necessary for justified belief in the existence of a kind of fact (the Explanatory Requirement), we are not justified in believing in moral facts. But despite the simplicity of the thought this argument attempts to capture, much work needs to be done if we are to have here a reasonably precise argument against Robust Metanormative Realism. Which possible explananda, for instance, count in shouldering the burden of the explanatory requirement? Only observations, as Harman himself seems to suggest? Why this restriction? Maybe explaining non-observational beliefs, or desires, or actions, or non-action sociological events, or more purely causal events suffices for satisfying the explanatory requirement or the intuitive condition it is meant to capture?[7] Do *moral* facts count as respectable explananda, such that if moral facts are required in order to explain other moral facts, moral realism is vindicated? This seems like cheating, but can moral facts be declared less than respectable explananda without begging the question against the realist?[8] What assumptions about the individuation of kinds of fact is it reasonable to read into the explanatory requirement? What kind of explanatory role must be played by a kind of fact in order to satisfy the explanatory requirement? And if the argument is generalized from the metaethical to the metanormative,[9] doesn't it flirt with self-defeat, given that the explanatory requirement itself is normative

[6] For his original statement of the problem, see Harman (1977: ch. 1). Harman does not claim originality for the general problem his text tries to capture.

[7] For discussion of these issues, see Sturgeon (1984: 54–5); Lycan (1986: 89), Railton (1986: 192); Brink (1989: 186–9); Sayre-McCord (1992), Wright (1992: 197–8), Shafer-Landau (2003: 102–3).

[8] For some discussion, see Nagel (1986: 146), Brink (1989: 182–3), and Shafer-Landau (2003: 104).

[9] A generalization along these lines has been suggested by Sayre-McCord (1988: 278; 1992: 70, footnote 21). It is also clear that normativity—and not just morality—is at stake in Nagel's (1986: chapter 8) and Dworkin's (1996) relevant discussions.

through and through?[10] These are some of the questions in need of answers if Harman's Challenge is to become a serious threat to Robust Realism.[11]

As I am not here specifically interested in Harman's Challenge but am rather using it as a way of introducing my argument for Robust Realism, for my purposes no such detailed discussion is needed. For regardless of the details, an intuitive challenge remains: We have, it seems, good reason to believe in electrons, and perhaps also in numbers, because they play an appropriate role in the best explanation of a respectable explanandum. Can belief in normative facts—or, say, in values—be justified in a similar way? If not, what reason *do* we have for believing in them? Shouldn't we avoid multiplying kinds of entities, facts and truths without sufficient reason?

Broadly speaking, two realist response-strategies suggest themselves.[12] The realist can, first, reject the No Explanatory Role Thesis, and argue—usually, by citing examples—that normative facts indeed do play an appropriate role in the best explanation of a respectable explanandum. Or, second, the realist can reject the Explanatory Requirement, arguing that we have reason to believe in normative truths even though (or even if) they do not play such an explanatory role.

The former strategy has been far more common in the literature.[13] Though one can find in the literature some hints and brief comments suggesting the second strategy,[14] it has not, to the best of my knowledge,

[10] For similar points, see Quinn (1986: 539), Simon (1990: 113 (footnote 27)), Putnam (1995: 71), McGinn (1997: 13–4), and Shafer-Landau (2003: 113).

[11] This is not necessarily a criticism of Harman. First, it's not clear he is interested in the metanormative generalization of his argument. And second, perhaps even without answering some of the questions in the text, Harman's Challenge poses a serious threat to some other kind of realism, like the Cornell Realist's. Indeed, much of the discussion of Harman's Challenge was conducted by such naturalist realists as Sturgeon and Brink.

[12] Zimmerman (1984: 81–2) and Leiter (2001: 88) draw a similar distinction between two strategies of coping with Harman's Challenge.

[13] And an extensive literature it is. For some of it, see: Audi (1997: chapter 5); Blackburn (1991a; 1991b); Brink (1989: 182–97); Copp (1990); Harman (1977; 1984; 1986; 1998); Harman and Thomson (1996: ch. 6, 9, and 10); Leiter (2001); Lycan (1986); McDowell (1985: 117–20); Moore (1992); Quinn (1986); Railton (1998); Sayre-McCord (1988; 1992); Shafer-Landau (2003: 98–115); Sturgeon (1984; 1986; 1991; 1992; 1998); Wiggins (1990); Wright (1992: ch. 5); Yasenchuk (1994); Zimmerman (1984).

[14] Remarks that are somewhat suggestive of the second strategy, either in rejecting the explanatory requirement, or in interpreting it liberally enough, can be found in Lycan (1986: 89), Wiggins (1990: 85), McDowell (1985: 118–19), Nagel (1986: 144–5), Dworkin (1996: 119–22), Platts (1980: 79), Korsgaard (1996: 96), and Shafer-Landau (2003: 114–15). A clearer statement of the second strategy and an initial attempt at

been pursued systematically. In a moment I will proceed to pursue the second strategy and so I can afford to remain largely neutral—at this point in my argument, at least—regarding the prospects of the first strategy, about which I am rather pessimistic. Very briefly, then, and without pretending that the following comment is a serious argument: My pessimism regarding the first strategy of vindicating Robust Realism comes not only from the implausibility, as it seems to me, of the claim that normative facts play an appropriate role in the best explanation of relevantly respectable explananda, but also from two further points: First, I suspect that even if normative facts do play such a role, the first strategy of coping with Harman's Challenge could at most vindicate a naturalist kind of realism, not the Robust Realism I am out to defend (this, of course, is not a reason to think the first strategy must fail, only that it can't get me all that I want). And second, and more importantly, it seems to me that even if normative facts do not play such an explanatory role, still this doesn't compromise their respectability in any way. If this is so, a serious attempt at the second strategy is certainly called for.

Let me put the first strategy to one side, then, and focus on the second.

3. PARSIMONY

A worry immediately threatens: What underlies the explanatory requirement is, after all, a highly plausible methodological principle of parsimony: Kinds of entities should not be unnecessarily multiplied, redundancy should be avoided.[15] And, it seems, without such a principle it is exceedingly hard—perhaps even impossible—to justify many of our negative existential beliefs. Taking this methodological principle as given, then, how

pursuing it can be found in Simon (1990: 105–6) and Sayre-McCord (1988: 278–80). An emphasis on the point of view of the deliberating agent—central to my employment of the second strategy of coping with Harman's Challenge—can be found in Regan (2003) and Rosati (2003). At times, Regan's claims are very close to my own, except he thinks such line of thought only defends realism 'for practical purposes' (2003: 656). I am not sure I understand this phrase, and to the extent that I do, I want more. And for an emphasis on the practical, deliberative relevance of the realist truth of normative judgments, see Fitzpatrick (2005: 685–6).

[15] I am perfectly happy talking here about ontological profligacy and parsimony, ontological commitment, entities, and facts. But if for some reason you find talk of normative *truths* rather than facts, or normative *properties* rather than objects much less threatening, feel free to paraphrase accordingly. Nothing much will then have to be changed. Notice, for instance, that the appeal of the parsimony requirement survives such a paraphrase.

can the explanatory requirement be consistently rejected? Assuming that (irreducibly) normative facts play no appropriate explanatory role, are they not redundant, and then isn't belief in them unwarranted?

In order to allay this worry, it is necessary to distinguish two different requirements of parsimony. First, there is the most general requirement not to multiply ontological commitments without sufficient reason. This requirement places a prima facie burden of argument on the party arguing for a belief in the existence of entities or facts of a certain disputed kind. Call this *the minimal parsimony requirement*.

Often, though, more is packed into the methodological principle of parsimony than the minimal parsimony requirement. It is often assumed that the only way of satisfying the minimal parsimony requirement is by showing that the relevant kind of fact is *explanatorily* useful.[16] With this assumption, the minimal parsimony requirement becomes the explanatory requirement.

I want, then, to reject the explanatory requirement while adhering to the minimal parsimony requirement. And, as is by now clear, the way to do this is to reject the assumption that the minimal parsimony requirement can only be satisfied by explanatorily indispensable facts, truths, properties, and entities.[17] In other words, I suggest we restrict, in accordance with the minimal parsimony requirement, our ontological commitment to just those things that are indispensable. But I suggest that we consider other—non-explanatory—kinds of indispensability as satisfying this requirement. So the line I'm about to take does not have the unacceptable counterintuitive result of admitting objectionable—and completely, not just explanatorily, redundant—things into one's ontology.

[16] In a somewhat different context (that of characterizing the realist-antirealist debate, not that of deciding it), Wright (1993: 73) notices this often-made assumption (explicitly referring to Harman), and expresses his doubts about it.

[17] Slors (1998: 243) makes a similar point about the mental, when he writes: 'But why shouldn't mental regularities have some other function than a causal-explanatory one? It might just be possible that the mental justifies its place in our ontology by other means than its causal efficacy.' And here is Grice (1975: 31): 'My taste is for keeping open house for all sorts and conditions of entities, just so long as when they come in they help with the house-work. Provided that I can see them at work, and provided that they are not detected in illicit logical behaviour … I do not find them queer or mysterious at all. To fangle a new ontological Marxism, *they work therefore they exist*, even though only some, perhaps those who come on the recommendation of some form of transcendental argument, may qualify for the specially flavoured status of *entia realissima*. To exclude honest working entities seems to me like metaphysical snobbery, a reluctance to be seen in the company of any but the best objects.' Honest working entities are, of course, those that satisfy the minimal parsimony requirement. And I would add only that explanatory work is not the only kind of work around the house that needs doing.

4. INDISPENSABILITY,[18] EXPLANATORY AND OTHERWISE

Why should we believe in, say, electrons? One common answer runs like this: There are many inferences to the best explanation the conclusion of which entails the existence of electrons: our best scientific theories quantify over electrons; we ought to believe that these theories are at least approximately true (they are, after all, our *best* theories, our best explanations of numerous phenomena; and they are also, it seems, fairly good), and so we ought to believe that electrons exist. If electrons play an appropriate role in the best (and good enough) explanation of respectable explananda—and it seems they do—we're justified in believing that electrons exist. Of course, Inference to the Best Explanation (IBE, for short) is not uncontroversial. For now, though, let us assume that IBE suffices to justify ontological commitment.

As I understand inferences to the best explanation, they are really particular instances of indispensability arguments.[19] Electrons are indispensable for our best explanations; so, by IBE, electrons exist. And it is important to note here, that instances of IBE are arguments from *explanatory* indispensability. Electrons are indispensable for our explanatory project, and for this reason we are justified in believing they exist.

As has already been argued, the availability of the second strategy of coping with Harman's Challenge depends on there being other, non-explanatory, kinds of indispensability that suffice to justify ontological commitment.[20] Later on I will suggest one such other kind, deliberative indispensability. For the moment, though, I want to make the following preliminary point: Given some other purportedly respectable kind of indispensability, the proponent of the explanatory requirement (who is also a proponent of IBE)

[18] A terminological apology: My use of the word 'indispensability' is without a doubt a stretch of ordinary usage. Seeing, however, that my use of this term is not completely discontinuous with ordinary usage, that I explicitly explain my way of using it, and that my way of using it echoes the way it is already used in the context of indispensability arguments in the philosophy of mathematics, I hope this stretch is not too misleading.

[19] This relation between IBE and indispensability arguments has been noticed by Field (1989: 14) and Colyvan (2001: 7–8, especially fn. 17). Interestingly, Harman (1977: 10) mentions indispensability arguments for mathematical realism as support for his *disanalogy* between mathematical and ethical facts. If I am right in what follows, these arguments in fact supply the material for an important analogy between the two.

[20] Field (1989: 14) and Colyvan (2001: 6) have noticed that there may be other kinds of indispensability that can ground prima facie respectable indispensability arguments. Resnik (1995; 1997: ch. 3) puts forward what seems to be an argument from a different (pragmatic) kind of indispensability, but his is still indispensability to the scientific project, broadly understood.

must find a non-arbitrary way of distinguishing between explanatory and that other kind of indispensability. She must show, in other words, why it is that explanatory indispensability ought to be taken seriously, but other kinds of indispensability ought not to be so taken; she must present a reason for taking explanatory indispensability to justify ontological commitment that does not generalize to other kinds of indispensability.[21] Now, my way of justifying the move from indispensability to belief will not be of that sort. It will apply to explanatory indispensability just in case it applies to other, non-explanatory, kinds of indispensability, and in particular to deliberative indispensability. This does not show, of course, that *no* rationale can be given for restricting respectable status to explanatory indispensability alone. So think of my point here as a challenge: Can you think of any reason for grounding ontological commitment in explanatory indispensability that is not really more general, a reason for grounding ontological commitment in indispensabilities of other kinds as well?

If there is no reason for taking explanatory indispensability seriously that is not a reason for taking some other kinds of indispensability seriously, then the move from the minimal parsimony requirement to the explanatory requirement is arbitrary and so unjustified.[22] If any other kind of indispensability can be defended, then the second strategy of coping with Harman's Challenge becomes promising: All that is then left to do is to show that (irreducibly) normative truths are indispensable in this other, non-explanatory, way.

5. INDISPENSABILITY: SOME DETAILS

Before doing that, though, more needs to be said about indispensability. As has been noted in the philosophy-of-mathematics literature,[23]

[21] Thus, I think Simon (1990: 105–6) accurately characterizes the dialectical situation when she writes: 'What one would like from the anti-realist is an argument for using explanatory necessity as a criterion of reality which is more compelling than the absence of a better one. On the other hand, what one would like from the realist is, if not an alternative criterion, at least some indication of how one is to go about evaluating claims concerning the reality of different purported existents.' My argument can be seen as an attempt to give Simon what she wants from the realist.

Later on, Simon writes (1990: 108): 'And, one might ask, is not the necessity of saving morality as compelling as explanatory necessity? Perhaps it is a necessity which itself warrants multiplying entities.' It is not entirely clear to me what Simon has in mind, but she may very well be anticipating here an argument not unlike my argument from deliberative indispensability.

[22] I suspect this is what McGinn has in mind when he accuses Harman's explanatory requirement of being arbitrary and dogmatically empiricist (1997: 13; see also at 17, 36). For a similar point, see Putnam (1995: 70).

[23] See, e.g. Field (1989: 14), Colyvan (2001: 6).

where discussions of indispensability are typically located, indispensability is always indispensability *for* or *to* a certain purpose or project. Quantifying over numbers and sets is arguably indispensable *for* doing physics. Quantifying over (possibly other) abstracta is arguably indispensable *for* doing metalogic.[24] Of course, one thing may be indispensable for one purpose or project but not for another.

Once this is noticed, it becomes clear that in order fully to understand what (the relevant kind of) indispensability comes to two distinct questions must be answered. First, it must be determined what it takes, given a purpose or a project, for something to be indispensable for it. As I will put things, the first thing that is needed is an account of *instrumental indispensability*. Second, it must be determined which purposes or projects are such that indispensability for them suffices to ground ontological commitment. That is, an account of what I will call *intrinsic indispensability* is likewise needed.

5.1 Instrumental Indispensability

Given a purpose (such as explaining) or a project (such as the scientific project), what does it take for something to be indispensable for it, in the sense relevant for ontological commitment?

Of course, being helpful is not enough. If, for instance, mathematical objects are only used in scientific theories as a means of simplifying inferences which could be drawn without numbers as well, then, it seems, mathematical objects are not indispensable for the scientific project in the relevant sense. What is needed here is something like Field's (1989: 59) distinction between being useful in, e.g., facilitating inferences on the one hand, and, on the other hand, being useful in being theoretically indispensable.[25] However exactly the latter is to be understood, it seems intuitively clear that the former cannot justify ontological commitment, even assuming that the relevant project is intrinsically indispensable; it is perfectly compatible, for instance, with a fictionalist attitude towards mathematics and a nominalism about abstract objects. Mere usefulness does not suffice for instrumental indispensability.

Nor does what I will call (merely) *enabling* indispensability. Presumably, we cannot successfully engage in the scientific project without sufficient sleep. But sleep is not indispensable to the scientific project in the sense that suffices for the justification of ontological commitment. Of course, if

[24] See Field (1991: 1).

[25] Brink (1989: 192) makes a similar distinction in the metaethical context between pragmatic and in-principle indispensability. For reasons that will emerge in what follows, I think Brink's terms are potentially misleading.

we cannot successfully engage in the scientific project without sufficient sleep, then that we have in fact so engaged in the scientific project is evidence that we did get sufficient sleep. But our engaging in the scientific project—though evidence for sufficient sleep—does not *commit* us to any claims about us having had sufficient sleep. The account of instrumental indispensability I am after should have this result. So enabling indispensability is not what I am after.

An initially attractive recourse is to restrict instrumental indispensability—indispensability for a *theory*, for now—to just those things that are ineliminable from the theory. However, as Colyvan (2001: 76–7) argues, this too will not do, for the following two reasons. First, it is not entirely clear what ineliminability is. Surely, just noting that once the disputed entities are eliminated the theory that is left is different from the one we started with is not sufficient for ineliminability, for this requirement is satisfied by all entities a theory invokes, talks of, or quantifies over. Second, it may very well be the case that no entity is strictly ineliminable for any theory, because the theory can be reformulated and reaxiomatized such that any given entity is eliminated.[26] Ineliminability as a criterion for instrumental indispensability thus also fails.

I want to follow Colyvan in offering the following criterion for instrumental indispensability. If a scientific theory T_1 quantifies over, say, electrons, and T_2 is the theory we get after eliminating all references to electrons from T_1, *and if T_2 is all-things-considered at least as attractive as T_1* (or is, at least, sufficiently attractive), then it seems clear that electrons are not instrumentally indispensable for our scientific project.[27] The relevant criteria of attractiveness are, of course, explanatory. An entity is explanatorily indispensable just in case it cannot be eliminated from our explanations without loss of explanatory attractiveness. Colyvan's condition is intuitively appealing, and may be considered simply a result of the policy of inferring only to the *best* explanation.

For my purposes, though, Colyvan's condition is not good enough as it stands, for I am interested in more than just *explanatory* indispensability, and in more than just indispensability *to a theory*. Luckily, though, Colyvan's

[26] Colyvan (2001: 77).

[27] This is a reformulation of Colyvan's (2001: 77) criterion. The term 'instrumental indispensability,' as well as the (explicit) distinction between instrumental and intrinsic indispensability are mine. Field nowhere puts an explicit definition or characterization of what it takes for an entity to be indispensable to a theory, but at times he says things that suggest that he too acknowledges something like Colyvan's condition. Colyvan (2001: 76, n. 16), for instance, quotes the following sentence from Field (1980: 8): 'we can give *attractive* reformulations of [the theories of modern physics] in which mathematical entities play no role' (Colyvan's emphasis). In the metaethical context, Wiggins (1990: 84) hints at such a condition.

condition—and its appeal—generalize nicely. Something is instrumentally indispensable for a project, I suggest, just in case it cannot be eliminated without undermining (or at least sufficiently diminishing) whatever reason we had to engage in that project in the first place; without, in other words, thereby defeating whatever reason we had to find that project attractive. The intuition underlying this criterion for instrumental indispensability is simple: The project itself is (intrinsically) indispensable for a reason, and if the only way to engage in it in a way that doesn't defeat that reason involves a commitment to an entity (or a fact, or a belief, or whatever), then the respectability of the project confers respectability on that commitment. Colyvan's condition is a particular instance of this condition, with the relevant project being the scientific one, and the relevant criteria of attractiveness being explanatory.

On this account, then, what is in the first instance indispensable to the scientific—and more generally explanatory—project are not electrons and numbers but, somewhat roughly, the validity of IBE, or indeed the belief in the by-and-large explanation-friendliness of the universe, so that our best explanations are likely to be true (I return to relevant doubts below). It is the belief that much of what is going on can in principle be explained, can be made sense of, and so that IBE is at least a reasonably good rule of inference, that is directly indispensable to the explanatory project. The commitment to electrons and numbers is both derivative (from the more general belief together with the specific scientific findings and theories) and tentative (for better explanations may be found in the future). We would lose whatever reason we had to engage in the explanatory project not if we ceased to believe in electrons, but rather if we ceased to believe that there is *some* explanation of many of the phenomena we try to explain.

5.2 Intrinsic Indispensability

So much, then, for instrumental indispensability. But that something is (instrumentally) indispensable for a project cannot justifiably ground ontological commitment without some restriction on the set of acceptable projects. Believing in evil spirits, for instance, may be indispensable for the project of sorcery, but this is no reason to believe in evil spirits (if anything, it is a reason not to engage in sorcery). And God may be indispensable for the project of achieving eternal bliss, but this does not give reason to believe in God—unless, that is, the project of achieving eternal bliss is of the kind that can justify ontological commitment; unless, in other words, it is an intrinsically indispensable project.

It has been noted in the philosophy-of-mathematics literature that some restriction on the set of admissible purposes is needed. Nevertheless, to the

best of my knowledge no criterion for intrinsic indispensability has been suggested. Colyvan (2001: 7), for instance, asks the right question, but fails to answer it:

> Which purposes *are* the right sort for cogent [indispensability] arguments? I know of no easy answer to this question.

Nor does he suggest an answer to this question that is not easy. Now, in discussions of the Quine–Putnam indispensability argument for Platonism regarding mathematical objects, neglecting to offer a criterion for intrinsic indispensability is not a serious dialectical flaw: As is often noted,[28] the argument is put forward by the mathematical Platonist in an attempt to convince scientific realists. And with these as the major interlocutors, both parties to the debate are happy to assume that, whatever the criterion for intrinsic indispensability, at least the scientific project satisfies it, at least the scientific project is respectable enough to justify ontological commitment. (Indeed, when both parties are also metaphysical naturalists, both are happy to assume also that the scientific project is the *only* one that is intrinsically indispensable.) The parties are typically so comfortable with such an assumption that it remains implicit.[29]

In our context, though, more needs to be done. I agree that the explanatory project is intrinsically indispensable. But I am not willing to grant that it is the *only* intrinsically indispensable project. And in order to establish the claim that our deliberative project is also intrinsically indispensable, it is necessary to answer the question Colyvan leaves unanswered. Which projects, then, are intrinsically indispensable?

Think of the explanatory project again. What is it that makes it—as we assume, for now—intrinsically indispensable? Why is it that if it is indispensable for our explanatory project that p we are justified in believing that p? What distinguishes the explanatory project from, say, sorcery, such that indispensability to science, but not to sorcery, justifies ontological commitment? Answering these questions satisfactorily requires more detail than I can supply here. Let me, then, put forward my answer in a preliminary and somewhat dogmatic way.

The explanatory project is intrinsically indispensable because it is one we cannot—and certainly ought not—fail to engage in, it is unavoidable for us; we are essentially explanatory creatures. Of course, we can easily refrain from explaining one thing or another, and it's not as if all of us have to be amateur scientists. But we cannot stop explaining altogether, we cannot

[28] See Colyvan (2001: e.g. 25).

[29] Colyvan (2001: 7) is a welcome exception, in that he explicitly notes this assumption.

stop trying to make sense—*some* sense—of what is going on around us. In an important sense, the explanatory project is not one that, like sorcery, is optional for me: I have no option of stopping (or not starting) to engage in it. If God (or believing in her, or both) is indispensable for the project of achieving eternal bliss, the rational thing to do seems to be either to believe in her or to abandon the project of achieving eternal bliss. But with non-optional projects like the explanatory one, there is no real option of abandoning them. If something is indispensable for such a project, it seems belief is the only rational way to go. And this line applies to all and only essentially unavoidable projects.

This is, then, my (largely unargued-for) suggestion for a criterion of intrinsic indispensability: A project is intrinsically indispensable if (and only if, quite plausibly; but my argument doesn't rely on the following condition being also necessary) it is non-optional in the relevant sense. Instances of IBE are justified, then, because they are arguments from indispensability to the explanatory project, which is essentially unavoidable.[30]

6. DELIBERATION AND INTRINSIC INDISPENSABILITY

But if that is right, it seems clear that our deliberative project is likewise intrinsically indispensable. For we are also essentially deliberative creatures. We cannot and should not avoid asking ourselves what to do, what to believe, how to reason, what to care about. We can, of course, stop deliberating about one thing or another, and it's not as if all of us have to be practical philosophers (well, if you're reading this paper, you probably are, but you know what I mean). But we cannot stop deliberating altogether. The deliberative project is not one we can opt out of, it is not optional for us.

If I am right, then, about what makes projects intrinsically indispensable, the deliberative project is one such project. But I acknowledge that much more needs to be said in support of this criterion for respectable projects.[31] Notice, then, that even if I am wrong, if you want to exclude deliberative indispensability as not-quite-as-respectable as explanatory indispensability,

[30] Here and below I remain undecided on whether intrinsically indispensable projects are those we *can*not disengage, or rather those we *should* not disengage, or perhaps some combination of the two. I believe that both the explanatory and the deliberative projects satisfy both conditions. But I concede that a fuller development of the argument would have to address this issue.

[31] For a little more, see Schechter and Enoch (forthcoming: section 6). For much more, see Enoch and Schechter (forthcoming).

you face the challenge of distinguishing between the two. What reason is there, then, to take the explanatory project seriously that is not equally a reason to take the deliberative project seriously? I cannot think of one. And so I tentatively conclude that the deliberative project is intrinsically indispensable if the explanatory one is, that the explanatory project is in no relevant way privileged compared to the deliberative one.[32]

The deliberative project is, then, intrinsically indispensable (or at least—it is intrinsically indispensable if the explanatory one is). If it is instrumentally indispensable for the deliberative project that *p*, we are justified in believing that *p*. At least, we are every bit as justified in so believing as we are in believing the conclusions of inferences to the best explanation (from warranted premisses). If, then, it can be established that irreducibly normative truths are deliberatively indispensable, we are every bit as justified in believing in them as we are in believing in the explanation-friendliness of the universe, and, derivatively, in electrons.

7. DELIBERATION AND THE INSTRUMENTAL INDISPENSABILITY OF NORMATIVE TRUTHS

Law school turned out not to be all you thought it would be, and you no longer find the prospects of a career in law as exciting as you once did. For some reason you don't seem to be able to shake off that old romantic dream of studying philosophy. It seems now is the time to make a decision. And so, alone, or in the company of some others you find helpful in such circumstances, you deliberate. You try to decide whether to join a law firm, apply to graduate school in philosophy, or perhaps do neither.

The decision is of some consequence, and so you resolve to put some thought into it. You ask yourself such questions as: Will I be happy practicing law? Will I be happier doing philosophy? What are my chances of becoming a good lawyer? A good philosopher? How much money does a reasonably successful lawyer make, and how much less does a reasonably successful philosopher make? Am I, so to speak, more of a philosopher or more of a lawyer? As a lawyer, will I be able to make a significant political difference? How important is the political difference I can reasonably expect to make? How important is it to try and make *any* political difference? Should I

[32] Indeed, there may even be some reason to think that the deliberative project is privileged compared to the explanatory one, because when explaining we evaluate competing explanations. See Sayre-McCord (1988: 277–81) and Wiggins (1990: 66, footnote 5).

give any weight to my father's expectations, and to the disappointment he will feel if I fail to become a lawyer? How strongly do I really want to do philosophy? And so on. Even with answers to most—even all—of these questions, there remains the ultimate question. 'All things considered', you ask yourself, 'what makes best sense for me to do? When all is said and done, what should I do? What *shall* I do?'

When engaging in this deliberation, when asking yourself these questions, you assume, so it seems to me, that they have answers. These answers may be very vague, allow for some indeterminacy, and so on. But at the very least you assume that some possible answers to these questions are better than others. You try to find out what the (better) answers to these questions are, and how they interact so as to answer the arch-question, the one about what it makes most sense for you to do. You are not trying to create these answers. Of course, in an obvious sense what you will end up doing is up to you (or so, at least, both you and I are supposing here). And in another, less obvious sense, perhaps the answer to some of these questions is also up to you. Perhaps, for instance, how happy practicing law will make you is at least partly up to you. But, when trying to make up your mind, it doesn't feel like just trying to make an arbitrary choice. This is just not what it is like to deliberate. Rather, it feels like trying to make the *right* choice. It feels like trying to find the best solution, or at least a good solution, or at the very least one of the better solutions, to a problem you're presented with. What you're trying to do, it seems to me, is to make the decision it makes most sense for you to make. Making the decision is up to you. But which decision is the one it makes most sense for you to make is not. This is something you are trying to discover, not create.[33] Or so, at the very least, it feels like when deliberating.

Deliberation, then, is the process of trying to make the decision it makes most sense for one to make. And, as the discussion above suggests, it has a distinctive phenomenological feel.

Thus, deliberation should be distinguished from the making of an arbitrary choice. You're in the supermarket, intending to get a cereal. You may have good reasons to pick Mini-Wheats rather than Raisin Bran (you just don't like Raisin Bran that much), perhaps even one brand over another (the Kellogg's one is usually fresher). But you have no reason, it seems, to pick one package of Kellogg's Mini-Wheats over another, and you know you don't. Of course, you have reason to pick one rather than none at all.

[33] 'In deliberation we are trying to arrive at conclusions that are correct in virtue of something independent of our arriving at them.' (Nagel 1986: 149). For a similar point, though restricted to the case of making a moral choice, see Dancy (1986: 172).

But you've already decided you'll pick one rather than none at all. All that remains to be done now is just to pick a specific package arbitrarily. I take it to be uncontroversial that sometimes we just pick.[34] And it is one lesson of the unfortunate fate of Buridan's ass that picking arbitrarily may often be the rational thing to do.[35] But it is clear, I think, that the phenomenology of arbitrary picking is very different from that of deliberation, of trying to make the right decision.

It is worth noting how *similar* the phenomenology of deliberation is to that of trying to find an answer to a straightforwardly factual question: When trying to answer a straightforwardly factual question (like what the difference is between the average income of a lawyer and a philosopher) you try to get things right, to come up with the answer that is—independently of your settling on it—the right one. When deliberating, you also try to get things right, to decide as—independently of how you end up deciding—it makes most sense for you to decide.

In the supermarket, you have no (normative) reason to pick one package of Mini-Wheats rather than another. With the only relevant decision to be made being which one to pick, there is no one option it makes most sense for you to pursue. More than that, it isn't even the case that one option is at all better than any other. And you know all this. Now, as mentioned before, this doesn't preclude your just picking a package of cereal. Though, if you come to reflect on your situation, you may feel some discomfort, we are not typically—certainly not always—paralyzed in such situations. We can just pick in the face of a known (or believed) absence of reasons. But we cannot, it seems, *deliberate* in the face of a believed absence of reasons. Knowing that there is no decision such that it makes most sense for us to make it, we cannot—not consistently, anyway, in a perfectly commonsensical sense of 'consistently'—try to make the decision it makes most sense for us to make. Deliberation—unlike mere picking—is an attempt to eliminate arbitrariness by discovering (normative) reasons, and it is impossible in a believed absence of such reasons to be discovered.

[34] For a discussion of such cases—and for references to some who question what I say in the text is uncontroversial—see Ullmann-Margalit and Morgenbesser (1977), from which the example is taken (though somewhat modified). They also introduce some helpful terminology: They suggest a distinction between choosing (for reasons) and picking (arbitrarily, in the kind of case described in the text), with 'selecting' being the generic term. For similar distinctions see Darwall (1983: 69), Kolnai (1962: 213) and Railton (1997: 64, n. 12).

[35] The interesting questions regarding Buridan's ass are, I think, not *whether* we can just pick (we obviously can), and not *whether* cases of just picking can be beneficial (they obviously can), but rather *how* it is that, rational creatures that we are, we can just pick, and *how* it is that just picking can be the rational thing to do.

Thus, in deliberating, you *commit* yourself to there being (normative) reasons relevant to your deliberation.[36] Now, this sense of commitment need not entail an explicit belief that there are such reasons, and it certainly doesn't preclude an explicit belief in their non-existence (this is psychologically possible, of course, because people are often inconsistent). Nevertheless, in a perfectly good sense of 'commitment', by deliberating you've already committed yourself to the existence of reasons. To see what I mean by commitment here,[37] think of a reasoner who routinely infers to the best explanation. Now, she may not be a very reflective reasoner, and so she may not have any beliefs *about* which inductive inference rules are valid and why. Or perhaps she's been convinced by some of the literature criticizing IBE, and she now explicitly believes that IBE is not a good rule of inference. Nevertheless, by routinely inferring to the best explanation, she commits herself to IBE being a good rule of inference. If she believes that IBE is not a good inference-rule, she is being inconsistent (though perhaps in a somewhat generalized sense of this term)—unless, that is, she has some story available to her explaining how her use of IBE is compatible with her explicit rejection of it (perhaps, for instance, by showing that IBE is, though generally fallacious, actually harmless in a privileged class of cases, and by restricting her own use of IBE to such cases). Similarly, I want to argue, by deliberating you commit yourself to there being relevant reasons; if you also believe there aren't any, you are being inconsistent in exactly the same sense, and just as irrational, too.

Notice that no such commitment is involved in cases of mere picking. Neither by picking one package of Mini-Wheats from all the others nor by going through some mental process beforehand, do you commit yourself to there being any reason that makes your package more worth picking than the others (you may commit yourself to there being reason to pick some package rather than none at all, but this is a different matter). It is, then, a result of the nature of deliberation—an attempt to eliminate the arbitrariness so typical in cases of mere picking—that by deliberating, by asking yourself which choice it would make most sense for you to make, you are committing yourself to there being reasons relevant to your choice. Suppose a friend of yours seems to undergo a process of deliberation, but

[36] As already mentioned, I am happy invoking explicitly ontological terms, talking about the existence of reasons. But if for some reason you find talk of normative truths less problematic than talk of the existence of normative entities (because, perhaps, less offensive to your naturalist leanings), feel free to paraphrase my claims along such lines. For myself, I cannot see why a commitment to the existence of (irreducibly normative) reasons is any more of an offense to naturalism than a commitment to (irreducibly) normative truths. But I need not develop this point here.

[37] I thank Stephanie Beardsman and Derek Parfit for pressing me on this issue.

then—when asked, perhaps—says that it really doesn't matter one way or another, that there is absolutely nothing to be said for or against any of the relevant alternatives, that there are no considerations counting in favor of any of his possible decisions. You would treat him either as having changed his mind ('Oh, he thought, until just a moment ago, that there was a point to his deliberation, but now he understands that this is not so'), or as being inconsistent. You would treat him as you would someone who professes to reject IBE and nevertheless infers to the best explanation—he has either changed his mind about IBE ('Oh, he thought, until just a moment ago, that we should not infer to the best explanation, but now he sees that he was wrong about that'), or he is being inconsistent. What explains this attitude of yours, I think, is precisely that both *are* being inconsistent. And this is also why, upon coming to believe that there are no relevant reasons, deliberation stops (though a decision may remain to be made).

Now, that something is a (normative) reason for you to join a law firm, a consideration that counts in favor of so doing, is a paradigmatically normative claim, as is that pursuing graduate studies in philosophy is the thing that makes most sense for you to do. So, by deliberating, you commit yourself to there being relevant reasons, and so to there being relevant normative truths (you do not, of course, commit yourself to the reasons *being* the normative truths). Normative truths are thus indispensable for deliberation.[38]

But I hear objections. *Don't we sometimes deliberate when we know that the weight of reasons is balanced, so that no option is the best?* Yes, but by doing so we betray our lack of confidence in this normative judgment, and our suspicion that there may be reasons we've overlooked, or ones to which we haven't assigned the right weight. *Well, aren't our desires enough for deliberation? Why do we need normative truths to settle deliberation, when we are moved by desires?* Because when you allow yourself to settle a deliberation by reference to a desire, you commit yourself to the normative judgment that your desire made the relevant action the one it makes most sense to perform. So even with desires at hand, you still commit

[38] 'The ordinary process of deliberation, aimed at finding out what I should do, assumes the existence of an answer to this question' (Nagel 1986: 149). For similar points, see Bond (1983: 60), Darwall (1983: 224) (though Darwall doesn't make this point regarding the deliberation of agents in general, but rather only regarding his 'ISIS', an internally self-identified subject), Kolnai (1962) (though Kolnai, being a skeptic of sorts regarding normative truths, draws skeptical conclusions about deliberation as well), and Pettit and Smith (1998: e.g. 97) (who argue that deliberation is a kind of conversation one has with oneself, and that adopting this kind of conversational stance—to oneself as well as to others—involves assumptions, one of which is rather close to the one in the text).

yourself to a normative truth. *Anyway, we don't necessarily explicitly invoke normative truths when deliberating.* True, but that doesn't mean we don't commit ourselves to normative truths when deliberating. The reasoner who routinely infers to the best explanation need not have explicit beliefs about IBE being a good rule of inference. But she is nevertheless committed to this claim. *Well, can't we deliberate even believing there are no normative truths, just like you can try to move a rock believing you will fail, indeed in order to prove to your friends that the rock is too heavy to be moved?*[39] Perhaps we can, but, first, this way of deliberating seems in an important sense parasitic on the more common one, where one believes that one is at least somewhat likely to succeed. So it's not clear that this line of thought can be applied to deliberation as a whole (rather than to some particular cases). And second, even if such deliberation is possible, it is clearly less attractive than the fuller deliberative projects, where one tries to find answers one believes are there to be found. And on the account of instrumental indispensability presented above, this suffices to establish the instrumental indispensability of the belief in normative truths. *So perhaps in order to deliberate you have to believe that there* may *be normative truths to be found. This is still no reason to believe that there* are *such truths.*[40] Some delicate modal questions are relevant here. If, for instance, the modality invoked in the objection is something like the possibility of everyday, practical, 'can-do' locutions, then at least given the robust modal status of normative truths (if they exist), the (practical) possibility of discovering them may entail their actuality. But even putting this point to one side, still the answer to the previous objection holds: For the retreat to the possibility (rather than actuality) of the existence of normative truths takes something, it seems, from the strength of the reason to engage in the deliberative project. And note, of course, that the line of thought expressed in the objection—whatever its ultimate strength—cannot serve to distinguish between the case of normative truths and the case of whatever is necessary for the explanatory project. If possibility suffices for the former, the mere possibility that the universe may be explanation-friendly should suffice for the latter. *Well then, what if deliberation is simply illusory? Perhaps what is needed in order to explain deliberation is not normative truths, but rather a good error theory.*[41] The first thing to note in reply here is that I do not argue that normative truths are needed for the *explanation* of

[39] I owe this example to John Gibbons.

[40] I thank Pete Graham and Josh Schechter for pressing me on this point.

[41] My argument for Robust Realism may be thought of as a kind of a transcendental argument. And this objection is close in spirit to Stroud's (1968) famous objection to transcendental arguments—namely, that at most they show that *belief* in the disputed claim is necessary, not that its truth is. See also note 45 below.

deliberation.[42] I want to remain neutral on this and all other explanatory questions. Normative truths are needed, so I've argued, not necessarily for the person observing the phenomenon of deliberation ('from outside', as it were), but for the deliberating agent herself. It is still possible, of course, that deliberation is illusory, that it essentially relies on a false belief in normative truths. But we would need a very strong argument to believe that, perhaps as strong as the argument we would need in order to believe that the universe is not even reasonably explanation-friendly. (*Can* there even be such an argument?) The arguments meant to show that there are no irreducibly normative truths—a huge promissory note coming up—show no such thing.

These objections—and others—deserve a more serious treatment than I can give them here. But at this point it should be fairly clear, I think, how the most pressing objections are to be dealt with.

8. A QUICK NOTE ABOUT THE MOVE FROM INDISPENSABILITY TO (JUSTIFIED) BELIEF

There is a very different kind of worry that needs to be addressed: Why think that indispensability—explanatory or deliberative—is any guide at all to ontology? 'Even granting you the details of your indispensability argument,' my interlocutor can say, 'all you've shown is that, in some sense, we *need* normative truths. But how is this any reason at all to believe there are such things? Perhaps you've established that it would be nice if there were normative truths, or that we deeply want them to exist. But concluding from this to the *belief* in normative truths is a clear instance of wishful thinking.'[43]

[42] My emphasis on deliberation is in some respects very close to some of the things Gibbard says in *Thinking How to Live* (2003), but our conclusions are very different. The point in the text explains, I think, why: one of Gibbard's major lines of argument against the realist is, I think, that expressivism can explain all that needs to be explained about deliberation ('we don't *need* queer properties to explain reasoning what to do'; 2003: 7). Even if this is so, though, my argument stands, for the reason given in the text: I do not claim that Robust Realism is what is needed in order to (third-personally) explain deliberation, but rather in order to (first-personally) engage in deliberation.

[43] Here is a similar accusation from Korsgaard (1996: 33): 'Having discovered that he needs an unconditional answer, the realist straightaway concludes that he has found one.' (Korsgaard doesn't, of course, address my argument; this sentence is taken from her criticism of realists (primarily Nagel) whose views and arguments are—though distinct from—nevertheless closely related to mine.) Zimmerman (1984: 95) makes a similar point in criticizing Platts. And Russ Shafer-Landau (2003: 29, n. 11) makes a similar point against Wiggins's different but related critique of (what he calls) non-cognitivism.

The first thing to note in reply to this objection is that it applies just as forcefully to arguments from explanatory indispensability: 'Even granting you the explanatory indispensability of numbers, or electrons, or whatever,' someone may argue similarly, 'all you've shown is that, in some sense, we *need* there to be electrons and numbers if we are going to make sense of the world. But how is this any reason at all to believe that there are such things? Perhaps you've established that it would be nice if there were electrons, or that we deeply want them to exist (because we want the world to make sense to us). But concluding from this to the *belief* in electrons is a clear instance of wishful thinking.' Of course, had we had independent reason to believe that the universe was such as to make sense to us, that it was at least by and large intelligible, that it was explanation-friendly, this problem would go away. Similarly, if we had independent reason to believe that the universe was deliberation-friendly, the analogous worry about my own argument would go away. But it does not seem like we have—or indeed can have—independent reason to believe such things. And so the worry stands.

Of course, justifications come to an end somewhere. And perhaps this is where: We usually take that something is theoretically useful to be reason to believe it, and perhaps we should rest content with that as a fairly basic epistemic procedure, as one place where epistemic justification comes to an end.[44] But as always, the justifications-come-to-an-end-somewhere reply is not very satisfying.

In fact, I think a somewhat more satisfying response can be given, one that while, in a sense, grounds epistemic justification in pragmatic utility, nevertheless respects the autonomy and uniqueness of epistemic justification. Elsewhere, Joshua Schechter and I present an account of the justification of basic belief-forming methods that has this feature, and that vindicates indispensability arguments—both explanatory and deliberative.[45] But because I cannot discuss this in detail here, let me settle for the following dialectical point.

[44] David Lewis, for instance, is not terribly worried. In introducing his argument for modal realism—a rather surprising ontological thesis, no less surprising than Robust Metanormative Realism, I would hope—he says: 'I begin the first chapter by reviewing the many ways in which systematic philosophy goes more easily if we may presuppose modal realism in our analyses. I take this to be a good reason to think that modal realism is true, just as the utility of set theory in mathematics is a good reason to believe that there are sets' (1986: vii).

[45] See Enoch and Schechter (forthcoming). Because my argument for Robust Realism may be thought of as a transcendental argument of sorts, I want to disassociate myself from one (other) way of understanding such arguments. Transcendental arguments are sometimes presented as attempts to show that the relevant sceptical position or argument is unstable, that the sceptic defeats herself. In this spirit, one may argue that the (deliberating) normative sceptic is shown by my reasoning to be inconsistent, and

Think again of indispensability arguments in the philosophy of mathematics, where numbers and sets are said to be indispensable to science. Someone who rejects IBE and the existence of electrons with it is under no pressure to acknowledge the existence of mathematical objects because of their role in scientific theories. A full defense of Mathematical Platonism will, I guess, have to address such philosophers as well. But this is not the work supposed to be done by the indispensability arguments themselves. These are targeted at the naturalist who accepts electrons but is reluctant to accept numbers. If the price one has to pay in order to reject numbers is a denial of the existence of electrons, Mathematical Platonism may not be completely vindicated, but it certainly gains plausibility points.

Analogously, then: If a complete defense of Robust Realism is to be presented, the general worry about the move from indispensability to belief has to be addressed. But this is not the work supposed to be done by the indispensability argument itself. This argument is targeted primarily at the metaphysical naturalist who accepts arguments from explanatory indispensability (along with electrons and perhaps also numbers) but is reluctant to accept arguments from deliberative indispensability (along with normative facts). If the price one has to pay in order to reject normative facts is a denial of the existence of electrons, and of the validity of IBE more generally, Robust Realism may not be completely vindicated, but it certainly gains plausibility points. And for now I am happy to settle for this result.

9. (FURTHER) SUPPORTING THE INDISPENSABILITY PREMISS

One more step is necessary for the indispensability argument for Robust Realism. Alternative metanormative views must be rejected, and in particular, it must be shown that no alternative metanormative theory can deliver the goods that are deliberatively indispensable. For if a non-robust-realist view of normativity and normative discourse can supply all that is needed for sincere deliberation, irreducibly normative truths are after all not deliberatively indispensable. Think again of indispensability arguments in the philosophy of mathematics: If a non-Platonist view of mathematical discourse and entities can supply all that is needed for scientific explanations

perhaps this justifies the belief in normative truths—the very belief our entitlement to which the relevant sceptic is questioning. But this line of thought, I think, fails. For some reasons, see Wright (1991), and Enoch (2006: section 4.3). The justification of basic belief-forming methods Schechter and I develop is not of this kind.

(and is adequate otherwise), numbers (Platonistically understood) are after all not explanatorily indispensable and the indispensability argument (as an argument for Platonism) fails.

In the context of my argument for Robust Realism, rejecting alternative metanormative views is thus not a luxury: It is not merely a further dialectical step, enhancing the plausibility of one view by reducing that of others. Nor is it an instance of the (purported) flaw that is typical of the writing of many realists—that of writing mostly negatively, rejecting other views while having very little by way of positive argument in support of their realism.[46] Rather, rejecting alternative views is part of the positive argument—the argument from deliberative indispensability—for Robust Realism.

So what is needed here is nothing less than a survey of the metanormative field, and arguments showing that each less-than-robust-realist view of normativity does not suffice for deliberation (either directly, or indirectly, in that it fails in some other way, and so on the whole cannot accommodate deliberation). Rather than go through such a survey, let me emphasize some of the general points such a survey would, I believe, bring to light.

Because only normative truths can answer the normative questions I ask myself in deliberation, nothing less than a normative truth suffices for deliberation. Furthermore, it seems like nothing but an *explicitly* normative truth suffices for deliberation. And because the kind of normative facts that are indispensable for deliberation are just so different from naturalist, not-obviously-normative facts and truths, the chances of a naturalist reduction seem rather grim.[47] For similar reasons, the chances of a Neo-Aristotelian metaethical or metanormative view that blurs the normative-natural distinction (perhaps utilizing, as Bloomfield (2001) does, an analogy with such concepts as *healthy*) do not seem promising. The gap between the normative and the natural, considered from the point of view of a deliberating agent, seems unbridgeable.

An honest non-cognitivist or expressivist—even a quasi-realist—will have to agree that there is *a* sense in which all normativity is grounded in the attitudes she just happens to find herself with.[48] Such views, then, do

[46] See Korsgaard (1996: 31) (referring to Clarke and Price). And Russ Shafer-Landau too (2005: 264) characterizes his defense of moral realism in his (2003) as essentially an argument from elimination.

[47] For an interesting discussion of what the reductionist claim comes to, see Schroeder (2005). Though Schroeder's discussion can help in rooting out some common mistakes about reduction, it does not, I think, successfully deal with the just-too-different intuition I use in the text. For some more statements of this 'just-two-different' intuition, and for some discussion, see Dancy (2005, mostly on 141) and Fitzpatrick (forthcoming).

[48] His many protests notwithstanding, this is true even of Blackburn. See his (1981: 164–5) parent metaphor. See also Gibbard's (2003: e.g. 82) characterization of

not allow for serious deliberation about these very attitudes. And while the unending ingenuity of non-cognitivists can make this problem less obvious, it cannot, it seems to me, make it go away. So expressivist views cannot allow for the full scope of deliberation. Furthermore, even if miraculously they can, still the indispensability argument for Robust Realism stands. Remember, on the account of instrumental indispensability I endorsed (following Colyvan), for something to be instrumentally indispensable for a project it is not necessary that it cannot be eliminated from that project. Rather, it is sufficient that it cannot be eliminated without defeating whatever reason we had to find that project attractive in the first place. And the deliberative project loses much of its initial appeal, it seems to me, once normativity is viewed as dependent on our attitudes. (In this respect, *cognitivist* subjectivist theories, of course, do no better.)

Can an error theory do better? In one respect, the answer seems to be positive. For, as is widely noted, error theorists at least acknowledge the full strength of the commitments of normative discourse. Nevertheless, error theories will not block the argument from deliberative indispensability to Robust Realism. Error theorists have a decision to make: They have to decide whether to continue engaging in the discourse they are error theorists about, presumably justifying doing so on instrumental grounds of some sort, or to abandon the discourse altogether. They have to choose, in other words, between Instrumentalism and Eliminativism. But Instrumentalism is or entails a normative claim—roughly, that it makes sense to continue using normative language even though normative discourse is systematically erroneous—and so is arguably unavailable to the error theorist about normative discourse. And global metanormative Eliminativism is simply not an option for deliberative creatures. Or so, at least, I have argued.

This is all very sketchy, of course. Proponents of Normative Naturalism (of many different kinds), or of Expressivism, or of Error Theory, may very well have retorts available to them, ones that need to be addressed. And there may be other alternatives as well: a revisionary account, perhaps, that is error-theoretic in a way, but that somehow avoids both Instrumentalism and Eliminativism; or perhaps a dispositional theory that somehow avoids a naturalist reduction;[49] or perhaps—though I doubt it—there is room somewhere in logical space for a constructivist view that avoids the

expressivism in terms of such explanatory priority. So I am not among those impressed by the ability of the expressivist to accommodate everything the realist wants to say. If he is to have a distinct position, the expressivist must concede *a* sense in which normativity is response-dependent (even if there are other senses in which he can argue it is not). And denying *this* sense is one of the things the realist wants to say.

[49] For an argument against dispositional theories that employ some idealization, see my 'Why Idealize?' (Enoch 2005).

classification above.[50] But enough has been said, I hope, to appreciate the challenge such views face if they are to supply what is needed for deliberation. And I do not see on the horizon an alternative metaethical view that successfully addresses this challenge.[51]

10. A TENTATIVE CONCLUSION

As it stands, my argument for Robust Realism is incomplete. What is needed to complete it is, first, a more detailed discussion of other metanormative views, and second, a more detailed defense of the claim that deliberation does commit one to the existence of normative truths against objections. Together, such discussions will complete the argument for the indispensability premiss. And for the argument to be *completely* complete, a justification of the move from indispensability—deliberative *or* explanatory—to belief should be vindicated.[52] Even this will not suffice for a full defense of Robust Realism. For that, traditional objections to the view—objections from metaphysical queerness, from epistemological and semantic access, from supervenience, from disagreement, and from the relations between normativity and motivations, to name the most influential—will have to be addressed. So I do not want to pretend that the work of the Robust Realist is done.

But it is not premature, I think, to draw the following conclusions. First, arguments from deliberative indispensability are prima facie as respectable as the more common arguments from explanatory indispensability. Absent some story distinguishing between the two, taking the latter but not the former seriously is an arbitrary and so unjustified philosophical move. And second, the opponents of Robust Realism are going to have to play the game on the Robust Realist's home court: They are going to have to show

[50] For an argument against the currently fashionable attempts to ground normativity in what is constitutive of action—either independently, or within a constructivist framework—see my 'Agency, Shmagency' (2006). Nothing in my argument counts against a more modest constructivist view, one that, for instance, attempts an account of some part of the normative domain in terms of another part *of that domain*. But such a modest constructivist view cannot, of course, be the full meta*normative* story (for all I've said, it may be the full meta*ethical* story).

[51] Perhaps Michael Ridge's 'Ecunemical Expressivism' (presented in his contribution to this volume) is one such new alternative? Perhaps so. And perhaps his view can cleverly avoid some of the pitfalls other expressivist views fall pray to. But it cannot, it seems to me, avoid the problem of (some objectionable kind of) response-dependence mentioned above. And the ideal-observer version of the ecunemical expressivist view may—on top of that—be subject to the objection to idealization I put forward in my (2005).

[52] For many details, again see Enoch (2003).

how their view is compatible with the phenomenology of deliberation. This is not a challenge they have been too enthusiastic to address (if I am right, not without reason[53]). But it is here that the battle is to be fought.

REFERENCES

Audi, Robert (1997) *Moral Knowledge and Ethical Character* (Oxford: Oxford University).

Blackburn, Simon (1981) 'Reply: Rule-Following and Moral Realism,' in S. H. Holtzman and C. M. Leich (eds.), *Wittgenstein: To Follow a Rule* (London: Routledge & Kegan Paul), 163–87.

____ (1991a) 'Just Causes,' *Philosophical Studies* 61: 3–17.

____ (1991b) 'Reply to Sturgeon,' *Philosophical Studies* 61: 39–42.

Bloomfield, Paul (2001) *Moral Reality* (Oxford: Oxford University Press).

Bond, E. J. (1983) *Reason and Value* (Cambridge: Cambridge University Press).

Bransen, J., and Cuypers, S. E. (eds.) (1998) *Human Action, Deliberation and Causation* (Dordrecht: Kluwer Academic Publishers).

Brink, D. O. (1989) *Moral Realism and the Foundations of Ethics* (Cambridge: Cambridge University Press).

Colyvan, Mark (2001) *The Indispensability of Mathematics* (Oxford: Oxford University Press).

Copp, David (1990) 'Explanation and Justification in Ethics,' *Ethics* 100: 237–58.

____ and Zimmerman, D. (eds.) (1984) *Morality, Reason and Truth: New Essays on the Foundations of Ethics* (Totowa: Rowman & Allanheld).

Cullity, G., and Gaut, B. (eds.) (1997) *Ethics and Practical Reason* (Oxford: Clarendon Press).

Dancy, Jonathan (1986) 'Two Conceptions of Moral Realism,' *Proceedings of the Aristotelian Society, Supp.* 60: 167–88.

____ (2005), 'Nonnaturalism,' in David Copp (ed.), *The Oxford Handbook of Ethical Theory* (Oxford: Oxford University Press), 121–44.

Darwall, S. L. (1983) *Impartial Reason* (Ithaca, NY: Cornell University Press).

Dworkin, Ronald (1996) 'Objectivity and Truth: You'd Better Believe It,' *Philosophy and Public Affairs* 25: 87–139.

Enoch, David (2003) *An Argument for Robust Metanormative Realism* (Dissertation, NYU), available at **http://law.mscc.huji.ac.il/law1/newsite/segel/enoch/index.html**

____ (2005) 'Why Idealize?,' *Ethics* 115: 759–87.

[53] Though see Stephen Finlay's heroic attempt to accommodate the phenomenology of deliberation on Humean premisses in his contribution to this volume. Let me remind you that I do not argue that irreducibly normative truths are necessary for the *explanation* of deliberation, but rather *in order to* deliberate. Perhaps Finlay can be understood as arguing that desires are sufficient for explaining deliberation. If, though, he also tries to accommodate deliberation itself, it is for the reader to judge how phenomenologically successful his attempt is.

Enoch, David (2006) 'Agency, Shmagency: Why Normativity Won't Come from What Is Constitutive of Action', *Philosophical Review* 115: 169–98.
____ 'How is Moral Disagreement a Problem for Realism?,' unpublished manuscript.
____ 'The Epistemological Challenge to Metanormative Realism: How Best to Understand It, and How to Cope with It,' work in progress.
____ and Schechter, Joshua (forthcoming) 'How Are Basic Belief-Forming Methods Justified?,' *Philosophy and Phenomenological Research*.
Field, Hartry (1980) *Science without Numbers: A Defence of Nominalism* (Oxford: Blackwell).
____ (1989) *Realism, Mathematics, and Modality* (New York: Basil Blackwell).
____ (1991) 'Metalogic and Modality,' *Philosophical Studies* 62: 1–22.
Fitzpatrick, William J. (2005) 'The Practical Turn in Ethical Theory: Korsgaard's Constructivism, Realism and the Nature of Normativity,' *Ethics* 115: 651–91.
____ (forthcoming) 'Robust Ethical Realism, Non-Naturalism and Normativity,' *Oxford Studies in Metaethics*, 3.
Gibbard, Allan (1990) *Wise Choices, Apt Feelings* (Cambridge, Mass.: Harvard University Press).
____ (2003) *Thinking How to Live* (Cambridge, Mass.: Harvard University Press).
Grice, Paul (1975) 'Method in Philosophical Psychology (From the Banal to the Bizarre),' *Proceedings and Addresses of the American Philosophical Association* 48: 23–53.
Harman, Gilbert (1977) *The Nature of Morality* (Oxford: Oxford University Press).
____ (1984) 'Is There a Single True Morality?,' in Copp and Zimmerman 1984: 27–48.
____ (1986) 'Moral Explanations of Natural Facts: Can Moral Claims Be Tested against Moral Reality?,' *Southern Journal of Philosophy* 24 (supp.): 57–68.
____ (1998) 'Responses to Critics,' *Philosophy and Phenomenological Research* 58: 207–13.
____ and Thomson, J. J. (1996) *Moral Relativism and Moral Objectivity* (Oxford: Blackwell).
Honderich, Ted (ed.) (1985) *Morality and Objectivity* (Boston: Routledge and Kegan Paul).
Kolnai, Aurel (1962) 'Deliberation is of Ends,' *Proceedings of the Aristotelian Society* 62: 195–218.
Korsgaard, Christine (1996) *The Sources of Normativity*, ed. Onora O'Neill (Cambridge: Cambridge University Press).
Leiter, Brian (2001) 'Moral Facts and Best Explanations,' *Social Philosophy and Policy* 18(2): 79–101.
Lewis, David (1986) *On the Plurality of Worlds* (Oxford: Blackwell).
Lycan, W. G. (1986) 'Moral Facts and Moral Knowledge,' *Southern Journal of Philosophy* 24, supp.: 79–94.
McDowell, John (1985) 'Values and Secondary Qualities,' in Honderich 1985: 110–29.
McGinn, Colin (1997) *Ethics, Evil, and Fiction* (Oxford: Clarendon Press).

Moore, M. S. (1992) 'Moral Reality Revisited,' *Michigan Law Review* 90: 2425–533.
Nagel, Thomas (1986) *The View from Nowhere* (New York: Oxford University Press).
Oddie, Graham (2005) *Value, Reality and Desire* (Oxford: Oxford University Press).
Parfit, Derek (2006) 'Normativity,' *Oxford Studies in Metaethics* 1.
Pettit, P., and Smith, M. (1998) 'Freedom in Belief and Desire,' in Bransen and Cuypers 1998: 89–112.
Platts, Mark, 'Moral Reality and the End of Desire', in M. Platts (ed.), *Reference, Truth and Reality: Essays on the Philosophy of Language* (London: Routledge & Kegan Paul, 1980), 69–82.
Putnam, Hilary (1995) 'Are Moral and Legal Values Made or Discovered?', *Legal Theory* 1: 5–19.
Quinn, W. S. (1986) 'Truth and Explanation in Ethics,' *Ethics* 96: 524–44.
Railton, Peter (1986) 'Moral Realism,' *Philosophical Review* 95: 163–07.
____ (1997) 'On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action', in Cullity and Gaut 1997: 53–79.
____ (1998) 'Moral Explanation and Moral Objectivity,' *Philosophy and Phenomenological Research* 58 (1998), 175–82.
Regan, D. H. (2003) 'How to be a Moorean', *Ethics* 113: 651–77.
Resnik, M. D. (1995) 'Scientific vs. Mathematical Realism: The Indispensability Argument,' *Philosophia Mathematica* 3: 166–74.
____ (1997) *Mathematics as a Science of Patterns* (Oxford: Clarendon Press).
Rosati, Connie (2003) 'Agency and the Open Question Argument,' *Ethics* 113: 490–527.
Sayre-McCord, Geoffrey (1988) 'Moral Theory and Explanatory Impotence,' *Midwest Studies* 12: 433–57; reprinted in G. Sayre-McCord (ed.), *Essays on Moral Realism* (Ithaca, NY: Cornell University Press, 1988), 256–81.
____ (1992) 'Normative Explanations', *Philosophical Perspectives* 6 (Ethics): 55–71.
Schechter, Joshua, and Enoch, David (forthcoming) 'Meaning and Justification: The Case of Modus Ponens,' *Noûs*.
Schroeder, Mark (2005) 'Realism and Reduction: The Quest for Robustness,' *Philosophers' Imprint* 5(1).
Shafer-Landau, Russ (2003) *Moral Realism: A Defence* (Oxford: Oxford University Press).
____ (2005) 'Precis of *Moral Realism: A Defence*,' *Philosophical Studies* 126: 263–7.
Simon, C. J. (1990) 'The Intuitionist Argument,' *Southern Journal of Philosophy* 28: 91–114.
Slors, M. V. P. (1998) 'Two Claims that Can Save a Nonreductive Account of Mental Causation,' in Bransen and Cuypers 1998: 224–48.
Stratton-Lake, Phillip (2002) 'Introduction,' in P. Stratton-Lake (ed.), *Ethical Intuitionism: Re-evaluations* (Oxford: Oxford University Press), 1–28.
Stroud, Barry (1968) 'Transcendental Arguments,' *Journal of Philosophy* 65: 241–56.
Sturgeon, N. L. (1984) 'Moral Explanations,' in Copp and Zimmerman 1984: 49–78.

Sturgeon, N. L. (1986) 'Harman on Moral Explanations of Natural Facts,' *Southern Journal of Philosophy* 24 (supp.): 69–78.

——(1991) 'Content and Causes: A Reply to Blackburn,' *Philosophical Studies* 61: 19–37.

——(1992) 'Nonmoral Explanations,' in J. E. Tomberlin (ed.), *Philosophical Perspectives* 6 (Ethics): 97–117.

——(1998) 'Thomson against Moral Explanations,' *Philosophy and Phenomenological Research* 58: 199–206.

Ullmann-Margalit, Edna, and Morgenbesser, Sydney (1977) 'Picking and Choosing,' *Social Research* 44: 757–85.

Wiggins, David (1990) 'Moral Cognitivism, Moral Relativism and Motivating Moral Beliefs,' *Proceedings of the Aristotelian Society* 91: 61–86.

Wright, Crispin (1991) 'Scepticism and Dreaming: Imploding the Demon,' *Mind* 100: 87–116.

——(1992) *Truth and Objectivity* (Cambridge: Harvard University Press).

——(1993) 'Realism: The Contemporary Debate—W(h)ither Now?,' in J. Haldane and C. Wright (eds.), *Reality, Representation and Projection* (New York: Oxford University Press), 63–84.

Yasenchuk, Kenc (1994) 'Sturgeon and Brink on Moral Explanations,' *Southern Journal of Philosophy* 32: 483–502.

Zimmerman, David (1984) 'Moral Realism and Explanatory Necessity,' in Copp and Zimmerman 1984: 79–103.

3

Ecumenical Expressivism: The Best of Both Worlds?

Michael Ridge

'Evaluative judgment has a decidedly Janus-faced character.'
Michael Smith

Michael Smith's thesis is about evaluative judgments, but he could just as easily be talking about normative judgments—judgments about reasons for action and judgments about what one ought to do. Indeed, on T. M. Scanlon's intriguing 'buck-passing' account of evaluative judgment, evaluative judgment just *is* normative judgment in disguise.[1] In what sense are normative judgments Janus-faced, though? In some respects, they seem like ordinary beliefs. We call them 'beliefs', as when we say things like, 'Britney believes that she ought to spend more time at the tanning salon.' We classify them as true or false. We sometimes think they constitute knowledge. They figure in apparently rational inferences. In other respects, though, normative judgments seem more like desires. Normative judgment is practical; it reliably guides action. Changes in normative view reliably track changes in motivation. We question the sincerity of someone who claims that she really ought to do something but shows no signs whatsoever of being motivated to do it, feel bad about not doing it, etc. Failure to act on one's all things considered normative judgment is irrational.[2] This contrasts with acting contrary to what one believes is required by merely conventional norms like those of etiquette. Finally, normative disagreement can without irrationality persist in the face of agreement on all the relevant facts. Nor is this disagreement well understood in terms of vagueness, at least

[1] See Scanlon (1998).

[2] This is not uncontroversial, but this is not the place to discuss the controversy. For a contrasting view, see Arpaly (2000).

in many cases.[3] Sometimes people have fundamentally different normative outlooks. For example, some people think the decisive question to ask about gay marriage is what God wills while others think the decisive question is whether gay marriage fosters relationships based on love, mutual respect, etc. This does not seem like a case of a shared normative conception with vague contours, so much as fundamentally and deeply different normative outlooks. Yet such people nonetheless disagree. Moreover, it is plausible to suppose that they will agree about what ought to be done not so much when they come to agree on some further fact about gay marriage, but instead when they take the same practical stance to it. This suggests that normative disagreement is what Charles Stevenson called 'disagreement in attitude', rather than disagreement in belief.[4] Once again, this suggests that normative judgments are better understood as desire-like states as opposed to belief-like states.

These competing characteristics of normative judgments have led to the formation of two diametrically opposed philosophical camps—the cognitivists and the expressivists. Cognitivism is traditionally defined as the doctrine that normative utterances express beliefs rather than desires. Expressivism, by contrast, is traditionally defined as the doctrine that normative utterances express desires rather than beliefs. For example, David Brink characterizes the expressivist as holding the view that 'moral judgments must express the appraiser's non-cognitive attitudes, *rather than her beliefs*'[5] while 'cognitivists interpret moral judgments as expressing cognitive attitudes, such as belief, *rather than non-cognitive attitudes*, such as desire.'[6] Frank Jackson and Philip Pettit characterize expressivism as holding that moral utterances express desires rather than beliefs (Jackson and Pettit 1998). Although these views are often formulated as views about moral judgment in particular, it is clear that they are very often taken to be plausible views about normative judgment more generally.

Unfortunately, the terms of this debate mask the following logical space:

The Ecumenical View: Normative sentences are conventionally used to express *both* beliefs and desires.

[3] Thanks to Joshua Gert for drawing me out on this point.

[4] The Janus-faced character of normative discourse is often noted. In addition to Michael Smith's discussion, Mark Lance and John O'Leary-Hawthorne note the dual aspects of normative judgment: 'normatives are in many ways just like ordinary declaratives. They take their place in the game of giving and asking for reasons, serving as premises and conclusions in reasoning … But in another crucial respect their consequences of application are like imperatives … one of the direct, and widely stable, consequences of application of a normative is the appropriateness of some act; to commit oneself to a normative is *ipso facto* to commit oneself to the propriety of some act' (Lance and O'Leary-Hawthorne 1997: 202–3).

[5] Brink 1997: 9, emphasis added.

[6] Brink 1997: 5, emphasis added.

I argue that a version of the Ecumenical View combines many of the best features of traditional (that is, Non-Ecumenical) forms of cognitivism and expressivism, while avoiding the worst vices of each. More specifically, I defend what I shall call 'Ecumenical Expressivism,' as opposed to what I shall call 'Ecumenical Cognitivism.' However, I shall not in this paper attempt to argue that Ecumenical Expressivism is superior to all forms of cognitivism; that would lead quickly to a discussion of familiar issues which would here distract from what is distinctive about Ecumenical Expressivism. Rather, the primary thesis I defend here is that Ecumenical Expressivism is superior to Non-Ecumenical Expressivism. If you are going to be an expressivist at all then you should be an Ecumenical Expressivist.

1. RECASTING THE DEBATE

The Janus-faced nature of normative judgment should make the Ecumenical View seem like an attractive if not obvious one. For the Ecumenical View is well poised to accommodate both the belief-like and desire-like features of normative judgment. Moreover, it promises to do so without abandoning a broadly Humean philosophy of mind; there is no need on the Ecumenical View to posit what are sometimes called 'besires'. A besire is supposed to be a single state of mind which at one and the same time represents the world as being a certain way and motivates the agent in a certain way all by itself, without the help of any independent desire. Many philosophers have, rightly in my view, found the very idea of a besire puzzling. However, this is not the place to rehearse the arguments on both sides of that debate. Suffice it to say that the Ecumenical View can accommodate the idea that beliefs and desires are, in David Hume's memorable terms, 'distinct existences'. The point is simply that normative utterances systematically function to express both.

However, on the traditional way of carving up the metanormative terrain, the Ecumenical View seems to imply that neither expressivism nor cognitivism is correct. In that case, one of the central metanormative debates of the past century has been a tempest in a teapot. This might seem welcome to those weary of apparently interminable debates about those doctrines. However, the issues at stake in that debate remain live ones even if the Ecumenical View is correct. We can usefully redraw the terms of that debate within an ecumenical framework as follows:

Cognitivism: For any normative sentence M, M is conventionally used to express a belief such that M is true if and only if the belief is true.

Non-Cognitivism: For any normative sentence M, M is not conventionally used to express a belief such that M is true if and only if the belief is true.

The distinction between cognitivism and non-cognitivism as drawn here is exclusive but not exhaustive. There is logical space for hybrid views according to which some but not all normative utterances express beliefs which provide their truth-conditions.[7] For present purposes I put these interesting hybrid views to one side, and focus on theories which treat normative discourse uniformly in these respects.[8] More germane here is that expressivism goes beyond non-cognitivism's purely negative thesis. Expressivism traditionally has been understood as making both a negative and a positive claim. The positive claim concerns the conventional role of normative sentences in expressing our pro-attitudes. Intuitively, this puts the 'express' in 'expressivism':

Expressivism: Non-Cognitivism, as defined above, plus the thesis that normative sentences are conventionally used to express pro-attitudes.

In any event, it should be clear enough that, characterized in these terms, there can be both cognitivist and expressivist versions of the Ecumenical View. The Ecumenical Cognitivist and the Ecumenical Expressivist agree that normative utterances express both beliefs and desires. They disagree about the connection between the truth of the belief expressed and the truth of the sentence which expresses it. The cognitivist insists that a given normative sentence is true if and only if the belief it expresses is true, whereas an expressivist denies this.

Ecumenical Cognitivism: Cognitivism, as defined above, plus the thesis that normative sentences are conventionally used to express pro-attitudes (as well as the beliefs which provide their truth-conditions).

Ecumenical Expressivism: Expressivism, as defined above, plus the thesis that normative sentences are also conventionally used to express beliefs (albeit not ones which are not thereby guaranteed to provide the sentences' truth conditions).

Crucially, this characterization of the debate fits well with at least one of the traditional arguments for expressivism. Most notably, it fits well with the idea that a modified version of G. E. Moore's Open Question Argument (henceforth the 'OQA') is one of the most important motivations for expressivism. Moore himself of course used the OQA to defend his own distinctive brand of non-naturalist cognitivism. However, generations of expressivists have argued that their own view is the real beneficiary of the argument. The basic point of the argument is that the expressivist is best placed to explain why it seems so plausible to competent speakers that

[7] Paul Edwards and David Wiggins have defended such views. See especially Edwards (1955).

[8] Elsewhere I defend a form of Ecumenical Cognitivism about rationality, so in a sense I too take a hybrid view of a sort.

for any proffered descriptive analysis *D* of a normative concept *N*, it will be possible for a speaker without conceptual confusion to admit that an action is *D* but still wonder whether it is *N*. Expressivism as glossed above should still be able to reap this dialectical dividend. For even if normative utterances do express beliefs, as the Ecumenical Expressivist insists, they do not express beliefs which are such that the utterance is semantically guaranteed to be true just in case the belief is true. Which is just to say that any representation of the world as being a certain way does not entail any particular normative stance. Much more would need to be said to make the OQA even halfway plausible, given the wide battery of objections lodged against it. However, this is not the place for such a defense. The point here is simply that, for better or worse, the OQA is an important motivation for expressivism.

On this revised way of understanding the debate between cognitivists and expressivists, there will be both cognitivist and expressivist forms of the Ecumenical View. The cognitivist version has in effect already been explored by Daniel Boisvert, David Copp, Matthew Millar, Jon Tresan, and others.[9] On at least one reading, James Dreier's speaker relativist theory is a species of Ecumenical Cognitivism.[10] By contrast, the expressivist version of the Ecumenical View has not been much explored.

2. TWO SPECIES OF THE ECUMENICAL EXPRESSIVIST GENUS

Here I want to explore two versions of Ecumenical Expressivism. The first version is the more simple one, according to which normative utterances express (a) a speaker's approval of actions in general insofar as they have a certain property, and (b) a belief which makes anaphoric reference to that property (the one in virtue of which the speaker approves of actions in general). The approval is *insofar* as the actions in question have the relevant property in the sense that having the property to a greater extent indicates greater approval, all else being equal, anyway. Just what the relevant property is can vary from one speaker to the next. I might approve of actions insofar as they promote happiness, while you might approve of actions insofar as they are in accordance with God's will. Indeed, it is the possibility of this sort of variability from one speaker to the next that allows Ecumenical Expressivism to accommodate the semantic intuitions which give Moorean OQAs whatever force they have.

[9] See Boisvert (2005), Copp (2001), Millar (2005), and Tresan (2006).
[10] See Dreier (1990).

In addition to expressing approval of actions quite generally insofar as they have a certain property, normative utterances on this account also express beliefs. The beliefs expressed make anaphoric reference back to the property in virtue of which the speaker approves of actions. The content of the belief expressed by any given normative utterance on this account can be given by a formula which takes the original sentence and replaces all uses of normative predicates with suitable anaphoric reference to 'that property' where 'that property' denotes the property of actions in virtue of which the speaker approves of actions quite generally. More schematically, on this account any given normative utterance expresses:

1. A suitable state of approval to actions insofar as they have a certain property.

and

2. A belief which makes suitable anaphoric reference back to that property.

For lack of a better name, call this first version of Ecumenical Expressivism the 'Plain Vanilla' version of the view. This will allow us to distinguish it from a slightly fancier Ideal Advisor version of Ecumenical Expressivism introduced below.

I say 'suitable state of approval' in (1) to mark the fact that on any plausible form of expressivism not just any old whim or urge will count as a normative judgment. Gibbard, for example, distinguishes the state of 'norm acceptance' from other sorts of pro-attitudes. I here remain officially neutral on what exactly the best candidate sort of pro-attitude is. In this respect, my proposal remains schematic, though I shall say more about this below (in section 6).

An example may help clarify the proposal. On the Plain Vanilla account, an utterance of 'If passive euthanasia is sometimes right then active euthanasia is sometimes right' expresses (1) a state of approval of actions insofar as they have a certain property *and* (2) the belief that if passive euthanasia sometimes has that property then active euthanasia sometimes has it too. Note that the utterance expresses a perfectly general pro-attitude to actions insofar as they have a certain property and a belief which refers to that property. Note too that the belief refers to the property in virtue of which the speaker approves of actions. This property will typically be the sort of thing that speakers take to make actions right, like reducing suffering or being approved of by God, or whatever. The belief does *not* refer to the property of being approved of by the speaker, unless the speaker holds a very odd normative outlook according to which something is worthwhile just in case she approves of it, thereby holding a fairly odd and self-referential pro-attitude. I take it that this is not the standard case. The crucial point is

that the anaphoric reference is to the content of the speaker's pro-attitude (the one expressed); it does not in the standard case make reference to the having of that pro-attitude itself.

A given speaker may well not know precisely in virtue of what property she approves of actions. A very rough and imperfect analogue might be the now idiomatic use of the sentences of the form, 'There's something about so-and-so,' which function to express a certain sort of attitude to the person in question without any commitment on the part of the speaker to being able to say just what the property is. Notice that such uses can be embedded in logically complex and unasserted contexts. For example, I can say, 'There's something about Mary, and, whatever it is, Lisa used to have it too.'[11]

In addition to the Plain Vanilla version, I also want to propose an Ideal Advisor version of Ecumenical Expressivism. Indeed, for reasons I cannot go into here, I actually think this version of the theory is much more promising than the Plain Vanilla version. On the Ideal Advisor version of the theory, normative utterances express:

1. A suitable state of approval to actions insofar as they would be approved of by a certain sort of advisor.

and

2. A belief which makes suitable anaphoric reference back to that that sort of advisor.

Like the Plain Vanilla version, this version of the theory also makes heavy use of anaphora. It might here be helpful to pause and work out the semantics for a particular normative term within the framework of the Ideal Advisor version of the theory.

Consider 'must' as it is used to indicate a sort of deontic necessity. Deontic necessities can be moral or non-moral. The sentence, 'one must not kill for profit' is typically taken to express a moral deontic necessity, while a sentence like 'You simply must try the risotto' more typically is taken to indicate a non-moral deontic necessity. In both cases, though, there is some sense in which the action in question is presented as required, whether the sort of requirement at issue is moral or non-moral. It is in this broad sense that both such uses of 'must' indicate a deontic necessity. Deontic necessity should, of course, be distinguished from metaphysical, conceptual, and logical necessity.

How would an Ideal Advisor version of Ecumenical Expressivism handle the semantics of 'must' as that term is used to introduce a deontic necessity?

[11] My apologies for the sexism of the example, but this familiar if sexist mode of discourse does provide a good structural model for the semantics I am developing.

The semantics would provide a recipe from any given sentence in which 'must' is used in this sense to an account of the states of mind expressed. Presumably, an ideal advisor would not merely recommend or suggest that one perform an action which is required. Instead, such an advisor would *insist* on one's performing the action. Insistence here should be understood as a non-normative concept, though. Otherwise we would need to give a further expressivist account of insistence, and we would be off on a regress. Fortunately, insistence does have a purely descriptive meaning as well as a richer normative meaning (the latter sense of 'insist' should indeed be understood in expressivist terms). In the sense in play here, to insist on something is to issue an imperative that the action be performed, and to do so in a way which is emphatic and conveys the idea that the speaker 'won't take no for an answer,' and that a failure to comply with the imperative will lead to a negative attitude by the speaker (anger, or perhaps disappointment or even pity).

How, though, should we generalize the semantics of the ideal advisor approach to cover more complex sentences in which 'must' in the deontic sense appears in unasserted contexts, such as the antecedent of a conditional? Here is one way in which this idea could be worked out in more detail, just to give a sense of how the approach could be made more concrete; I do not necessarily mean to endorse this particular way of working out the theory, though.[12] Take any sentence '*p*' in which 'must' is used in a deontic sense. The utterance of '*p*' expresses approval of a certain sort of advisor and the belief that *p**, where *p** is the content you get when you take p and replace all occurrences of '*A* must Φ' with '*A* is/are such that such an advisor would be disposed to insist that *A* Φ, where 'such an advisor' makes anaphoric reference to the sort of advisor the approval of which the speaker's utterance expresses. Deontic uses of 'must' which do not appear in the form '*A* must Φ' are taken to be elliptical for something which is more properly put in this form. Compare the way in which 'the jeans don't fit' is

[12] I develop a more detailed and considered account elsewhere in a book-length treatment of Ecumenical Expressivism I am currently writing, provisionally entitled *Impassioned Belief*. There I argue that we should allow for the possibility that someone might approve of a variety of different sorts of advisors as ideal, and that this might constitute a sort of interesting normative pluralism. Fortunately, the semantics for normative terms developed here can rather smoothly make room for the possibility of such views. The basic idea is to understand normative utterances as expressing approval of a suitable *set* of advisor types, where the set may or may not have more than one member. The belief expressed then will make suitable anaphoric reference back to the members of that set. A full presentation of the details of this more complicated version of the theory would go beyond the present scope, though, and would distract from the more general advantages of Ecumenical Expressivism. For expository reasons, I therefore work with a somewhat oversimplified version of the theory in the text here.

elliptical for the claim that the jeans don't fit someone or other, depending on the context. Putting such elliptical uses to one side, though, consider the following specific example. Take the sentence, 'If Robin made it, then you simply must try the risotto.' On the proposed semantics, an utterance of this sentence expresses (a) approval of a certain sort of advisor, and (b) the belief that if Robin made it then such an advisor would be disposed to insist that you try it (the risotto).

This should be enough to get across the basic semantics for Ecumenical Expressivism, both in its Plain Vanilla guise and in its (more plausible, in my view) Ideal Advisor guise. However, before we can usefully discuss the dialectical advantages of Ecumenical Expressivism over its Non-Ecumenical rivals, we must first consider one more important divide within the Ecumenical Expressivist camp.

3. TRUTH AND ANOTHER DISTINCTION

According to Ecumenical Expressivism, a normative sentence is not semantically guaranteed to be true if and only if the belief it expresses is true. Here we come to yet another divide, for this failure to provide truth-conditions can itself be understood on either of the following two models. First, following in the tradition of A. J. Ayer and Bertrand Russell, we could hold that normative sentences simply are not truth-apt and so trivially are not true just in case the belief expressed is true. We might call this version of expressivism 'cave man' expressivism, to mark its association with the very early history of the doctrine.

The cave man approach is straightforward, but it does force us into the uncomfortable position of claiming that a great deal of ordinary discourse in which we classify normative sentences as true or false is deeply confused. A second version of expressivism tries to avoid this counterintuitive consequence. On this approach, normative sentences are truth-apt but are not semantically guaranteed to be true just in case the belief they express is true. This position is much more delicate, but if it can be made to work then it also holds out the promise of a much more plausible view. For such a view would not force us to abandon nearly as much of ordinary discourse as the maverick views of Air and Russell. This approach is associated with what Simon Blackburn has famously called 'quasi-realism'—the attempt to show how realist sounding discourse can be made intelligible within what ultimately is an expressivist framework.[13]

[13] Actually, I prefer the label 'quasi-descriptivism' since the 'realism' in 'quasi-realism' suggests that the metaethical view on offer is committed to construing normative talk

One of the more promising versions of the quasi-realist strategy invokes the following two doctrines about truth. First, it invokes Deflationism about truth, which *very* roughly holds that to say that '*p*' is true is no different from saying *p*. This is a very rough first approximation indeed, as it tells us nothing about the more interesting uses of 'true' in various forms of indirect discourse (e.g. 'Everything he says is true,' 'the third sentence on the page is true,' etc.). Second, this strategy invokes Deflationism about truth-aptness, according to which there is nothing more to being apt for truth than being such that sentences of the form '"*p*" is true' are well formed. The basic idea is then to hold that normative sentences are trivially truth-apt, but insist that to call a given normative sentence true is really to do nothing more than reiterate it. On the expressivist view, to reiterate a normative sentence is in part to give voice to a suitable non-cognitive attitude. This in turn means that I can agree with the content of the belief expressed by your normative utterance without being forced, on pain of inconsistency, to admit the truth of what you said. For I can admit the truth of the belief you have expressed but refuse to share your non-cognitive attitude.

Actually, things are a bit more complicated than this. For we must now distinguish two senses of 'belief'. In the first sense of 'belief', beliefs have a representational direction of fit—they aim to fit the world. Beliefs in this sense also stand in inferential relations and have various other features which distinguish them both from desire-like states and from other representational states (e.g. perceptions). Moreover, in this sense, beliefs are a natural kind that will figure in a mature theory of human psychology. Filling out this sense of 'belief' in more detail or vindicating the hypothesis that there are beliefs in this sense would take us too far afield. It is in this sense of 'belief' that I can acknowledge that the belief you expressed is true but deny the truth of your normative utterance. Crucially, in this strict sense of 'belief' there are no normative beliefs. In this sense of 'belief' so-called normative beliefs are really just belief/desire pairs.

However, we can allow a wider notion of belief which includes normative beliefs as well, though this notion of belief will not pick out a natural kind. Beliefs in this sense will include beliefs in the strict sense as well as beliefs qua normative beliefs as suitable belief/desire pairs. The basic idea is that 'belief' in this sense refers to whatever causally regulates our actual use of

in realist terms. Realism, though, is often understood in terms of a kind of mind-independence, according to which the truth in a given area of discourse can outrun even our best judgments in that area. As I understand Blackburn's view, though, whether morality should be construed in realist terms in this sense is a first-order question. To take a realist view is to adopt one sort of set of attitudes, whereas to take an anti-realist view is to adopt another rather different sort of set of attitudes. However, I will not fight over the word here.

'belief', and given Ecumenical Expressivism this will include certain belief (in the natural kind sense)/desire pairs. In this sense of 'belief', I cannot admit the truth of the (normative) belief expressed by your utterance and at the same time deny the truth of what you have said. I cannot, for example, admit that your belief that abortion is wrong is true but deny that your utterance of 'abortion is wrong is true'. The upshot is that our definition of expressivism and cognitivism must be understood as working with 'belief' in the strict sense. Throughout the rest of this paper, unless I explicitly note otherwise, I shall be using 'belief' in the strict sense.

Working through an example should help illustrate the details of the proposed account. Suppose that you say that it was right to divert the trolley. According to the Ideal Advisor version of Ecumenical Expressivism, this will amount to your expressing a pro-attitude to actions insofar as they would be approved of by a certain sort of advisor and the belief that such an advisor would be disposed to insist on the diversion of the trolley. Suppose I agree with you that the sort of advisor you have in mind would be disposed to insist on the diversion of the trolley. Perhaps I independently know that you approve of a sort of utilitarian saint as an ideal advisor, and I believe that diverting the trolley will maximize the total amount of happiness in the world. Does this admission also force me to admit that what you said is true? No. For to admit that what you said is true I must also share your approval of actions insofar as they would be approved of by a sort of advisor who would insist on the diversion of the trolley. I may simply not take any such attitude, in which case my recognition of the truth of the belief you have expressed does not force me, on pain of inconsistency, to admit that what you have said is true. Of course, if I agree that your belief that diverting the trolley is right is true then I cannot deny that what you have said is true. In that case, though, I must in some sense either have or be committed to having a suitable attitude to a sort of advisor who would be disposed to insist on trolley diversions in such cases. The adoption of such an attitude will not be forced on me by the mere recognition that you approve of such an advisor.

The deflationist approach is promising when viewed in large frame, but the devil is in the details. In particular, giving a plausible deflationist account of how the truth predicate works in indirect discourse that is compatible with expressivism is extremely tricky. Expressivists cannot simply take over without modification some of the leading deflationist accounts in the literature. Paul Horwich's interesting account, for example, crucially insists that the truth predicate applies in the first instance to propositions rather than sentences. This would sit very poorly with the expressivist idea that ultimately there are no normative facts, and hence no normative propositions. Of course, a quasi-realist expressivist may well (and Blackburn

has) go on to give a deflationist sense of 'proposition' and try to show how a quasi-realist can 'earn the right' to talk about normative propositions and facts as well. Even if this can be made to work, however, the notion of proposition invoked will not be the more substantial one Horwich has in mind.

Although I think the issues about how the expressivist should handle truth are extremely important, they would in the present context very quickly take us too far afield from what is distinctive about Ecumenical Expressivism. To put my cards on the table, I think that Ecumenical Expressivism should be developed in a way that accommodates the truth-aptness of normative discourse, but that this accommodation should not be contingent on a thoroughgoing deflationism about truth.[14] Naturally, much more would need to be said about this approach even to convey the basic ideas behind it, much less to defend its plausibility. However, this is a very long story, and requires another paper altogether. Therefore, for present purposes I must put these very thorny issues firmly to one side. To simplify matters, let us just assume that we are here trying to see what advantages a 'cave man' version of Ecumenical Expressivism would have, even though I ultimately prefer to defend a more quasi-realist version of the view. So we can here put the challenge of accommodating the truth-aptness of normative discourse to one side, albeit with the understanding that this issue must eventually be revisited. I now discuss three advantages of Ecumenical Expressivism in turn.

4. FIRST ADVANTAGE: AVOIDING THE FREGE–GEACH PROBLEM[15]

Old fashioned 'boo-hooray' forms of expressivism tell us nothing about utterances in which normative predicates are used in unasserted contexts, such as 'If lying is wrong then getting little brother to lie is wrong'. To this extent they are incomplete. Moreover, one otherwise tempting strategy for completing the expressivist theory runs into another problem. For one might hold that while normative predicates function to express non-cognitive attitudes in asserted contexts, they serve a very different function and have a different meaning in unasserted contexts. The obvious problem with this approach is that if normative predicates do not express attitudes

[14] I develop this approach at some length in my 'The Truth in Ecumenical Expressivism' (forthcoming) and also in *Impassioned Belief*.

[15] In this section I draw heavily on Ridge (2006), where my proposed solution to the Frege–Geach problem is developed in more detail.

when they occur in unasserted contexts then apparently valid arguments turn out to commit the fallacy of equivocation. Consider the following famous toy argument:

1. Lying is wrong.
2. If lying is wrong then getting little brother to lie is wrong.

So,

3. Getting little brother to lie is wrong.

If the meaning of 'wrong' in (1) is cashed out in terms of the expression of an attitude, whereas the meaning of 'wrong' in (2) is not understood in these terms, then the meaning of 'wrong' shifts from (1) to (2). In that case, though, the argument commits the fallacy of equivocation, and is invalid. This, though, is implausible. This argument certainly need not be fallacious or invalid. The expressivist needs to account for the meaning of normative terms in such a way that they are not systematically ambiguous between asserted and unasserted contexts. This, in brief, is the 'Frege–Geach' problem for expressivism. The problem was first noted by P. T. Geach, and his characterization of the problem drew heavily on an analogous problem pressed by Frege for certain theories of negation.[16]

There are various standard expressivist strategies for dealing with the Frege–Geach problem, but these strategies all have numerous problems which have been discussed at some length in the literature. Here I want to explore a rather different strategy which becomes available once we make 'the ecumenical turn'.

According to Ecumenical Expressivism, normative utterances express both beliefs and desires. Since we are here dealing with some of the belief-like features of normative judgment, an obvious strategy is to let the beliefs do the lion's share of the work in meeting the Frege–Geach challenge. First, consider why the charge of incompleteness—that expressivism has nothing to say about uses of normative predicates in unasserted contexts such as the antecedent of a conditional—does not threaten Ecumenical Expressivism.

Ecumenical expressivism gives a systematic and unified semantics for both asserted and unasserted uses of normative predicates. According to Ecumenical Expressivism, for any declarative sentence *p* in which 'required' is used, an utterance of *p* expresses (a) an attitude of approval to all and only actions insofar as they would be approved of by a certain sort of advisor, and (b) the belief that *q*, where *q* is what you get when you take *p* and replace all occurrences of 'required' with 'such that it would be insisted on

[16] See Geach (1965).

by such an advisor'. This account is perfectly general and applies across the board to both asserted and unasserted contexts.[17]

So Ecumenical Expressivism avoids the charge of incompleteness. One might still worry, though, that it cannot preserve logical validity. It is perhaps

[17] Actually, one very important qualification must be added, which raises issues I must put to one side here. For the general account of the meaning of normative predicates laid out in the text does not plausibly extend to contexts in which normative predicates figure in the contents of a propositional attitude attributed to someone (e.g. when I say, 'She believes that abortion is wrong.'). The point is that in these contexts we are not typically assuming that the person to whom we attribute the propositional attitude associates the same cluster of descriptive properties with a given normative predicate that we do. For example, when I, as a utilitarian, say that Jones believes that abortion is wrong and need not be presuming that Jones believes that abortion fails to maximize happiness. So we should understand such attributions in terms of the attribution of a suitable belief/desire pair without taking a position on whether the speaker shares our conception of the good. So when I say that she believes that abortion is wrong I am making a purely descriptive claim, namely that she has the belief that abortion is wrong. It turns out (though a given speaker may not realize this, of course) that the belief that abortion is wrong is really in one sense a belief/desire pair—a general pro-attitude of the right kind and a belief which makes suitable anaphoric reference back to the content of that pro-attitude.

In itself this need not pose any special problems. The meaning of normative predicates in propositional attitude ascriptions is connected in obvious and systematic ways to their meanings in other contexts. Moreover, there need be no special problem about the validity of arguments employing such ascriptions as premisses, since we cannot in general draw any inferences about the contents of such ascriptions from the ascription itself (apart from the fact that someone believes or desires that content, and that inference is valid on the account developed here).

An instructive analogy is with pejorative terms (also discussed in Copp (2001)). Plausibly, to call someone a 'nigger' is at least in part to express contempt toward certain people in virtue of their race. However, intentional attitude ascriptions need not involve any such expression of contempt. For example, someone who sincerely says, 'David Duke just thinks of me as a nigger,' certainly does not thereby express contempt for people in virtue of their race. Instead he ascribes to Duke an attitude of contempt and a belief that he (the speaker) has the features to which this contempt is cued. Here we have a nice parallel with the account developed here, for on the Ecumenical Expressivist account we should also say that such contexts involve the ascription of a suitable attitude/belief pair. Moreover, this shift in expressive meaning (in the case of pejoratives) from intentional attitude ascriptions to other contexts seems to create no insuperable problems in this context, and this point should be common ground. So if there are general problems lurking here then they are problems for everyone and not just the expressivist.

The only real difficulties emerging for Ecumenical Expressivism on this front arise when we combine ascriptions of normative beliefs with claims about the truth of what the subject believes, which should allow us to infer a normative conclusion. For example, we have inferences like, 'She thinks abortion is wrong, and everything she thinks is true, so abortion is wrong.' However, I shall not here go into the details of how Ecumenical Expressivism is best extended to deal with these further cases. For this would require a full theory of truth (for a start) and would therefore take us too far afield from an outline of the basic ideas and advantages of the ecumenical approach. I explore these issues in my 'The Truth in Ecumenical Expressivism.' Thanks to Timothy Williamson and John Hawthorne for pressing me on this point.

no surprise that here the fact that normative utterances express beliefs as well as desires does some real work for us. For the ability of normative judgments to figure in logically valid inference is, after all, a belief-like feature of normative judgment par excellence. However, we will still need a suitable account of logical validity. Since we are here working with a 'cave-man' form of Ecumenical Expressivism, we cannot define validity in a standard truth-conditional way. However, there is a close cousin of the truth-conditional conception which will work. Let us define validity as follows:

> An argument is logically valid just in case it is such that, necessarily, anyone who accepts the premisses and at one and the same time denies the conclusion is thereby guaranteed to have contradictory beliefs.

The definition should be extensionally equivalent with truth-conditional accounts in the context of arguments with purely descriptive premisses and conclusions.[18] Hence it marks a less radical departure from standard views of logical validity than the usual expressivist stories. For these more traditional accounts are usually cast in terms of a so-called 'logic of attitudes' which defines logical validity in terms of the avoidance of having a 'fractured sensibility'. The availability of a less radically reversionary account of logical validity therefore already highlights one advantage of Ecumenical Expressivism, given the notorious difficulties associated with the logic of attitudes approach.[19]

More to the point, this account of logical validity provides Ecumenical Expressivism with an easy and straightforward explanation of the validity of arguments in which normative predicates are used in unasserted contexts. Begin with the simplest form of argument, reiteration—'*p*, therefore *p*'. Let '*p*' be an atomic normative utterance such as 'Charity is required'. On the proposed conception of validity, the argument is valid just in case any agent who accepts the premiss but denies the conclusion would thereby be guaranteed to be caught in an inconsistency. Since the denial of the conclusion would simply be 'Charity is not required' the question is whether anyone who accepts (C) 'Charity is required' and who accepts (not-C) 'Charity is not required' is thereby caught in an inconsistency. On the model proposed here, any possible agent who accepts (C) and accepts (not-C) both believes that charity is such that the relevant sort of advisor

[18] The only slight complication here is what to say about sentences which employ conventional implicature words like 'but' and 'even'. I have dealt with these issues at length elsewhere (in Ridge 2006: 327–8), though, and must for present purposes put them to one side.

[19] See Dorr (2002), Hale (1986), Hale (1993), and van Roojen (1996). I explore the difficulties raised by these authors for non-ecumenical approaches in more detail in Ridge (2006).

would be disposed to insist upon it, and at the same time believes that charity is not such that the relevant sort of advisor would be disposed to insist upon it. This clearly is an inconsistency of a familiar kind—inconsistency in belief. So the argument is valid on the proposed account.

It is straightforward to see how this account can be extended to deal with other logically complex sentences. The general scheme for any logically complex sentence in which 'required' appears is as follows. Let '*p*' stand for a logically complex sentence in which 'required' is used. An utterance of '*p*' expresses (a) the agent's approval of actions insofar as they would be approved of by of a certain sort of advisor, and (b) the agent's belief that p^*, where p^* is identical to *p* save that all occurrences of 'is required' are replaced by 'is such that such an advisor would insist on it', where 'such an advisor' makes anaphoric reference back to the sort of advisor approval of which was voiced in (a).

It should be clear by now how this account can explain the validity of arguments with normative predicates quite generally. Consider the standard case of modus ponens:

1. Telling the truth is required.
2. If telling the truth is required, then not getting your little brother to lie is required.
3. Therefore, not getting your little brother to lie is required.

On the proposed account, the acceptance of (1) requires the belief that a certain sort of advisor would be disposed to insist on telling the truth. The acceptance of (2) involves the belief that if such an advisor would be disposed to insist on telling the truth then such an advisor would be disposed to insist on not getting your little brother to lie. To deny (3), though, involves believing that such an advisor would not be disposed to insist on not getting your little brother to lie. This is obviously an inconsistent triad of beliefs. The general strategy works across the board in an elegant way, no matter how complicated the judgments. The Frege–Geach problem simply does not arise.

Before leaving the Frege–Geach problem, it is worth noting that it is crucial to the tenability of the proposed definition of logical validity that it ranges over all possible believers. Suppose I am a utilitarian, so I approve of actions insofar as they maximize utility. In that case it would be contradictory for me to think that an action maximizes utility yet is not morally right. However, the inference, '*X* maximizes utility, therefore *X* is morally right' had better not be valid, on pain of contradicting the very intuitions which underlie the OQA. Fortunately, on the account offered here this argument is invalid. For while it is true that a utilitarian who believes both that an action maximizes utility and that the action is not

morally right is thereby caught in an inconsistency, it is *not* true that any possible believer who believes that an action maximizes utility and that the action is not morally right is thereby guaranteed to be caught in an inconsistency. Anyone not committed to utilitarianism can accept the premiss and reject the conclusion without inconsistency. So the inference is not valid on the proposed definition of validity. This is why it is crucial that validity is defined in terms of whether *anybody* who accepted the premisses and at one and the same time denied the conclusion would thereby be caught in a contradiction.

5. SECOND ADVANTAGE: AKRASIA

Let us understand akrasia as someone S's judging that she ought to *X* but failing to *X* (or even to intend to *X*) when she knows she could. So analyzed, akrasia seems both possible and irrational. A plausible analysis of normative thought and discourse should explain its possibility and irrationality.

Standard forms of cognitivism notoriously have trouble explaining how a representation of the world as being a certain way can make it irrational not to act in one way rather than another. Unless we abandon a broadly Humean philosophy of mind, the cognitivist will have to tell a very special story about the contents of normative judgments in order to explain how such judgments rationally commit a speaker to a course of action. This is not to say that such special stories have not been told, but it does at least present a prima facie challenge to cognitivists. I say this is a problem for *standard* forms of cognitivism because taking the 'Ecumenical Turn' can help cognitivists here too, a point to which I return below.

At least some versions of Non-Ecumenical Expressivism also have trouble making good sense of the idea that failure to be motivated by one's normative judgment involves irrationality. Here I put to one side those versions of Non-Ecumenical Expressivism which understand normative judgments in terms of higher-order attitudes (e.g. accounts inspired by Harry Frankfurt's work). These accounts are in a better position to deal with this particular set of issues.[20] Another important strand of expressivist thought understands normative judgments about a given action as an occurrent pro-attitude in favor of that very action. For example, on A. J. Ayer's classic account, a speaker's judgment that a particular instance of stealing was wrong just is that speaker's having certain occurrent feelings of disapproval of that very action.[21] On these sorts of accounts, whenever someone judges that she

[20] Thanks to Joshua Gert for useful discussion here.
[21] See Ayer 1953.

ought to *X* in *C*, she is on these accounts thereby guaranteed to have at least some motivation right then and there to *X* in *C*. Such Non-Ecumenical Expressivists cannot therefore explain akrasia in terms of a simple absence motivation. Instead, it seems that they must explain akrasia in terms of the presence of conflicting motivations, and this is indeed how expressivists typically have explained akrasia insofar as they allow that it is possible at all.[22]

This approach is problematic for at least two reasons. First, it seems ad hoc to posit a conflicting motivation in every possible instance of akrasia. In some cases it is more plausible to suppose that the person simply lacks suitable motivation to do what she nonetheless believes she should. People who are clinically depressed or listless come to mind.

Second, the invocation of conflicting motivations has difficulty making sense of the idea that akrasia is both possible *and* irrational. On the one hand, if the conflicting desires were really stronger than the desire which constitutes one's normative judgment then it begins to look mysterious why it would be irrational to act on the stronger desire. While acting on one's strongest desire can be irrational, it certainly need not be. In at least some cases, the mere fact that an agent wants *A* more than *B* can be enough to make her action both intelligible and rational. If, on the other hand, the conflicting desires are not motivationally stronger then it becomes mysterious how they could explain your acting on them as opposed to your motivationally stronger normative judgment.

Ecumenical Expressivism has more resources with which to explain the possibility and irrationality of akrasia. Most importantly, because of its insistence that normative judgment involves both belief and pro-attitude, Ecumenical Expressivism allows for a sort of division of labour. My normative judgment is constituted by a perfectly general pro-attitude to actions that would be approved of by a certain sort of advisor and a belief which makes anaphoric reference to such an advisor. One way of unpacking this division of labor is to say that the pro-attitude functions as my normative conception (my conception of what it is to be required, say) and the belief functions as the application of that conception to the world. This division of labor is the key to the Ecumenical Expressivist's account of akrasia.

Pre-theoretically, it is very plausible to suppose that one's very general pro-attitudes can fail to transfer motivational 'oomph' to the particular situation one faces. For example, I might have the general aim of getting some work done today. In spite of this general aim, I might just sit around

[22] In recent work, Gibbard instead opts for the Socratic option that akrasia in the sense articulated in the text is simply not possible. See Gibbard (2003).

and do nothing all day, in a state of depression, distraction, or listlessness. Indeed, one way of understanding depression and listlessness is as preventing one's standing intentions or plans from issuing in the sorts of proximate intentions which are causally efficacious in actually getting one to act. This failure of transfer of motivation between the general and the particular can therefore explain the possibility of akrasia without the ad hoc assumption of conflicting motivation. For my general pro-attitude in favor of acting a certain sort of advisor would want me to act can fail to issue in action simply because that general pro-attitude does not transfer its motivational 'oomph' to the situation at hand in the form of a proximate intention due to my depression or listlessness. This story does not require the presence of any conflicting motivation. This, of course, is not to deny that conflicting motivations are not often present and explanatory, but rather explains why in these sorts of cases involving depression and listlessness they need not be present in order for the agent to be akratic. This approach is not available to the Non-Ecumenical Expressivist precisely because on their account there is no suitable division of labor between the general conception and the particular application of that conception. There is instead just one's pro-attitude to the action before one, and that seems enough to ensure at least some motivation to perform that very action.

Why is akrasia as understood by the Ecumenical Expressivist irrational, though? Such a failure of motivational transfer represents a failure to be motivated to take what you believe to be a constitutive means to your more general end while still holding the end. In Kantian terms, it represents the violation of a hypothetical imperative, which is often taken to be the paradigm case of practical irrationality. For this explanation of the irrationality of akrasia to be plausible, though, we must suppose that the pro-attitude which partially constitutes one's normative judgment is an executive state, such as an intention or plan, as opposed to a mere desire or preference. That is, the pro-attitude in question is partly constituted by the agent's being disposed to exert some real willpower in pursuit of its object. For the plausibility of the Kantian idea that it is irrational to violate a hypothetical imperative depends crucially on understanding what it is to make something one's end in terms of committing one's *will* to that end in some important sense.

Interestingly, this is just the sort of account Alan Gibbard has independently developed, arguing that one's normative judgments just are constituted by one's plans. However, Gibbard's account is a form of Non-Ecumenical Expressivism. Hence, Gibbard cannot invoke a division of labor between the pro-attitude and the belief which constitute one's normative judgment to explain akrasia in the way I have suggested. Indeed, Gibbard himself seems to think that akrasia is impossible if one wholeheartedly

believes that one ought to do something.[23] All the same, Gibbard seems right to analyze normative judgments (at least, judgments about what one ought to do, all things considered) in terms of executive states like plans rather than mere preferences.

To be fair, Ecumenical Cognitivists can also profit from this dialectic. For on at least some versions of Ecumenical Cognitivism, a speaker's normative judgment is partly constituted by a suitable general pro-attitude. This more general pro-attitude could be an executive state cued to whatever the cognitivist takes to be the content or character essential to normative judgments as such. For example, an egoist version of Ecumenical Cognitivism could hold that normative judgment is partly constituted by an intention to do whatever is most in the speaker's interest. Such versions of Ecumenical Cognitivism can tell the same sort of story about why a failure to be motivated in the case at hand is irrational as the one told here. If this is right then the advantage expressivists can claim for their view in terms of the action-guiding aspects of normative judgment must be heavily qualified. The Ecumenical Expressivist does have an advantage here, but only over Non-Ecumenical Cognitivists (assuming the prima facie challenge laid out above cannot be met, contra e.g. Michael Smith) and some versions of Non-Ecumenical Expressivism. The real advantage of Ecumenical Expressivism over Ecumenical Cognitivism must, therefore, be found elsewhere. Presumably, the ability of Ecumenical Expressivism to accommodate Moorean 'Open Question' intuitions will be highly germane on this score, but that is a complex set of issues that I must here put to one side.

6. THIRD ADVANTAGE: CERTITUDE, ROBUSTNESS, AND IMPORTANCE

Michael Smith has posed an important but previously unappreciated challenge for expressivists. The challenge is to distinguish certitude, robustness, and importance, and explain how each plays a characteristic role in motivation. Very roughly, my certitude that charity is right is a measure of how certain I am about this thesis. Decision theorists gloss this in terms of how much I would be willing to gamble on its truth. Importance is how strong I take the relevant reason(s) to be—how strong I think the reasons in favor of charity are, for example. Finally, robustness is the stability of my belief in the face of further information and deliberation. The problem for expressivists is well put by Smith:

[23] See Gibbard 2003: 153.

desires possess just two structural features that look like they will be of any use in the present connection. Desires differ from each other in terms of their strength … And the strength of an agent's desires may vary over time under the impact of information and reflection … degree of strength can represent something, presumably either Importance or Certitude, and since the strength of the subjects' desires can vary over time under the impact of information and reflection, that too can presumably represent something, presumably Robustness. But that leaves one thing, either Importance or Certitude, not represented at all. (Smith 2004: 354–5)

Smith poses his challenge in terms of evaluative judgments as to what is good or bad, but the challenge works just as well when posed for normative judgments. The challenge is a good one. Indeed, I think it is a difficult and perhaps impossible challenge for Non-Ecumenical Expressivists to meet precisely because their theories are cast entirely in terms of desires.

Ecumenical Expressivism has more resources. The challenge is to explain how expressivism can distinguish some of the more subtle belief-like features we associate with normative judgment. By holding that normative judgments are partially constituted by beliefs as well as desires, Ecumenical Expressivism is better situated to explain this without giving short shrift to the desire-like features of normative judgments. Moreover, the linkage between the belief and the desire which constitutes one's normative judgment ensures that certitude, robustness, and importance will play the sorts of motivational roles that we ordinarily suppose they do. Here is how Ecumenical Expressivism can meet Smith's challenge:

Certitude

Certitude: An agent's certainty that he should Φ is represented by two factors:

(a) his certainty (in the ordinary sense) that Φ-ing would be approved of by the relevant sort of advisor,

and

(b) the relative strength of his pro-attitude in favour of actions insofar as they would be approved of by the relevant sort of advisor.

It is very tempting for the Ecumenical Expressivist to represent certainty in terms of (a) alone, the certainty the agent has in the relevant belief. For example, if a given agent were a utilitarian, then her certainty that an action is right would correspond directly to her certainty that the action maximizes utility (or her certainty that an act-utilitarian saint would approve of it, which should amount to the same thing if the agent is rational). Prima

facie, this seems like a simple and elegant solution to the problem of how to make sense of certainty in normative judgment.

However, this approach buys its simplicity at too high a cost. Consider the fact that someone can be very sure that a given action has the features which he takes to be reason-providing but not especially sure that these features really are reason-providing. For example, a utilitarian could be very sure that a given action would maximize utility, but be less sure that utilitarianism is true, even though that is her current defeasible view. Uncertainty about whether the action has the relevant properties can plausibly represent the first sort of uncertainty, but not the second.

So we need something like (b) to represent this dimension of certainty/uncertainty in one's fundamental normative conception (e.g. one's commitment to utilitarianism or whatever). Greater certainty along other dimensions will increase the likelihood that the agent will act as he thinks he ought, all else being equal. If I am more certain in my belief that an action has features I desire to instantiate then I shall be more likely to perform the action, all else being equal. Furthermore, if my commitment to instantiate a given sort of action is stronger then I shall be more likely to perform the action, all else being equal. So the proposed account meets Smith's challenge of showing how the account of certainty on offer fits with an intuitive view of how certainty in normative judgment plays a role in an agent's motivation.

To be clear, the account on offer is cast in terms of *relative* motivational strength for a reason. For if we instead glossed certainty in one's fundamental normative conception in terms of absolute motivational strength then perfectly general motivational maladies (depression and listlessness, say) would count as undermining an agent's certainty in all of her normative judgments, but that is implausible. Whereas if we understand certainty in one's normative judgment in terms of relative motivational strength then this counterintuitive result does not follow.

Robustness

This is perfectly straightforward. Since Ecumenical Expressivism offers an account of what it is to think one should Φ then it is obvious what it should say about the robustness of that judgment:

Robustness: The robustness of an agent's judgment that he should Φ just is the stability, in the face of new information and deliberation, of his being such that he approves of actions insofar as they would be approved of by a sort of advisor whom he at the same time believes would approve of his Φ-ing.

Clearly, this will track motivation in the way Smith suggests it should. If my judgment is not very robust then it shall shift easily in the face of further information and deliberation. Since my judgment so analyzed constitutes a motivation to Φ, this explains how less robust judgments are less reliable and providing motivation over time in the face of further information and deliberation.

Importance

Importance: How much reason an agent takes there to be in favor of Φ-ing is represented by how much the agent thinks the relevant sort of advisor would want him to Φ.

In a way, this is also the obvious move for an Ecumenical Expressivist to make. Here the idea of an ideal advisor component of the analysis does some real work. For we can understand the strength of a reason in terms of the motivational strength of a suitable advisor without thereby understanding the reason itself as the fact that such an advisor would want one to perform the action. Instead, it is much more plausible to hold that the reason an agent takes there to be for a given action is a fact that the action has some feature *F*. Which feature *F*? That feature *F* such that, by the agent's lights, the sort of advisor of whom he approves would approve of the action (to some extent) in virtue of its being *F*. Crucially the reason on this account is not that such an advisor would approve of the action, but some such fact as the fact that it would promote pleasure, for example. The point is that the agent takes this fact (that it would promote pleasure) as a reason only in virtue of approving of a sort of advisor who approves of the action to some extent in virtue of its promoting pleasure.

This conception of reasons allows Ecumenical Expressivism to accommodate the intelligibility of pluralism about the fundamental kinds of reasons there are without running into rampant incommensurability. For the strength of the preference of the ideal advisor provides a sort of common coin by which we can measure the strength of various reasons without thereby being driven into a monistic conception of reasons themselves. This, in itself, is an important advantage of the proposed account.

Moreover, it should be clear enough that on this account of importance the stronger an agent takes a reason for a given action to be, the more strongly motivated she shall be to perform the action, all else being equal. For if I desire to act as a certain sort of advisor would want me to act and I believe that such an advisor would prefer that I Φ rather than Ψ then, all else being equal, I should prefer Φ-ing to Ψ-ing. To be clear,

this works quite generally only if we understand an agent's pro-attitude as giving a ranking of options which matches the relevant sort of advisor's ranking. So my aim to act as such an advisor would want me to should not simply be understood as a desire to perform the action that such an advisor most wants me to perform. It should instead be understood as a ranking of options, with the advisor's top choice as the option I aim for as my top preference, the advisor's second top choice as the option I aim for as my second preference, and so on. So long as we understand normative judgment in just this way, though, there should be no problem explaining how importance as analyzed here can play just the motivational role we pre-theoretically associate with importance.

CONCLUSION

Normative judgment is indeed Janus-faced, and our best theories of normative judgment and discourse should reflect this. It is for this reason that I have argued that metanormative theory should take 'the Ecumenical Turn.' On the Ecumenical View, normative utterances express both beliefs and desires and normative judgment is itself constituted by both beliefs and desires. This naturally leads to a reconceptualization of the debate between cognitivists and expressivists according to which the real sticking point between the partisans of those views is no longer whether normative utterances express beliefs at all, but rather whether the beliefs expressed are such that the utterance is true just in case the belief is true. I have outlined an ideal advisor version of Ecumenical Expressivism and argued that this account has the main advantages of Non-Ecumenical Expressivism while at the same time having at least three substantial advantages over its non-ecumenical rivals.

First, Ecumenical Expressivism avoids the Frege–Geach problem without any of the problems associated with Non-Ecumenical Expressivist's solutions to that problem. Second, Ecumenical Expressivism can more plausibly account for both the possibility and irrationality of akrasia than Non-Ecumenical Cognitivism and at least some versions of Non-Ecumenical Expressivism. However, Ecumenical Cognitivists and those Non-Ecumenical Expressivists who analyze normative judgments in terms of higher-order attitudes may be equally well suited to accommodate these phenomena. So the scope of this second advantage is limited. Third, Ecumenical Expressivism can more plausibly distinguish certitude, robustness, and importance in our normative judgments than Non-Ecumenical Expressivism. On the whole, then, there seems to be a strong case for the following

conditional conclusion: *If* you are going to be an expressivist of any kind then you ought to be an Ecumenical Expressivist.[24]

REFERENCES

Arpaly, N. (2000) 'On Acting Rationally against One's Best Judgment.' *Ethics* 110/2: 488–513.

Ayer, A. J. (1953) 'A Critique of Ethics,' *Language, Truth, and Logic* (New York: Dover).

Boisvert, D. (2005) 'Expressive Assertivism and the Embedding Objection.' Unpublished Manuscript. **www.csub.edu/~dboisvert/Online_Papers/EA_and_The_Embedding_Objection-Long_Version.pdf**

Brink, D. (1997) 'Moral Motivation,' *Ethics* 108: 4–32.

Copp, D. (2001) 'Realist-Expressivism: A Neglected Option for Moral Realism,' *Social Philosophy and Policy* 18: 1–43.

Dorr, C. (2002) 'Non-cognitivism and Wishful Thinking.' *Nous* 33/4: 558–72.

Dreier, J. (1990) 'Internalism and Speaker-Relativism.' *Ethics* 101/1: 6–26.

Edwards, P. (1955) *The Logic of Moral Discourse* (Glencoe, Ill.: Free Press).

Geach, P. T. (1965) 'Assertion,' *Philosophical Review* 74: 449–65.

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

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

Hale, B. (1986) 'The Compleat Projectivist.' *Philosophical Quarterly* 36: 65–84.

____ (1993) 'Can There Be a Logic of Attitudes?' in J. Haldane and C. Wright (eds.), Reality, Representation and Projection (New York: Oxford University Press).

Horgan, M., and Timmons, M. (1992) ''Troubles for New Wave Moral Semantics: The ''Open Question Argument'' Revived'', *Philosophical Papers* 21: 153–75.

Jackson, F., and Pettit, P. (1998) 'A Problem for Expressivists,' *Analysis* 58/4: 239–51.

Korsgaard, C. (1996) *The Sources of Normativity* (Cambridge: Cambridge University Press).

Lance, M., and O'Leary-Hawthorne, J. (1997) *The Grammar of Meaning* (Cambridge: Cambridge University Press).

Millar, M. (2005) *Making Moral Judgements: Internalism and Moral Motivation*. Ph.D. Thesis: University of Edinburgh.

Ridge, M. (2006) 'Ecumenical Expressivism: Finessing Frege,' *Ethics* 116: 302–36.

____ (forthcoming) 'The Truth in Ecumenical Expressivism.'

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

[24] Thanks to the participants of the Second Annual Madison Metaethics Workshop and the audiences at Oxford, Glasgow, York, and Sheffield at which an earlier version of this paper was given for useful comments and suggestions. Thanks also to Matthew Chrisman, David Copp, Joshua Gert and Geoffrey Sayre-McCord for helpful correspondences on earlier versions of the material discussed here.

Smith, M. (2004) 'Evaluation, Uncertainty, and Motivation,' in his *Ethics and the A Priori* (Cambridge: Cambridge University Press).
Tresan, J. (2006) 'De Dicto Internalist Cognitivism,' *Nous* 40: 143–65.
van Roojen, M. (1996) 'Expressivism and Irrationality,' *Philosophical Review* 105: 311–35.

4

Cognitivism, Expressivism, and Agreement in Response

Joshua Gert

> 'But it seems to me that if there is even *one* ethical use such as Mr. Stevenson holds that there is, then probably *all* ethical uses are like it'
>
> Moore (1968: 551)

Hume believed that value claims were distinct from claims about, for example, the sizes and shapes of objects, in that the latter were the deliverances of reason, while the former had their source in the passions.[1] Considering this, Hume held what might appear to be surprisingly dogmatic views as to which traits count as virtues and which as vices, and he often used quite vehement language against those who denied the existence of the moral distinctions that he was trying to explain.[2] The reconciliation of these two aspects of Hume can be found in his belief, which many would now reject, that the constitution of human nature is so uniform that the appropriate sentiments are guaranteed to be present in virtually any representative of the species. But if human nature really guaranteed the same affective responses to the same actions, objects, and characters, then a case could be made that the terms that seem, on the basis of surface grammar, to refer to properties of those actions, objects, and characters, really do refer to them. At least there would be obvious candidates for their referents. True, these terms—'immoral', 'good', 'virtuous'—would have their origin in the passions in some sense. But to claim that they therefore

Thanks to Al Mele and Mike Ridge for written comments on the initial version of this paper, and to the audience at the Second Annual Metaethics Workshop, especially Nicholas Sturgeon, David Sobel, James Dreier, and Sarah McGrath. Thanks also to a pair of anonymous reviewers.

[1] See Hume (1975: 285–94). [2] See Hume (1975: 169–74).

must *refer to* passions, or serve *essentially to express* passions, would simply be to commit the genetic fallacy.

So there is a potential line of argument from a completely uniform agreement in affective response to a realistic interpretation of normative terms. It is partly because of the availability of this line of argument that advocates of expressivist views of value have so often stressed the degree of disagreement that can be expected, even between two normal people. A similar emphasis on interpersonal disagreement plays a central role in the arguments offered in favor of subjectivist views of secondary qualities such as color. What this suggests is that the actual degree of agreement in response will be a matter of some importance in explaining whether a more realistic view, or a more expressivist view, is correct for any domain in which it is undisputed that the final account must feature human responses as a crucial element. But it also suggests the view, which this paper will attempt to clarify and defend, that accounts of the normative that feature human responses as a central explanatory element (let us here call them 'response-featuring', since 'response-dependent' currently suggests a more narrowly cognitive type of account) can be expected to fall in a spectrum. The distinction between cognitivism and expressivism may therefore be regarded not as an illusion—as an emerging position in meta-metaethics would have it—but as a matter of degree.[3] One tempting conclusion at this point would be that it is then the job of the moral philosopher to stake out and defend a particular location on this spectrum. But another conclusion, and the one I will defend in this paper, is that different normative notions almost certainly fall at different locations.[4] 'Harmful' and 'beneficial' might well be best construed as objective referring words, even if 'beautiful' or 'funny' are best viewed along expressivist lines. If this is true, it will

[3] Sturgeon (1994: 95–6) makes a similar suggestive remark about the difference between moral relativism and moral realism. I believe that a more satisfying conception of the spectrum of plausible views includes expressivism rather than relativism as one of the poles, and for a reason that Sturgeon himself mentions: when the relevant disagreement becomes very restricted and local, relativism even for these cases becomes unattractive. This has the odd consequence that relativism, at some point on the spectrum, must 'pop in wholesale.' On the other hand, the view on offer in this paper allows the expressivist role to be present at every point in the spectrum—even when normative language is also completely objective. It is only that this role takes on a much-diminished practical importance.

[4] This point is closely related to one made by Michael Smith, to the effect that our response-featuring concepts—not necessarily normative ones—can be expected to fall on a spectrum from clearly representational to clearly non-representational. See Smith (1993: 245). Smith focuses on the richness of the platitudes about the concepts, and the utility of such platitudes in supporting a relevant is/seems distinction. I focus on degree of interpersonal agreement.

allow us to combine expressivism and response-dependent cognitivism in a motivated way.[5]

For those interested only or primarily in moral matters I should make it clear immediately that the implications of this paper will not be direct. This is because the terms here considered will be what Moore tended to call 'simple' or 'unanalyzable'. 'Morally wrong', on the other hand, seems quite obviously conceptually complex, so that a response-dependent account of it, along the lines of an account of 'yellow', seems to be a non-starter.[6] The proper analysis of moral terms will be a genuine *analysis*, and will make use of both normative notions (harm and benefit, for example) and non-normative notions (causation and consistency, for example).[7] The notion of moral wrongness may therefore inherit a certain expressivist aspect from some of the relevant normative terms, and a certain objective aspect from others.

The general strategy of this paper will be first to describe, in a domain that excludes the complexifying element of normativity, one way in which uniformity of response might help to explain the presence of a certain kind of referring term in the language. This simpler domain is the domain of color. Still considering color and color responses, the 'perturbation' of disagreement in response will then be introduced and gradually increased.[8] It will then be argued that at a certain point one plausible result of the increased disagreement in response will be a change in the way it is most fruitful to think of the semantics of color terms, from a way that emphasizes the object that elicits the response to a way that emphasizes the response itself. And this change will happen despite the fact that the underlying changes are merely matters of degree. When the same story is applied to

[5] This point is distinct from, though consistent with, the idea that there is no sharp distinction between the descriptive and the normative. See Jackson (1998: 120).

[6] Johnston (1989: 42) expresses a similar worry about complex normative notions. See also Sturgeon (1994: 109–10). Jeremy Koons's (2003) criticisms of response-dependent views depend on taking them to be offering accounts of moral notions.

[7] This point allows me to agree with the arguments of Horgan and Timmons (1991) against what they call 'new wave moral realism', and also with Sean Holland's (2001) extension of their point against response-dependent versions of moral realism, *without* having to extend their conclusions to *all* normative properties. See Gert (2006).

[8] Here and throughout 'agreement' should be taken to mean 'agreement in response' and not 'agreement in belief'. Agreement in response is prior to agreement in belief in the sense that the relevant agreement in response explains the development of the terms in the language that are required before any explicit beliefs can be held. This is not to claim that language is required before beliefs can be held. But it is to claim that language is required before one can hold a belief with the content that, for example, grass is green. This should only seem counter-intuitive to those who are unaware of the variety of ways in which cultures develop color terms, and of the fact that some cultures do not have a word that corresponds to our 'green'. Compare Wiggins (1998: 202 point (b) and 205).

the normative domain, the resulting account will do justice to many of the leading intuitions that stand behind the non-cognitivism of expressivist views, and the cognitivism of response-dependent realist views.

As the reader will have gathered by this point, the emphasis of this paper is more directly on semantics than metaphysics. But I do not take these two domains to be easily separable. This, in combination with my views regarding the existence of a continuum of possible semantics for response-featuring terms, means that I do not think that there will always be clear answers to metaphysical questions regarding the ontological status of the various putative properties to which such terms seem to refer. This will be especially true if those questions are phrased in such a way that only a limited set of answers are available, as in 'Is the word "beautiful" a referring word that refers to an objective, mind-independent property of beauty?' To some readers this will be frustrating: surely ascriptions of normative property *N* are either truth-apt or not; surely *N* is either an objective property or not; surely if *N* is an objective property, it is a certain *kind* of objective property: a disposition, the basis for the disposition, or something else. While I think there will often be answers to these questions, I take semantics to be prior to metaphysics in the sense that, if a clear view of the semantics fails to provide answers, there are no metaphysical answers to be found.

1. THE CASE OF COLOR

Consider broad color terms: terms like 'blue' and 'green'. In the philosophical literature devoted to these terms there is a debate that is very similar to the debate between ethical realists and expressivists. This is the debate between color objectivists and color subjectivists. Color objectivists hold, at the very least, that it is very often simply true that a certain object is, for example, blue.[9] What is meant by calling an object blue varies widely among color objectivists. Some hold that such an ascription is true if and only if the object has a reflectance profile that falls in a certain (possibly quite messy-looking) class.[10] Others hold that it means only that the object

[9] If fact this is not quite true. Some who call themselves color objectivists relativize basic color judgments to circumstances and viewers, so that nothing is simply blue or green. The closest one would come to such a claim, in terms of basic color ascriptions, would then be a quantified claim such as 'for any person *P* who is normal, and for any circumstance *C* which is normal, the object is green-for-*P*-in-*C*'. See Cohen (2004) and McLaughlin (2003). I argue against such views in Gert (forthcoming) and for current purposes I will not consider them.

[10] See Byrne and Hilbert (2003).

is disposed to produce a certain distinctive sort of phenomenal response under certain conditions.[11] And there are a number of other proposals, including the view that colors are *sui generis* properties, not reducible to anything else.[12] The plausibility of all of these realist proposals relies, whether explicitly or not, on the idea that there is a very high degree of uniformity in color perception among humans. For example, the plausibility of the view that the color blue is a class of reflectance profiles relies implicitly on the idea that normal humans will agree to a very great degree on which objects ought to be classified as blue. For if there were not a high degree of agreement on this, it would be impossible to motivate the idea that any particular classification was correct. What would 'correct' mean here? Because of the implicit reliance on a high degree of agreement, a favorite strategy of subjectivists about color is to draw attention to the surprising degree of disagreement regarding color even among people with normal vision. For example, the location of spectral unique green—a green that is neither yellowish nor bluish—varies from observer to observer over a range of about 30 nanometers.[13] This is a huge proportion of the visible spectrum, which runs approximately from 400 to 700 nanometers. While I defend a version of color objectivism elsewhere, the present point does not depend on any particular resolution of the dispute between color objectivists and subjectivists.[14] Rather, it is to show, in a relatively simple context, how the actual degree of agreement in response is relevant to the plausibility of various semantic views.

Continuing to consider the semantics of color terms, let us imagine, quite counterfactually, that all human beings always have precisely the same phenomenal color responses to the same objects, all the time. This does not even approximate our actual situation, since it ignores contrast effects and the effects of variations in illuminant. Now, in the world we are imagining, it would be plausible to expect terms to appear in the language that applied to objects that were saliently similar visually: terms such as 'blue', 'red', 'green', and so on. One reason for this is that the uniform phenomenal response could be expected to figure in the processes of language learning. Admittedly, agreement in phenomenal response cannot by itself explain the emergence of a referring expression. But it can, almost by itself, explain how the meaning of a referring expression continues to be passed on from generation to generation. For whether one assumes that explicit ostension

[11] This is the view often attributed to Locke. See also Johnston (1997).

[12] See Campbell (1997) and McGinn (1996). G. E. Moore also seems to have regarded colors in this way. See Moore (1922: 268–72; 1993: 7, 10). On the other hand, Moore also took color terms to name natural properties. See Moore (1968: 588).

[13] Hardin (1993: 76–9) makes use of this statistic for this purpose.

[14] Most of that defense is contained in Gert (forthcoming).

and correction play the major role in such teaching and learning, or that children pick up on the proper use of terms simply from watching their elders, it remains true that agreement in response—including in what is experienced as saliently similar—will facilitate the process.[15] As to the initial emergence of any particular color term, this seems to be a contingent matter. In the actual world, for example, not all human cultures have words for all of the same colors. Some, for example, have only terms that correspond, rather roughly, to the English 'black', 'white', and 'red', and others have, in addition to these, only 'yellow' and 'green', lacking a term that corresponds to 'blue'.[16] And this is not to be explained by any corresponding differences in color vision at the physiological level.[17] So there is no hope of arguing that color terms, much less *our* color terms, are inevitable for creatures like us. But it seems plausible to assume that part of what explains the existence of color terms is the usefulness of being able to ask for or order someone to fetch an object of a certain color, or of advising them to avoid objects of some other color. So in a world in which this genuinely is useful, it should not be surprising if words for these colors appeared, and had objective referents.

Let us grant, then, that in the world we are imagining, color words are referring words. Certainly there is a fact of the matter, for any given object, whether it ought to be called by any given color word, so that the color words have, in a fairly straightforward sense, fixed extensions. Given our stipulated uniformity of response, such words will function almost like 'marked with an X', or 'marked with an O'. Despite complete agreement as to which objects are red or blue in this world, it might well remain true

[15] This story about ostensive teaching both explains and coheres nicely with McDowell's (1985: 111) suggested understanding of secondary qualities as qualities, true ascriptions of which must be understood to be true in virtue of an object's disposition to produce experiences that themselves require reference to the property. For if ostensive teaching is a canonical way of teaching such terms, and if there are no independently authoritative means for determining the color of an object (as there are for determining, say, the shape of an object), then there will be no way of characterizing the content of the characteristic phenomenal experience caused by colored objects except by descriptions involving colors. McDowell seems to be wrong in claiming that one needs to specify the content of the characteristic experiences in terms of the *very property* that typically causes them, since one might describe the content of a certain color experience by reference to *other* colors. But this is a very minor technical point with which I do not think McDowell would disagree.

[16] See Hardin (1993: 165–6).

[17] However, even in cultures without our color words, the boundaries between the colors are not an arbitrary matter. The phenomenology of our color visions imposes structural constraints, evidently, on our color concepts, even though it does not fully determine whether, say, we have four basic color concepts or eleven. See Hardin (1993: 155–9, 168).

that the *nature* of redness remains a matter of philosophical and scientific dispute. And so the internal fighting amongst the color objectivists might continue.[18] But there will be as little dispute over the claim that grass is green as there is over the claim that raindrops are made of water. Let us call this world of maximally uniform response 'world 1'. To repeat the point of the above discussion: in world 1 it is very easy to imagine the emergence of objective color terms. And while phenomenal responses play a crucial role in explaining their emergence, there is little temptation in this case to say that the color terms *refer to* anything other than properties of external objects.[19]

Let us now make a slight move in the direction of the actual world. In the next world we will consider, which we can call 'world 2', contrast effects and variations in lighting sometimes cause different observers to have different phenomenal responses to the same object. It remains true however that if two observers view the object in the same circumstances, they will have the same response. Now, if disagreements about the applicability of broad color terms such as 'blue' or 'green' could occur when each of two observers was in what was regarded as normal conditions, problems would arise in determining which response revealed the true color of the object. So let us stipulate that this does not occur, or occurs only very infrequently. That is, let us stipulate that—in world 2—the only variations in lighting or context that typically cause disagreements about the application of a broad color term are variations so great as to make one of the contexts count as non-standard. In world 2 there will be disagreements on some occasions. But they can, at least in principle and up to a small degree of vagueness, be resolved. For they are almost always the result of at least one observer being in non-standard conditions. So it remains reasonable to continue to think of color words as referring words. Indeed, in world 2 color words function much like shape words.[20] For while there is sometimes disagreement regarding the shape of an object, these disagreements can almost always be resolved by making sure that none of the observers are in non-standard circumstances for the observation of shape. In world 2 there will be uniform agreement as to whether two objects match in color under

[18] My own view, which I do not defend here, is that any reductive account will force us to embrace controversial counterfactual claims to which the semantics of color terms need not commit us, and which we would do better to resist. For this reason I endorse a primitivism about color. In the domain of metaethics, the corresponding position is misleadingly called 'non-naturalism'.

[19] This may be more convincing if one considers that phenomenal color responses are here playing precisely the role that sensations of heat play in determining the extension of 'heat'. Compare Kripke (1980: 129–30).

[20] The similarity should not be exaggerated, however. Shape, for example, will often be determinable through more than one sense.

the same conditions, since everyone has the same visual dispositions. It is true that if the visual systems of people in world 2 are anything like ours, then two objects that are perfect matches under one illuminant might not be perfect matches under another. This is the phenomenon of metamerism.[21] As a result, it will not always be clear whether two objects ought to be said to be *perfect* color matches or not. But the question before us is not the question of perfect color matches. It is the question of whether objects in world 2 count as having such colors as green, red, blue, and so on. As far as that goes, there will be as much agreement as there is about the application of any minimally vague term such as 'square' or 'circular'.

Now the interesting part of the thought experiment begins. Let us introduce and increase the amount of interpersonal variation in phenomenal response. If we stipulate just a small quantity of such variation, it is reasonable to expect that this will have no very drastic effect on the semantics of the terms. If the variation is slight—as it is, for example, in world 3a—it is likely enough that speakers will not even notice it. Of course the semantics of a term might change without speakers taking any notice. But even if we want to say that the semantics of the terms changes with the introduction of this sort of disagreement, the most plausible view of such a change is merely that the vagueness of the term—already present in world 2 on account of the range of viewing conditions taken as normal—increases. As the variation increases—as we can stipulate that it does in worlds 3b, 3c, and so on—the term's degree of vagueness increases. But as long as this degree of vagueness is small enough, there seems little reason to think that the essential semantic properties of the terms will change. If they were referring terms in world 2, they continue to be referring terms in worlds 3a, 3b, 3c, and so on. It is worth noting that up to this point there will be no problem in explaining how color terms function grammatically as predicates that apply, truly or falsely, to objects. Deductive inferences involving color predicates will preserve truth, at least if we except the kinds of cases that appear in discussions of vague terms, such as the inferences that make up sorites arguments.

It is true that with the introduction of even relatively slight interpersonal variation in color perception, we should expect fairly widespread disagreement as to whether or not two objects *match perfectly in color*. But we can safely ignore such disagreements, since we are concerned with the semantics of the broad color terms, and not with the truth of claims about indistinguishability, and we are continuing to stipulate that there is very little disagreement about the applicability of the broad color terms. Inhabitants of worlds 3a, 3b, and 3c will admit that not all green objects look the

[21] See Hardin (1993: 27–8).

same, that any given green object looks slightly different in slightly different viewing conditions, and that the same object may look slightly different to different people. They will therefore distinguish, as they should, between the color of an object—green, say—and the particular phenomenal quality of the experience a particular person has when viewing that object. These need to be distinguished since one is a property of an object, and the other is a property of an experience.[22] Each person's visual equipment is a way of telling what color an object is, by way of the quality of the phenomenal experience the person has in viewing it. These different ways of determining color will result in slightly different results, and may therefore result in different answers to questions such as 'Are these precisely the same color?' But this should cause the color objectivist no worries yet. There are, after all, also various ways of measuring the length of the coastline of a state, and these ways will also yield slightly different answers. This does nothing to detract from the objective truth of the claim that the coastline of Maine is longer than that of New Hampshire, or that the coastline of New Hampshire is between 25 and 30 miles long.

As we continue to increase the degree of disagreement, let us also assume that the distribution of responses is normal. That is, let us assume that when we plot the percentage of the population against the range of phenomenal responses that members of that population have to a particular object, we get a clustering around some central response, tapering off in both directions. So, for example, a certain leaf might appear unique green to the majority of people, while a few will see it as somewhat yellowish or somewhat bluish, and a very few will see it as almost entirely yellow or almost entirely blue. If the outliers make up only a very small proportion of the population, then it would be natural, if not inevitable, for the semantics of the term 'green' to remain more or less the same. But it would also be natural for there to emerge a new term that classified the outlying responses—and those who regularly have them—as defective. For it will be relatively easy, and relatively important, to identify these outlying responses: easy, because they conflict with the responses of the vast majority of other people, and important because we do not want to rely on those who regularly have them. Let us suppose that in world 3c there has emerged a term to describe those who are visually defective in the relevant way. The easy identification of these visually defective people means that even such people can use color

[22] I do not mean to beg questions about color here: I mean only that in worlds 3a, 3b, 3c, etc., interpersonal variation is sufficiently limited that the semantics of color terms ought to be regarded as very similar to what it was in world 2. Thus, for these worlds, the claim that 'green' picks out a property ought not be controversial. Whether our own world is best thought of in the same way is a different question.

words for many of the purposes for which they are useful. For example, if they know the colors of objects based on the testimony of others, they can identify those things by color when asking for them. Similarly, they can make use of their memory, or solicit testimony, when they need to act or make decisions based on the colors of objects.

Once a term for the visually defective has emerged, the objectivity of colors can remain proof against a certain amount of variation in the responses of human beings. For example, in world 3c we can identify red objects with those objects that cause a certain phenomenal response in normal—that is, non-defective—human beings. And this class of objects may well remain unchanged even if the distribution of responses in human beings changes to a fairly significant degree. For there is no reason to think that the classification of responses as defective need be a strictly statistical matter. Rather, a host of pragmatic considerations will have their influence on the formation of the relevant concepts. For example, many standard kinds of color-blindness result in a diminished capacity to make color distinctions, and this might very easily manifest itself in some important practical tasks. Thus, even if more and more people begin to suffer from such a condition, the judgments of those who retain a fuller discriminatory capacity might plausibly continue to be regarded as authoritative.[23]

James Dreier (1990: 12–13) has objected to the idea that objectivity can be preserved by the means just described, since if this were so, we could generate an objective property out of any relatively common human response: we need only call those who do not share the common response 'defective'. One answer to Dreier's worry is to remind ourselves that, except in exceptional circumstances, it is not up to us to determine the meanings of the words we use in our native languages. Suppose, for example, that undergoing a certain procedure typically hurts. If a particularly insensitive person happens not to be caused any discomfort by the procedure, and sincerely says that it doesn't hurt, our only option is to concede that he has

[23] The relevance of such extra-statistical factors makes the relation between responses and properties, on this view, more like that involved in McDowell's example of a quality that is response-dependent, but not secondary: fearfulness. McDowell takes the relevance of such extra-statistical factors in determining what might count as a *mistaken* response to justify a relatively sharp distinction between secondary qualities and evaluative ones: a value 'is conceived to be not merely such as to elicit the appropriate 'attitude' (as a colour is merely such as to cause the appropriate experiences), but rather such as to *merit* it' (1985:118). Wiggins (1998: 187) takes a similar view. But my view is that there is no very sharp distinction to be made here. If it is true that harms *ought* to elicit aversion, so too is it true that green things *ought* to be seen as green. The 'ought' here is not an additional normative term, brought in from outside the response-dependent framework, or constructed from within it. Rather, it is a marker of the objectivity of the relevant property and of the consequent possibility of mistaken responses.

spoken truly. Our language doesn't contain the resources for ruling out this person's response as defective in the way it does when someone says of a red object that it is grey. Of course we could coin a new term, 'turts', and stipulate that it is to be true of anything that *typically* hurts. In that case, if our insensitive person claimed that the procedure didn't turt, we could legitimately correct him. But if we want to talk about colors, or whether or not something hurts, or—as we will do later on—whether something is harmful or beneficial, we must speak in accord with the semantics of the relevant terms as they are given to us.[24]

This same point—that it is not typically up to us to determine the meanings of existing terms in public languages—also provides some reason to be skeptical about inferences from the semantics of one secondary quality term (say 'bitter') to the semantics of others (say, color terms). In arguing for a kind of radical relativization of basic color concepts, some philosophers make use of an example from the domain of taste. One popular example is that of phenol-thio-urea, which, it is claimed, tastes bitter to three-quarters of the population but is tasteless for the remainder.[25] In this case it does seem extremely plausible to claim that neither of these groups has a monopoly on correctness of response. Rather, the appropriate thing to say seems to be that phenol-thio-urea is bitter for those in one group, while it is tasteless for those in the other. But it simply does not follow from the appropriateness of this claim that we should expect the same evenhandedness regarding the greenness of grass: for one thing, there is nothing that plays the role of phenol-thio-urea in the domain of broad color categories. That is, there is nothing that appears paradigmatically green to most of the population, but does not appear at all greenish to a statistically significant remainder. Moreover, while cognates for the term 'color-blind' appear in many languages, there is typically no term to describe people whose sense of taste differs markedly from the norm. And, as we will see in what follows, there is even less reason to expect the same evenhandedness in the case of the properties of being beneficial or harmful. One explanation for this fact may be the following: in the case of these normative properties it is extremely important to identify those who have anomalous responses, since we can expect them to behave in ways that are very importantly different from those who have the standard responses.

[24] That the meanings of the terms are given to us, and not invented by us, also helps defuse a worry about a latent circularity in any attempt to explain redness in terms of the responses of normal people in normal circumstances (people who can, that is, and among other things, discern red things from non-red ones). For we learn the extension of 'red' at the same time that we learn what 'normal' means in connection with the perception of colors. In this connection see Wiggins (1998: 212 n. 19).

[25] See McLaughlin (2003) and Jackson and Pargetter (1997).

Although the notion of 'visually defective' is not purely a statistical matter, it remains true that the task of identifying the visually defective will be easier when they are comparatively rare. In that case a few comparisons with the color judgments of other individuals will typically be sufficient. But as the distribution of phenomenal responses to any given object flattens out—as it does, let us say, in worlds 4a, 4b, and 4c—this sort of identification becomes more and more difficult. One possible result of such a statistical flattening is that only very extreme differences from the central tendency will be labeled as defective, and that lesser degrees of variation will count as normal. Another result might be that the terms that identify defective responses simply disappear, so that the term functions as 'hurts' did in the above example. In both cases it will become more and more common for two people to apply two different color terms to the same object, without either of those people's responses being regarded as defective. This need not entail that there is any drastic change in the semantics of color terms. Rather, an increase in the number of cases of faultless disagreement can be seen as a mere widening of the penumbra of vagueness in what was initially only a slightly vague term.

However, as the penumbra of vagueness of color terms increases, their usefulness will change, and it will become possible to understand their semantics in a different way. One important change in usefulness results from the fact that it will become increasingly difficult to rely on the assumption that other people will classify objects as one does oneself. Color claims will give less and less information about objects, and more and more information about the people who make color claims. The usefulness of color terms will therefore increasingly depend on the usefulness of knowing that some particular person happens to see a certain object in a certain way. If there is little use in such knowledge, we might expect color terms to disappear from the language. However, such knowledge might well be of practical interest to us. For example, it might turn out that seeing something as bluish green is typically pleasant, while seeing it as yellowish green is not.[26] If this were true we would want to know whether our friend sees a particular object as bluish green or yellowish green before we bought it as a housewarming gift. In general, if phenomenal responses are correlated in a relatively robust way with something importantly affective, it will remain useful to have a term in the language that functions grammatically like a

[26] For some evidence that this actually is true, see Hardin (1993: 163). Mark Leon (2002: 184–5) appears to think we have a practical interest in how things phenomenally appear that is independent of such considerations, and that this could underwrite a view on which color terms are in a fairly literal sense semantically ambiguous. But concerns about the privacy of experience make me doubt Leon's premises here.

predicate, but whose application to an object gives more information about the person speaking than about the object. The semantics of such terms will fruitfully be explained on expressivist lines, though what will be expressed will be a phenomenal color experience, and not an affective response. What is interesting is that it may well remain possible to think of the term as a referring expression with a significant degree of vagueness.[27] In such cases we should not be surprised to note that the term functions unproblematically in logical inferences. That is, this account of the emergence of terms that are properly understood as serving to express non-cognitive attitudes such as phenomenal color response provides a ready solution to what has been termed 'the embedding problem' for expressivism.[28] At least it provides such a solution for those expressivists willing to accept the cognitivist element in this story.

2. NORMATIVE TERMS

Even if one is persuaded by the story just told about color terms, the application of the lessons of that story to the domain of the normative is not straightforward. There are a number of significant disanalogies between color terms and normative terms that might prevent a similar story from going through. One difference is that our access to the colors of objects is, for practical purposes, entirely mediated by visual experience. We cannot, that is, use our fingers to discern the color of an object. Color objectivists all agree that the colors of objects supervene on their physical microstructure.[29] But knowledge about that physical microstructure is not available to us when we apply color terms. Rather, we apply the terms on the basis of our visual experience, or the testimony of someone who has had such experience, or induction involving such experience. In the case of normative terms such as 'harmful' and 'beneficial,' on the other hand, the subvening

[27] It seems to me therefore that the proposal of this paper can accommodate Peter Railton's useful reminder that not all secondary qualities are the same. 'Bitter', for example, seems actually to function in something like the way that affectively loaded color terms would function at the extreme of variation described in the text, since it is acceptable for different people to apply it, in certain cases, to quite different things. See Railton (1998: 77). It is worth noting, however, that the presence of the word 'bitter' in the language plausibly depends on the existence of a class of unambiguously bitter things, for it seems correspondingly *implausible* that the word 'bitter' could be taught effectively based *solely* on the behavioral manifestations of experiences of bitterness.

[28] Geach (1958, 1965). See also Searle (1962, 1969).

[29] This is most obviously true for categorical-basis physicalists like Frank Jackson. But it is equally true for reflectance physicalists like Alex Byrne and David Hilbert. Even Moore admits it.

base is also typically available to us. That is, let us suppose that the loss of freedom counts as harmful. When we are witness to (or imagine) some behavior that results in a loss of freedom, and apply the term 'harmful' to it on this account, we do this because of an awareness of the loss of freedom. But this difference in our epistemic relations to the supervenience base is no reason to abandon the relevant analogy with color.[30] For the point of the analogy is only that we learn to apply the terms because there is a common salient response evoked by a certain class of objects (in the case of color terms) or events (in the case of the normative terms 'harmful' and 'beneficial'). The disanalogy does, however, explain how it is that we can use stories as part of the teaching of the normative terms, while we cannot do the same in teaching color terms. The relevance of imagined and counterfactual situations also serves to defuse a worry about relativism that I will address later.

Suppose now, as a slightly tempered Hume might have done, that the affective responses of human beings are virtually uniform with respect to the sorts of things that are attractive and aversive, if not in the degree of attraction and aversion.[31] And imagine that we learn the meanings of terms that apply to the uniform objects of these affective responses through some combination of ostensive teaching and passive observation. Further, imagine that these terms are 'harmful' and 'beneficial'. Then two things will be true. First, these terms will be objective referring terms, just as color terms were in worlds 3a, 3b, and 3c. But also, the salient response, by means of which the learning of these terms was made possible, is an affective response, so that at least typically we can expect people who spontaneously classify something as harmful to have the appropriate affective response. Of course this is not necessarily true. Just as people can have visual problems—either permanent or temporary—and still manage to apply color terms correctly on many occasions, so too might people who have affective disorders both learn, and use, our two normative terms correctly on many occasions. Now, in the case of color, those whose responses fall outside of a certain range are called 'color-blind'. In the case of the harmful and beneficial, those whose responses are outside of an analogous range might be called, for example, 'crazy', 'silly', 'stupid', or 'irrational'.[32] It is no surprise that there are terms

[30] Compare Smith (1993: 247).

[31] 'Slightly tempered' because of the insertion of the weakening 'virtually'. See, e.g., Hume (1975: 212–33, 268–83).

[32] I do not mean to beg any questions by including the term 'irrational' in this list. There is an emerging technical philosophical usage, stemming from T. M. Scanlon's use of the term, according to which irrationality always involves a form of *akrasia*, since it must involve acting against one's own normative judgment. See Scanlon (1998: 25–32). I criticize this technical usage in Gert (2004: 214–16). Here I want only to make it

for people who respond anomalously in this respect, since it is important to identify them. Such people do not respond to common incentives and disincentives in predictable ways, or in the ways presupposed by formal and informal public policies such as those that govern driving automobiles or coming across the private property of other people. As a result, even if we do not particularly care about their well-being, we need to be wary of them. Indeed, it is not surprising that we have a spectrum of such terms, since some anomalous responses are more alarming than others. The existence of the terms 'crazy', 'silly', 'irrational', and so on, and their relation to the terms 'harmful' and 'beneficial', explains how a certain version of the internalism requirement can be true, at least for reasons that involve harms and benefits.[33] This is the version of internalism according to which any agent who believes that there is a reason of such a sort to perform a certain action will be motivated accordingly, at least insofar as that person is rational (that is, not crazy, stupid, etc.).[34]

This story about the meanings of the terms 'harmful' and 'beneficial' parallels the story about the meanings of color terms on world 3c. It entails that there is a certain amount of vagueness to the term 'harmful'. But it also entails that virtually everyone is averse to the same kinds of things. Whenever there is any significant amount of disagreement in response, with the consequent disagreement in application of the term 'harmful', this will count as a case in which there is no truth of the matter as to whether or not something counts as a harm. Now, as we move from the equivalent of world 3c to the equivalents of worlds 4a, 4b, and 4c, there is much less agreement. It may be useful, in considering these worlds, to make use of a distinct normative term: one for which one of worlds 4a, 4b, or 4c approximates the actual world. So let us change from 'harmful' and 'beneficial', about which there actually is a very great deal of agreement, to 'funny', about which there is substantially less.

clear that this is not the sense of 'irrational' I mean to indicate. Rather, I want to capture the more generic concept of a failure in practical mental functioning.

[33] In fact, I do not think the response to benefits is exactly attraction, so that the version of internalism will be slightly different from the simple version suggested by these remarks.

[34] The current proposal therefore allows us, at least in this case, to go a step beyond David Wiggins's explanation of what he, following Stevenson, calls the 'magnetism' of value terms, and what others have discussed under the heading of 'internalism'. Wiggins's points are that if a value property and a response are related in the way he and I suggest, then of course it will be 'strange for one to use the term for the property if he is in no way party to the attitude' and that this can be true even if 'he regards it as a matter of keen argument what it takes for a thing to count as having the property'. Wiggins's explanation is of course available to us in this and other cases as well. See Wiggins (1998: 198–9).

Despite a very high degree of variation in what people find funny, it is at least plausible that (in the present context) my inhaling normally, or my taking a sip of coffee, is *not* funny. On the story I am trying to motivate, there is no objective *positive* core to funniness. That is, there is nothing that one *must* find funny, in order not to be classified as relevantly defective. This position, however, is consistent with the claim that there are things that one *cannot* find funny without being so classified. If you find my normal inhalation of air funny, something is wrong with you. If you have just taken some drugs that have altered your sense of humor but not your intellectual powers, what you might appropriately say is that even though you know it isn't really funny, the drug is making my normal inhalation *strike you* as funny. But this doesn't mean that you must claim that it *is* funny, any more than someone who knows he is wearing rose-tinted glasses, or experiencing an after-image induced by a green light, must say that his milk is pink.

The above picture of the semantics of 'funny' is one according to which it is never determinately true that something is funny, though it is sometimes determinately false. Another technical way of putting this is that the term is *maximally vague*: it has only a penumbra, and a zone outside the penumbra, but nothing *inside* the penumbra. What this means is that when people disagree about what is funny then, at least typically, they ought to acknowledge that neither has any special claim to correctness. Because the term 'funny' has no solid core of determinate reference, and because its penumbra of vagueness is so large and plays such a significant role in its semantics, we can call the whole penumbra the 'pseudo-reference' of the term. For while no particular applications of the term 'funny' to things within the pseudo-reference are correct in a sense that implies that failure to apply the term in that way is incorrect, it is also true that applications of the term within the pseudo-reference are correct in the sense of not being *incorrect*.

What someone is willing to call 'funny' gives some indication of her response to it. That a term is vague need not mean that there is no reason why a certain speaker applies the term to a borderline case as she does.[35] 'Green', for example, is vague, because when there are disagreements as to where green ends and blue begins it is often the case that there is no fact of the matter as to who is correct. But it would not be surprising if the explanation for such a disagreement turned out to lie in the fact that the red/green channel of the observer who favors 'green' is signaling green more strongly

[35] This fact, together with the idea, discussed below, that arguments can sometimes affect how one applies a normative term, provides an explanation of what has sometimes been called the 'essential contestability' of normative notions.

than the yellow/blue channel is signaling blue, while the reverse is true of the observer who favors 'not green'. That is, it would not be surprising if we could explain the first observer's application of the term 'green' by appeal to the fact that something is going on in her that is similar to what goes on in normal people when they look at something unambiguously green. But in fact it is extremely important to note that despite the plausibility of such an explanation in any particular case, we need not insist on its availability. Just as it is possible to sincerely, competently, and correctly apply color terms to objects that do not, at the moment, appear to have the color named by the term (or even to objects that do not even appear at all, since they are in the closet), it is also possible, once one has learned the term 'funny' in the normal way, to apply it sincerely, competently, and correctly even in the absence of the characteristic response. Of course if one seldom or never has the characteristic response, this may well make it impossible to achieve competence. But this fact is irrelevant to the current points. These points are (a) that the presence of a characteristic response can sometimes explain our use of a term even though (b) that response need not, as a matter of semantic necessity, be present for our use to count as sincere, competent, and correct.

In the case of the term 'funny' one objection to the above story is that it locates the vagueness in the wrong place. As I have admitted, and as is surely true, when you and I disagree about whether or not a certain joke is funny, very often both of us are absolutely clear that neither of us is misapplying the word. The objector takes clarity about this matter to show that we are not within the penumbra of vagueness that I am suggesting makes up the pseudo-reference of 'funny'. Where is the vagueness of 'funny', according to the objector? It appears only in those relatively rare cases in which we have a different kind of disagreement: a disagreement as to whether or not the term 'funny' has been used correctly or not. Consider: if I claim that, in the present non-remarkable context, it is funny that water covers most of the Earth, then I have misapplied the word. On the other hand, if I claim that Jay Leno's monologue was funny, I have not misapplied it. But neither do you misapply it in *denying* that the monologue was funny. In each of these three cases it is clear whether the word 'funny' has been used acceptably or not. According to the objector, the vagueness of 'funny' appears only in cases of a different sort: cases in which it is not clear whether the term 'funny' is being applied acceptably. Let us call one such case 'case V'. It is only case V, and similar cases, that make up the penumbra of vagueness belonging to the term 'funny': that penumbra does not include all of the unambiguously *correct* applications, as my account suggests. The response to this worry is that there is certainly a difference between the disagreement over Jay Leno's monologue, and the disagreement that arises in case V.

But the distinction between first and second-order vagueness is sufficient to account for this difference. As in the case of baldness, for example, there will be cases in which it is clear that it is not a mistake either to apply, or to refuse to apply, the term 'bald'. This is an instance of first-order vagueness, and it is analogous to a disagreement over Jay Leno's monologue. On the other hand, there are cases in which there is disagreement as to whether or not someone is in the penumbra of 'bald' or not. That is a disagreement as to *whether or not there could be faultless disagreement* in the application of 'bald'. This is an instance of second-order vagueness, and it is analogous to the disagreement to be found in case V.

On this account many paradigmatic normative disagreements are in essence the same as disagreements we explain by reference to vagueness. This claim might spark the worry that it seems to leave the demand for normative consistency unexplained. If faultless disagreement over whether or not Jay Leno's monologue was funny is merely a special instance of the faultless disagreement we see in less controversial cases of vagueness, why is it that I cannot move back and forth in my assessment of that monologue without being criticized as inconsistent? But in fact this same demand of consistency is found in the application of many vague terms. If I call X bald, then I may well be forced to call Y bald too, given that X has noticeably more hair than Y. And yet someone might sincerely and competently deny that Y is bald. Indeed, I myself might have classified Y differently, had I been in a different mood when called upon to make my initial judgment.

So it remains fruitful to conceive of the fact that we can disagree, without error, about the funniness of Jay Leno's monologue as an instance of vagueness. Perhaps there might be some reason not to *call* it vagueness, but if this is so it may well merely be a result of the fact that there is a difference in degree that is worth marking: the same difference that made it useful to introduce the technical term 'pseudo-reference' as a name for the penumbra of first-order vagueness for these terms. Our reluctance to use the word 'vagueness' here need not indicate any essential difference in the nature of the phenomenon. As in more standard cases of vagueness, such as the case of baldness, there is a range of cases to which the term applies, and it is partly the result of the way that those cases strike the speaker—the way those cases seem saliently similar to other cases in the same range—that determines whether the speaker will apply the term or not. And when there is less uniformity in how a given case will strike different people, and more practical consequences of its striking someone this way rather than that (as is more likely to be the case when the response is an affective one) we have an increased interest in learning about how it strikes any given person. All this has the result that we will focus more on the expressivist element in the semantics of the relevant words. But even in such cases we can

continue equally usefully to apply the semantics of vague referring words. And even in the case of more objective words, such as color words as they function in worlds 3a, 3b, and 3c, we can often see the use of a word as the expression of a non-cognitive attitude.[36] The two semantics need not be seen as incompatible.

Accounting for the usefulness of an expressivist semantics by joint appeal to vagueness and to the role of a particular salient response allows, I think, for a nice resolution to an interesting dispute between Nicholas Sturgeon (1991) and Simon Blackburn (1991a, 1991b). The dispute begins with Blackburn's acknowledgment that, in the absence of a unifying attitude that must be involved in the sincere expression of normative claims of any particular sort, it becomes a problem to explain what normative disagreement amounts to: it isn't disagreement over whether or not some objective property applies in a given case, but it isn't a specifiable kind of disagreement in attitude either. The dispute is essentially over whether or not this point is fatally destructive to Blackburn's project. According to the view on offer in this paper, it need not be seen as fatally destructive, since a certain kind of affective response can be seen as somehow *central* to a given normative term (say, funny), without the expression of that response being essential to each particular sincere competent use of the term. But for such normative terms there may still remain a referring function—'softened' to whatever degree the term is vague—so that the realism for which Sturgeon is arguing is also preserved. The only normative terms for which this resolution cannot be offered are those that are located so far down the 'normative spectrum' that they cannot plausibly be regarded as having *any* objective limits to their application. But I do not know of any normative terms that function in this unrestricted way.[37]

3. WORRIES AND OBJECTIONS

The proposal this paper is making is in many respects quite radical. It is making the claim that normative terms are, from a semantic point of view, really not very different from many other ostensively taught terms for descriptive properties, such as color terms. Many normative terms refer, in a more or less standard way, to those things that are used in the processes

[36] In the case of color words, this will be possible only when the person is applying the word on the basis of a visual inspection of the object to which he is applying it.

[37] Foot (1978a, 1978b) expresses a similar skepticism about such unrestricted normative terms in a number of places.

of teaching and learning them, and to things that are such as to provoke a saliently similar response. The fact that salient similarity is, in this case, a matter of eliciting a similar affective attitude means that normative terms are bound up with affect in a way that has a persistent distorting effect on theorizing about such terms, and can make them appear mysterious and certainly not open to a straightforward realistic interpretation. For example, it is sometimes claimed that some sort of motivational internalism is true for normative judgments, while color judgments have no analogous connection to motivation, so that the analogy between normative properties and secondary qualities must be flawed.[38] But by itself this is no objection to the analogy at all: the plausibility of motivational internalism for normative properties is simply a direct result of the fact that the characteristic response is affective, while for color it is phenomenal. Normal people are *moved* by normative properties of which they are aware, just as normal people are *caused to have certain phenomenal experiences* by color properties of which they are aware.

It is not to be expected that this paper will have completely dispelled the air of mystery that surrounds the normative. In this final section I want to address a number of worries and objections, but I am fully aware that even if my answers are satisfactory, many other doubts will arise. Perhaps the most that can be hoped is that this section will provide some inductive evidence that seemingly destructive criticisms can be dealt with satisfactorily.

1. What is the Account of Vagueness?

A good number of philosophers have suggested that normative terms cannot receive a straightforward realistic analysis because of the following feature of normative talk: disagreement can persist, even between two competent speakers, and even in the face of total agreement on all relevant non-normative facts. For some this 'remarkable fact' leads to expressivism, while for others it leads to relativism. This paper, on the other hand, has tried to suggest that there is nothing in such disagreements that we do not see, perhaps in a somewhat milder form, in the case of uncontroversially descriptive terms. We can account for such irresolvable disagreement simply by appeal to vagueness. Given the variability in human visual systems, and the way in which color words are taught, it is inevitable that there will be disagreements in spontaneous judgments as to whether a certain pair of green dress socks are slightly bluish, or slightly yellowish. And this sort of disagreement may well occur even in the face of total agreement on all

[38] See, for example, Blackburn (1993: 160).

relevant non-chromatic facts. This does not mean, however, that 'green', 'bluish', and 'yellowish' cannot receive straightforward realistic analyses. Rather, it only means that they are to some degree vague, and that the particular pair of green dress socks at issue falls within the penumbra of both 'bluish' and 'yellowish'. I have therefore also suggested that there is no fact of the matter as to whether or not the socks really are bluish or yellowish, and that when a particular speaker spontaneously applies the term 'bluish' to them, we can take this as an indication that she is having a phenomenal experience that has a certain quality. This highlights the expressive element in response-dependent terms—even non-normative ones. This expressive element takes on more semantic importance as the degree of vagueness increases. In the case of normative terms, it is associated with affective states rather than phenomenal ones, and this link with affect accounts for many of the phenomena that philosophers have taken to be mysterious or problematic features of normative thought and talk: for example, that it is not a contingent matter that a rational person who regards something as harmful is averse to it. Our interest in knowing about other people's affective states also helps to explain the persistence of normative terms in the language, even when they become 'maximally vague' and therefore give relatively little information about the objects to which they are applied. In worlds 4a, 4b, 4c, and so on we might well expect color terms to disappear, since their primary functions depend on their reference being more or less determinate, and we have relatively little practical interest in knowing what sort of chromatic experiences someone else is having. But the analogous claims are false in the case when the relevant responses are affective, and can be expected to bear fairly directly on choice and action.

In all of the preceding argument, the operative notion of vagueness was left almost entirely unexplained. Indeed, all that was really said was that faultless disagreement in the application of a term indicates the presence of vagueness. But this is clearly insufficient. Indexical terms also have the same feature. I might apply the phrase 'accessible by car' faultlessly in my description of Buenos Aires, while you might faultlessly apply the phrase 'not accessible by car'. For I might be in La Plata, while you might be in Madrid. I therefore need to say something to explain why I assimilate paradigmatic disputes over the application of a normative term to vagueness, rather than interpreting them as the relativist might. In order to do this, it might seem that I am required to offer some substantial theory of vagueness. Indeed, given the radical nature of my conclusions, one might even suspect that I am covertly depending on a very controversial notion of vagueness. But I do not think this is the case. All that I think I need to do is to appeal to a pair of quite plausible ideas. The first idea is

that color terms are reasonably thought of as vague, and not as hiding an implicit reference to the speaker. This should strike most readers as a benign assumption, inasmuch as an overwhelmingly standard illustrative example of vagueness—second perhaps only to baldness—involves the terms 'red' and 'orange'. The second is that if the response-dependent story I am telling about the terms 'harm' and 'benefit' is plausible, then these terms will have the same sort of semantics as color terms, at least as far as the question of vagueness-versus-indexicality goes. And I have already argued that if this is the right view of more objective normative terms, then there is no reason to view less objective normative terms, such as 'funny', as different in essence. Rather, the distinctive character of such 'maximally vague' terms is that their positive core of reference has disappeared, swallowed up in the penumbra that, for other vague terms, separates the class of objects to which the term definitely applies from the class to which it definitely does not.[39]

Another response to this worry that is at least worth mentioning is that not all philosophers would accept the distinction between vagueness and indexicality upon which it depends. That is, there is one well-represented conception of vagueness according to which vague terms *are* a species of indexical. Diana Raffman, for example, has a view according to which the applicability of a vague predicate to an object depends partly upon a feature of the speaker she calls an 'internal context'. Such a context is provided by the (current) disposition of the speaker either to classify the object as in the positive or negative extension of the predicate, or to move to another context in which such as classification will then be made.[40] Raffman's standard example is color. If her account provides a correct explanation of the vagueness of color terms, my story about the nature of normative terms strongly suggests that her account will apply equally well to them.

[39] It is an independently interesting result of this picture of increasingly vague predicates that some vague predicates—maximally vague ones—will not lend themselves to sorites arguments. This is for a reason that Nicholas Smith (2005: n. 27) nicely explains. His point is that in order to generate a sorites argument using the predicate '*F*', we need a series of objects that begins with an object *a* such that '*Fa*' is definitely true, and ends with an object *b* such that '*Fb*' is definitely false, but for a term to count as vague we only need a series that starts with an object *a* such that '*Fa*' has one truth value, and ends with an object *b* such that '*Fb*' does not have that truth value. It seems to me that 'funny' and 'unique green' are both instances of such predicates, and that they therefore falsify the plausible hypothesis that we can identify vague predicates as those that are susceptible to the sorites paradox. It is worth noting that, though Smith identifies the conditions that would have to be met by a vague predicate that is not susceptible to the sorites paradox, he is skeptical that any predicate meets these conditions.

[40] This is a very crude representation of her view. See Raffman (1996).

2. The Responses Relevant to Normative Terms are Influenced by Argument

My response to the preceding worry depended on the plausibility of understanding normative terms such as 'harm' and 'benefit' as having essentially the same sort of semantics as color terms. The plausibility of this claim depended, in turn, on the plausibility of the idea that the affective responses associated with 'harm' and 'benefit' play the same role in determining the extensions of these words as do phenomenal color responses in determining the extensions of color words. This may give rise to a new worry. Phenomenal color responses cannot—one might suggest—be influenced by argument, while the affective responses associated with normative terms are commonly influenced in this way. For example, one might initially regard the loss of one's inherited title as a harm, but become persuaded that it really doesn't matter. And as one becomes so persuaded, it is plausible that one's basic attitude towards the loss changes. This change in attitude, it might be argued, finds no parallel in the case of color. It is the result, one might suggest, of the fact that the relevant attitudes are, in Thomas Scanlon's (1998: 20–4) terminology 'judgment-sensitive'. This means that the relevant attitudes are not usefully thought of as the basic extension-determining elements in an account of the related normative terms, but are somehow dependent on a cognitive judgment: a judgment that one might actually express by using the relevant normative term. That is, the attitude of aversion one characteristically has towards harms is the result of the judgment that they are in fact harms. If this is right, the story I am telling about the emergence and semantics of normative terms will not be at all plausible, since it will lack an account of the content of this more basic normative judgment: the one that stands behind the non-cognitive attitude. Worse: even if we could supply such an account, the response-dependent story I have been telling would lose its point. For we would already have an account of the content of a harm-judgment.[41]

This is an interesting challenge, but not, I think, very destructive to the analogy I want to make. First of all, the challenge is most persuasive when one thinks of moral notions. Surely it is true that people are sometimes rationally persuaded to change their moral beliefs, and one common result

[41] It should be evident that advocates of the view underlying the objection cannot make use of a response-dependent account of the basic normative notions that figure in the judgments to which our attitudes are sensitive. They must therefore either provide such an account, or acknowledge that the response-dependent account is at least attempting to explain something that their own accounts take as basic. Scanlon takes the latter route.

is that their attitude towards certain behaviors undergoes a corresponding change. For example, upon being convinced that homosexuality really isn't a moral matter at all, a former homophobe might well find that he no longer has the same constellation of negative attitudes towards homosexuals. But this fact is not relevant to my project, since I do not mean to offer an account of moral terms at all. Moral terms are not *basic* in the relevant sense. Rather, it is plausible that there is a correct moral *theory*. Such a theory will explain moral notions in terms of other, *more basic*, notions. Moral argument is often a matter of convincing someone that the moral view upon which he has been relying has consequences that he himself is unwilling to accept, and that he should therefore adopt a different view, and come to a different conclusion about the case at hand. Given the social benefits of people acting according to their moral convictions it is no surprise that our parents and other members of our community try to ensure that we act only in ways we judge to be morally acceptable. And it is no surprise that this training quite often has an effect. So if we come to change our views about the moral acceptability of some particular type of action, it will also be no surprise if a number of our attitudes undergo a corresponding change. But none of this is relevant to an account of the basic normative notions, such as 'harm' and 'benefit', that are the subject of this paper, and that we apply without relying, even implicitly, on any theory or definition.[42] The difference between the two cases can be seen in the fact that actual people (I mean 'non-philosophers') often do try, sometimes with success, to persuade other actual people to revise their moral views. But it is rather rare to encounter a situation in which a real person is in good faith trying to persuade someone else that something that person regards as a basic harm is not really a basic harm. Of course there are disputes about whether and to what extent a certain activity (say, taking drugs) involves the *risk* of harm. Similarly there are disputes as to whether or not a certain event (say, death) is essentially a harm (permanent loss of consciousness) or is, rather, the means to a benefit (eternal bliss). But these are distinct issues, best thought of as disputes about causal matters.

This last point also provides a distinct way to address the current challenge, even when that challenge is cast in terms of more basic normative notions. It is a commonplace that many apparent moral disagreements do not actually involve any disagreement at all about the most basic relevant moral principles. Rather, what looks like a difference in basic moral principle is often nothing more than a difference in opinion regarding the likelihood

[42] This claim is consistent with the fact that philosophers have offered explicit theories of harm and benefit, and that these theories may have, on some rare occasions, influenced their basic attitudes.

of various consequences, or the truth of certain claims about the motivations or intentions of the person whose action we are judging. For example, two people who do not disagree in any way about what makes acts morally wrong might still very easily disagree about the morality of supporting a certain kind of redistributive social policy. Similarly, even if 'harm' is an objective referring term, the extension of which is related to the affective responses of normal people in the way I have suggested, there still might be arguments between two competent speakers as to whether or not insulting someone involves doing him a harm. For there is a distinction between *basic* harms, such as pain and loss of freedom, and *derivative* harms, such as being the butt of a joke. Derivative harms count as harms because they increase the likelihood that one will suffer one of the basic harms. Because there can be dispute as to whether something that is not a basic harm really does increase someone's likelihood of suffering a basic harm, there can be argument about whether or not something counts as a harm—in the derivative sense. And if someone who had been convinced that a joke was really harmless becomes convinced of the opposite, this could easily bring about a change in attitude towards such jokes. But this does not imply any change in attitude towards the *basic* harms.

Finally, I can even concede that argument can, on occasion, influence the affective attitudes one has toward basic harms. But this concession need not trouble me unless the scope for such reasoned change in attitude is so great that it makes sense to say that we might be systematically wrong in a large portion of our most basic normative judgments. Here my response is simply a denial that this is a real possibility, just as it is not a real possibility that we are systematically wrong in a large portion of our confident judgments of color. It is true that there might be *local* errors: the responses that underlie the teaching of normative notions might well be perverted in some social context, so that honor, for example, is taken to be a basic benefit, and hair loss a basic harm. Such mistaken views might be corrected by argument later on. But this might also be the case for our judgments of the colors of certain objects. There is a certain amount of top-down processing in color vision. One's beliefs about an object—that it is a tree and not a rock formation, or that it is lying in the shade and not in the sun, and so on—can have a substantial influence on the actual phenomenal experience one has of the object.[43] Yet no one takes this phenomenon, by itself, to undermine a response-dependent realist account of color. Color judgments take place in an interpretative context. Because of this, if we represent the world incorrectly in color-independent ways, our color perceptions may

[43] That is, not merely on one's *judgment* of the color of the object. See, e.g. Hardin (1993: 104–6).

well deviate from the norm. So too do normative judgments take place in the context of a picture of the way the world is in non-normative aspects. Argument can alter this picture for us, and as a result, some of our normative responses may change. The effectiveness of argument is not problematic for my account. Rather, it is only one of the influences that explain the variability that contributes to the vagueness of normative terms.

Here again the analogy with color should be useful. For it can help us reconsider—and reject—the naïve view that in arguments over the applicability of a normative term, one of the disputants must be wrong. When two people have different phenomenal responses to a pair of green dress socks, there can be an argument that seems very genuine. The disputants will urge each other to reconsider in a better light, or from closer up, or in comparison to some less controversial paradigm. Here too both parties may be convinced that one and only one must be right. Nevertheless, a little knowledge of color science ought to dispel this naïve view. So too ought we discard the view that the possibility of argument over the applicability of a basic normative notion indicates anything very special about such terms that would prevent us from viewing them as I have suggested: as referring terms with a significant degree of vagueness.

3. Response-Dependence Leads to Relativism

A general worry about the view I put forward here is that it appears to conceal a kind of relativism. I do not mean the kind of speaker-relativism that suggested itself as an account of the faultless disagreement that occurs in the application of normative terms. Rather, the present worry would persist even if one granted that these faultless disagreements are the result of vagueness, and not of indexicality or some other form of relativity. Rather, because I hold that the responses of the vast majority of human beings somehow determine the extension of normative terms such as 'harm', there seems to be a kind of relativism to the current human population. Despite the size of the relevant group, the troubling aspects of relativism remain. For example, it seems I must hold that if an interstellar gas descends on Earth and changes our most basic affective attitudes, this will change what counts as harmful and beneficial.

This is an important worry, and it reappears in various guises. Whenever it arises, however, it gets its apparent force from a misconception of the relation between the affective responses of human beings, and the meanings of response-dependent normative terms. For it is a mistake to think that the view of objective normative terms on offer in this paper is one according to which facts about human desires, motivations, and the like, 'confer value'

on those things that possess it.[44] This would either assimilate such views to sophisticated versions of subjectivism or cultural relativism, or would invest human desire with an unexplained ground-level normativity. But rather than 'conferring value', the relevant affective responses, according to the response-dependent view here described, merely help explain the development of a referring term in the language that also has the features we think of as characteristically normative.[45] Thus there is no mysterious fount of normativity to be found in human wills. Nor is it true that normative claims are descriptive claims about human wills. Rather, just as agreement in phenomenal color responses facilitates the teaching of color words, agreement in affective responses facilitates the teaching of normative terms that then refer—directly—to a property of the things that elicit that response.[46] This sort of talk may seem to some to commit me to an unrealistically strong form of realism about the normative properties to be found at the more cognitive end of the spectrum. After all, I have just said that such properties play a causal role in eliciting relevant responses, so they must be there independently of such responses. But while I do indeed accept this sort of talk—indeed, I take it at face value—I do not think the realism it commits me to is very problematic. Consider the corresponding claims in the realm of color: it is the yellowness of a lemon that typically causes the characteristic phenomenal experience I have when I look at it. We have a word for this color property because the structure of our phenomenal color space is as it is, and is sufficiently uniform to allow the teaching of color words. But to say that the lemon is yellow is not to say anything about human beings or their visual equipment—much less about a uniformity in human visual responses. The lemon would have had the property picked out by the word 'yellow' even if human beings had never existed. We cannot change the colors of objects by changing human beings or human language: we must change the things themselves. Nor does this strong form of realism undermine the appropriateness of the label 'response-dependent' for the account of color suggested in this paper, for it remains true that the reason we have a word with the extension of

[44] Jackson (1998: 157) characterizes response-dependent views in this way. See also Koons (2003: 276).

[45] Of course these affective responses are not *by themselves* sufficient to explain the development of normative terms, any more than agreement in phenomenal responses is sufficient to explain the development of color terms. In both cases such terms must also serve some useful purpose: allowing us to describe the world in useful ways ('The red berries are poisonous'), or to indicate something about our own state that other people may be interested in knowing ('Jay Leno is funny').

[46] I should therefore be taken to be in agreement with David Wiggins (1998: 188–9) in his view that views of the sort I am calling 'response dependent' need not be seen as attempting to provide reductive analyses.

the word 'yellow' has everything to do with typical human responses being what they are.

Suppose now that the terms 'harm' and 'irrational' are referring words, explained as I have suggested, and that 'irrational' is a remainder term, much like 'color-blind'. This would explain why the following claim is true: It is irrational to fail to be averse to harms. Now consider the case in which an interstellar gas changes our basic affective responses so that we are no longer averse to some of the things we currently call 'harms'. The appropriate description of this event is that the interstellar gas has driven everyone mad (has made everyone irrational), not that it has changed what counts as a harm. My account can allow us to say this, simply by rigidifying the reference of 'harm' and 'irrational'.[47] Of course rigidifying in this way may seem an ad hoc maneuver. Let me now explain why I do not think that it is.

I have already mentioned one interesting difference between color terms and normative terms. While we have no direct perception of the microphysical surface properties that elicit our phenomenal color responses, we are independently aware of the features of situations that elicit our affective responses. For example, when we are averse to the prospect of pain, we are aware that it is the prospect of pain to which we are averse. In the case of color, however, when a certain object elicits a phenomenal response of a certain sort, we typically have no idea what properties of that object are responsible for this. Now, though I do not endorse it, I do not deny the plausibility of the view that color words refer rigidly to certain microphysical properties of objects. But the fact that there are distinct possible worlds in which distinct microphysical properties elicit the very same phenomenal response means that at least most of us have no idea which world we inhabit. Even if the English word 'red' rigidly designates the property of having a surface spectral reflectance in reflectance-class μ, we should admit that there is a possible world in which people superficially like us use the word 'red' to designate quite a different reflectance property, although still a property that, in *that* world, is associated with phenomenal responses just like the responses we associate with objects with reflectances that fall in class μ. Because we do not know which world we

[47] Talk of rigid designation for predicates cannot be taken too strictly on analogy with similar talk regarding particulars. I take the claim that a predicate rigidly designates to indicate that its extension is not determined by the descriptions that competent speakers might associate with it, but by the most general description, of the appropriate sort (physical, psychological, biological), that *actually* picks out a large enough proportion of the sample used to fix the reference of the term. For a brief and illuminating discussion of the problems of uncritically extending talk of rigidity from particulars to kinds, see Soames (2003: 423–56).

inhabit, there is something persuasive in the idea that 'red' might mean something like '*whatever* property typically elicits phenomenal responses like *this*'.[48] This would make 'red' a non-rigid designator. But in the case of a term like 'harm', we *do* know which world we inhabit: it is one in which death and pain, among other things, are designated by 'harm'. Because of this, there is less temptation to think of 'harm' as a non-rigid designator that is equivalent to 'whatever property typically elicits aversion'. Rather, it is more plausible that 'harm' rigidly refers to the things used to teach the term. After all, we are *aware* of those things when we learn the term, and it is our responses to such things that make other harms seem sufficiently similar that we spontaneously apply the word 'harm' to them as well. Moreover, we can use non-actual scenarios as part of the process of teaching and learning such normative terms. This also strongly suggests that it is our *actual* responses that determine the referents of normative terms even in counterfactual situations in which we have different responses: that is, it strongly suggests that normative terms rigidly designate.

Now, even if we grant that 'harm' is a rigid designator, something like a worry about relativism remains. Again the analogy with color is useful. Suppose, for the sake of argument, that color words rigidly designate reflectance classes. And consider a world in which the color receptors in the human retina have spectrally shifted sensitivity profiles, although the rest of the human visual system remains the same. That is, while the three actual human color receptors have peak sensitivities at X, Y, and Z, the three receptors of human beings in this world have peak sensitivities at $X + 30$, $Y + 30$, and $Z + 30$. What this means is that these people have the same subjective color space as we do, but classify particular objects differently than we do.[49] Suppose that they develop the same color vocabulary that we have. Then while we would classify monochromatic light of wavelength 580 as paradigmatically yellow, these people would use this phrase to describe monochromatic light of wavelength 610. But this does not mean that we should regard the inhabitants of that world as making systematic *errors* about the colors of objects. For when they say 'this

[48] Of course there are huge problems in speaking so freely about intersubjective similarities and difference between phenomenal experiences in the absence of—or indeed in the presence of—behavioral differences. The current point could be made without this implicit appeal to qualia, but to do so would obscure what is essential: that our epistemic position with regard to the microphysical bases of colors makes a non-rigid interpretation of color words at least plausible.

[49] When people in this world look at the very same spectrum cast on a white wall by a prism—one that we would claim was 10 centimeters in length—they see exactly what we see, but shifted one centimeter to the right. Thus we will classify the color of a given illuminated point on the wall differently.

object is yellow', they are not expressing the same belief we would express with the same utterance. Rather, the truth condition for their utterance of 'this object is yellow' is that the surface of the object has a reflectance that falls in class θ, while the truth condition for our utterance of the phonetically similar sentence is that the surface has a reflectance that falls in class ρ.

Consider now a world in which people are uniformly averse to quite different things than we are, and in which they are not averse to such things as future pain, loss of freedom, and so on. And imagine that they have a pair of words, 'harm' and 'irrational', which are learned via the same ostensive methods that I have claimed facilitate our own learning of the words 'harm' and 'irrational'. Despite the way these people use the terms 'harm' and 'irrational', *we* can still say—using our own words with their standard meanings—that they are not averse to harms, and that because of this they often act irrationally. So far so good. But shouldn't we also say that they make no error in classifying their odd collection of things using their word 'harm'? Shouldn't we simply say that they are applying a distinct normative concept? Just as we should have said that neither of us had the *right* color concepts, shouldn't we be equally evenhanded in our assessment of their normative concepts? The answer is that whatever we say in answer to this semantic question, our own responses to their behavior will and should remain the same, just as our own phenomenal response to objects whose surface reflectance falls in class θ remains an experience of blue, and not of red. Because of this, our practical attitude towards people in the 'crazy' world will be the same as our attitude towards massively irrational and self-destructive people in our world. Whether we also see them as making a *semantic* error in failing to apply their word 'harm' to future pain is neither here nor there.

I have said that this worry about relativism can appear in different guises. Here is another version of the same worry, which may not initially seem to rest on the same underlying mistake. On my view the extension of 'irrational' is determined by the affective responses of the vast majority of human beings. Suppose, however, that I am not a member of this vast majority. What weight should the claim that my proposed course of action counts as irrational have with me? After all, it simply indicates that the vast majority of other people have a certain attitude towards it. But who are they to me? The answer to this question is that it again wrongly assumes that my view is that irrational actions are *made* irrational by the fact that the vast majority of people have a certain attitude towards them. But this view, which invests the attitudes of human beings with a mysterious normative power, is one I emphatically reject. Rather, just as the objective nature of color words makes it the case that my visual

response counts as *defective* if I see something as green when it is really yellow, so too does the objective nature of the word 'harm' make it the case that my affective response counts as *defective* (our special word for this kind of defect is 'irrationality') if I am not averse to something that is a harm. Of course merely being informed that my phenomenal color response is defective is unlikely to correct it. Similarly, someone who is not averse to a certain harm, and who therefore acts irrationally, is unlikely to be persuaded to act differently (*cured*, one might say) merely by being informed that his action is irrational, much less by being informed of the semantics of the terms 'harm' or 'irrational'. But that is no criticism of the view.

4. The Whole Paper is a Just-So Story

A final objection I anticipate to the whole strategy of this paper is that it amounts to little more than a speculative history of language development and similar speculation about the processes of language acquisition within a language that is already a going concern. I have three responses to this charge. The first is simply to point out that I have a good number of partners in at least a similar crime. Saul Kripke (1980: 134–9), Hilary Putnam (1975), and David Lewis (1997), to give three very prominent names, engage in essentially the same practice. Kripke's talk of initial baptism, for example, and reference to the causal-historical processes by which the reference fixed by such a baptism continues to be passed along, is the same sort of speculative linguistic history as that which I engage in. Putnam's discussion of the linguistic division of labor also belongs to this category. And Lewis's explanation of the naming of the colors also endorses a certain hypothesis about language acquisition. Beyond these three examples, it may also be worth mentioning that theories of proper function such as that advocated by Ruth Millikan make the function of, say, an organ, dependent on contingent historical processes that we will never witness. It remains plausible, though, that such a theory counts as offering an explanation of the fact that it is the function of the heart to pump blood. The role of nearly universal agreement in phenomenal color response in the teaching of color terms seems hardly more controversial.

Of course it remains possible that Kripke, Putnam, Lewis, and I are all of us telling unjustified and pointless just-so stories. So the second response to the charge that I am merely engaging in armchair historical linguistics is that there is a philosophical point to be gained merely by pointing out the existence of a certain just-so story. For example, the story told about harms and rationality here explains how it could be that

there are naturalistic properties that have the apparently 'queer' property of to-be-avoidedness built into them: for on the story told, it is explicable why it is necessarily true that all but irrational people will avoid harms, even though 'harm' is as naturalistic a concept as 'blue'. In that sense, what the just-so story does is help to dispel the worry that one is cheating in describing the semantics for one's target terms in a certain way. For example, Paul Horwich once claimed that the embedding problem for expressivism is no problem at all. We can simply say, according to Horwich (1994), that normative terms serve essentially to express non-cognitive attitudes, but *also* claim that they function like descriptive predicates in the context of inferences. James Dreier (1996) reasonably objected to the idea that one could so easily solve the embedding problem: one needs to show how the semantics could *possibly* work out in the correct way. But in any case, the point of the paper was not to argue that 'harmful' and 'beneficial' are referring terms. Rather, the point was that the difference between a cognitivist and an expressivist account of a term, the origin of which lies in certain characteristic human responses, may best be thought of as a difference in degree, which itself depends on a difference in the degree of uniformity of that response across human beings.

The third response is that though I am indeed relying on some empirical hypotheses about language acquisition, the hypotheses are very uncontroversial. With regard to how we learn the meanings of words like 'red', 'round', 'harmful', and so on—that is, with regard to how we learn to use these words correctly—surely the two most plausible mechanisms are the following. First, we are explicitly trained by our parents and other linguistic authorities, using ostensive teaching, and are corrected when we go wrong, until we get things right. Second, we passively observe the linguistic behavior of those around us, and pick up on the regularities by means of a kind of innate hypothetico-deductive process in which many of the parameters have been set to optimize learning. However these two mechanisms, together or separately, explain our acquisition of the relevant bits of language, the story I have been telling remains plausible. Of course it is possible to avoid any contentious commitment to even these relatively uncontroversial hypotheses about language acquisition by completely ignoring the question of how we learn the meanings of words. Much of contemporary ethical theory adopts this strategy. The danger of such a strategy, however, is a liability to think of language speakers as appearing out of nowhere, and this carries with it the danger of ignoring the constraints on possible semantics that are placed there by the fact that we—somehow or other—have to *learn* the meanings of the words we use.

REFERENCES

Blackburn, Simon (1991a) 'Just Causes', *Philosophical Studies*, 61: 3–16.
____ (1991b) 'Reply to Sturgeon', *Philosophical Studies*, 61: 39–42.
____ (1993) 'Errors and the Phenomenology of Value', in *Essays in Quasi-Realism* (New York: Oxford University Press).
Byrne, Alex and Hilbert, David (2003) 'Color Realism and Color Science', *Behavioral and Brain Sciences*, 26: 3–21, 52–63.
Campbell, John (1997) 'A Simple View of Color', in A. Byrne and D. Hilbert (eds.), *Readings on Color, Volume 1: The Philosophy of Color* (Cambridge, Mass.: MIT Press).
Cohen, Jonathan (2004) 'Color Properties and Color Ascriptions: A Relationalist Manifesto', *Philosophical Review*, 113: 451–506.
Dreier, James (1990) 'Internalism and Speaker Relativism', *Ethics*, 101: 6–25.
____ (1996) 'Expressivist Embeddings and Minimalist Truth', *Philosophical Studies*, 83: 29–51.
Foot, Philippa (1978a) 'Moral Beliefs', in *Virtues and Vices* (Oxford: Basil Blackwell).
____ (1978b) 'Goodness and Choice', in *Virtues and Vices* (Oxford: Basil Blackwell).
Geach, Peter (1958) 'Imperative and Deontic Logic', *Analysis*, 18: 49–56.
____ (1965) 'Assertion', *Philosophical Review*, 74: 449–65.
Gert, Joshua (2004) *Brute Rationality* (Cambridge: Cambridge University Press).
____ (2006) 'Problems for Moral Twin Earth Arguments', *Synthese*, 150: 171–83.
____ (forthcoming) 'A Realistic Color Realism', *Australasian Journal of Philosophy*.
Hardin, C. L. (1993) *Color for Philosophers: Unweaving the Rainbow*, expanded edition (Indianapolis: Hackett Publishing Co.).
Holland, Sean (2001) 'Dispositional Theories of Value Meet Moral Twin Earth', *American Philosophical Quarterly*, 38: 177–95.
Horgan, Terence, and Timmons, Mark (1991) 'New Wave Moral Realism Meets Moral Twin Earth', in J. Heil (ed.), *Rationality, Morality, and Self-Interest* (Lanham, Md.: Rowman and Littlefield, 1991).
Horwich, Paul (1994) 'The Essence of Expressivism', *Analysis*, 54: 19–20.
Hume, David (1975) *Enquiries Concerning Human Understanding and Concerning the Principles of Morals*, 3rd edition, ed. L.A. Selby-Bigge and P. H. Nidditch (Oxford: Clarendon Press).
____ (1978) *A Treatise of Human Nature*, 2nd edn, ed. L. A. Selby-Bigge and P. H. Nidditch (Oxford: Clarendon Press).
Jackson, Frank (1998) *From Metaphysics to Ethics* (Oxford: Oxford University Press).
____ and Pargetter, Robert (1997) 'An Objectivist's Guide to Subjectivism about Colour', in A. Byrne and D. Hilbert (eds.), *Readings on Color, Volume 1: The Philosophy of Color* (Cambridge, Mass.: MIT Press).
Johnston, Mark (1989) 'Dispositional Theories of Value', *Aristotelian Society Supplement*, 63: 139–74.
____ (1997) 'How to Speak of the Colors', in A. Byrne and D. Hilbert (eds.), *Readings on Color, Volume 1: The Philosophy of Color* (Cambridge, Mass.: MIT Press).

Koons, Jeremy (2003) 'Why Response-Dependence Theories of Morality are False', *Ethical Theory and Moral Practice*, 6: 275–94.
Kripke, Saul (1980) *Naming and Necessity* (Cambridge, Mass.: Harvard University Press).
Leon, Mark (2002) 'Colour Wars: Dividing the Spoils', *Philosophy*, 77: 175–92.
Lewis, David (1997) 'Naming the Colors', *Australasian Journal of Philosophy*, 75: 325–42.
McDowell, John (1985) 'Values and Secondary Qualities', in T. Honderich (ed.), *Morality and Objectivity* (London: Routledge and Kegan Paul).
McGinn, Colin (1996) 'Another Look at Color', *Journal of Philosophy*, 97: 537–53.
McLaughlin, Brian (2003) 'The Place of Colour in Nature', in D. Heyer and R. Mausfeld (eds.), *Colour Perception: Mind and the Physical World* (Oxford: Oxford University Press).
Moore, G. E. (1922) 'The Conception of Intrinsic Value', in *Philosophical Studies* (London: Routledge and Kegan Paul).
____ (1968) 'Reply to My Critics', in P. A. Schilpp (ed.), *The Library of Living Philosophers, Vol. IV: The Philosophy of G. E. Moore*, 3rd edition (La Salle, Ill.: Open Court).
____ (1993) *Principia Ethica*, 2nd edn, ed. T. Baldwin (Cambridge: Cambridge University Press).
Putnam, Hilary (1975) 'The Meaning of "Meaning" ', *Philosophical Papers: Volume 2* (Cambridge: Cambridge University Press).
Raffman, Diana (1996) 'Vagueness and Context-Relativity', *Philosophical Studies*, 81: 175–92.
Railton, Peter (1998) 'Red, Bitter, Good', *European Review of Philosophy 3: Response-Dependence*, 67–84.
Scanlon, T. M. (1998) *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press).
Searle, John (1962) 'Meaning and Speech Acts', *Philosophical Review*, 71: 423–32.
____ (1969) *Speech Acts* (Cambridge: Cambridge University Press).
Smith, Michael (1993) 'Objectivity and Moral Realism: On the Significance of the Phenomenology of Moral Experience', in J. Haldane and C. Wright (eds.), *Reality, Representation and Projection* (Oxford: Oxford University Press).
Smith, Nicholas (2005) 'Vagueness as Closeness', *Australasian Journal of Philosophy*, 83: 157–83.
Soames, Scott (2003) *Philosophical Analysis in the Twentieth Century: Volume 2* (Princeton: Princeton University Press).
Sturgeon, Nicholas (1991) 'Contents and Causes', *Philosophical Studies*, 61: 19–37.
____ (1994) 'Moral Disagreement and Moral Relativism', *Social Philosophy and Policy*, 11: 80–114.
Wiggins, David (1998) 'A Sensible Subjectivism' in *Needs, Values, Truth*, 3rd edn (New York: Oxford University Press).

5

Moral Obligation and Accountability

Stephen Darwall

Philosophers typically characterize morality's distinctive normativity in terms of the categorical character and overriding weight or absolute priority of the reasons for acting that a moral obligation purports, at least, to provide. Attempts within the Kantian tradition to vindicate moral obligation in these terms generally proceed from the practical standpoint; they argue that an agent is committed by constraints of rational deliberative thought to treating the moral law as supremely authoritative. These attempts are almost universally judged to fail, however, except by their most enthusiastic proponents.

I believe that they all fail for a common reason, namely, because it is impossible to establish the supreme authority of moral obligation from a purely first-person point of view. In what follows, however, I shall be concerned less to present this diagnosis than to argue that even were the arguments to succeed in their own terms, they would not yet account for a central aspect of moral obligation. The reason is that moral obligation's normativity essentially includes an irreducibly second-personal element. Moral obligations do not simply purport to provide supremely authoritative reasons. They are also what we are *responsible to* one another for doing, what members of the moral community have the authority as such to demand that we do by holding us accountable second-personally. Even if an argument could show that moral obligations invariably provide overriding reasons of whatever weight or priority, it would not yet establish any responsibility *to* anyone for complying with them, since no authority to demand compliance would yet follow. There is simply no way, I believe, to establish accountability except within a second-personal framework.

Much of the material in this essay has been drawn from Darwall 2006. I am grateful to Harvard University Press for allowing it to be reprinted here. I am especially indebted to the participants at the Second Wisconsin Metaethics Workshop for their comments

Moreover, although I cannot present it here, I believe that an argument that moral obligations necessarily do provide supremely authoritative reasons can be made from a second-person perspective.[1] Whenever we address any claims and demands to anyone at all, we are committed to the assumption that we and they share a common competence and authority as free and rational agents and, therefore, to the validity of any demands that derive from this authority. As I say, I shall not present any argument for that claim here. My aim here will be to argue that the distinctive normativity of moral obligation includes an irreducibly second-personal aspect. I begin by displaying the second-personal character of moral responsibility and then argue that this transfers to moral obligation owing to a conceptual connection that exists between moral obligation and moral responsibility.

SECOND-PERSONAL REASONS

To get the flavor of the kind of point I shall be trying to make, compare two different ways in which you might try to give someone a reason to stop causing you pain, say, to remove his foot from on top of yours.

One would be to get him to feel sympathetic concern for you in your plight, thereby leading him to want you to be free of pain. Were he to have this desire, he would see your being in pain as a bad thing, a state of the world that there is reason for him (or, indeed, for anyone who can) to change. And he would most naturally see his desire that you be pain free, not as the source of this reason, but as a form of epistemic access to a reason that is there anyway.[2] *In* desiring that you be free of pain, it would seem to him that this would be a better way for the world to be, that it is a possible outcome or state that, as Moore put it, 'ought to exist for its own sake' (Moore 1993: 34). Were he to credit the way things seem from the perspective of his desire, he would accept an *agent-neutral* (and *state-of-the-world-regarding*) reason for removing his foot.[3] The reason would not be essentially *for him* as the agent causing another person pain.

[1] I try to make this argument in Darwall 2006.

[2] On this point, see Darwall 1983; Bond 1983; Pettit and Smith 1990; Quinn 1991; Hampton 1998; Scanlon 1998: 41–55; and Dancy 2000.

[3] Agent-neutral reasons contrast with agent-relative reasons, those whose formulation includes an ineliminable reference to the agent for whom they are reasons (like 'that it will keep a promise I made,' 'that it will avoid harm to others (i.e., people other than me,' and so on). Agent-neutral reasons can be stated without such a reference: 'that it would prevent some pain from occurring to someone (or some being).' On the distinction between agent-relative (also called 'subjective' or 'agent-centered') and agent-neutral (also called 'objective') reasons, principles, values, etc., see Nagel 1970; Scheffler 1982; Parfit

It would exist, most fundamentally, for anyone who is in a position to effect the state of the relief of your pain and *therefore* for him, since he is well placed to do so.[4] Finally, in 'giving' him the reason in this way, you wouldn't so much be *addressing* it to him, as getting him to see that it is there anyway, independently of your getting him to see it or even of your ability or competence to do so.[5]

Alternatively, you might lay a claim or address a purportedly valid demand. You might say something that asserts or implies your authority to claim or demand that he move his foot and that simultaneously expresses this demand. You might demand this as the person whose foot he is stepping on, or as a member of the moral community, whose members understand themselves as holding one another to a (moral) demand not to step on each other's feet, or as both. Whichever, the reason you would address would be agent-relative rather than agent-neutral. It would concern, most fundamentally, his relations to others (and himself), viewed from within those relations, in this case, that his keeping his foot on yours causes another person pain, causes inconvenience, and so on. The reason would not be addressed to him as someone who is simply in a position to alter a bad state, whether of someone's pain or even of someone's causing another pain. If he could stop, say, two others from causing an identical gratuitous pain by the shocking spectacle of keeping his foot firmly planted on yours, this second, claim-based (hence second-personal) reason would not recommend that he do so. The reason would be addressed to him as someone *causing* gratuitous pain to another person, something we persons normally assume we have the authority to demand that we not do to one another.

What is important for our purposes is that someone can sensibly accept this second reason for moving his foot, one embodied in your claim or demand, only if he also accepts your *authority to demand* this of him

1984; Nagel 1986; and McNaughton and Rawlings 1991. For a discussion that raises a question about the value of this distinction, see Korsgaard 1996a.

I argue for the claim that sympathetic concern involves its seeming that there are agent-neutral reasons to further someone's welfare in Darwall 2002: 68–72. I do not deny, of course, that someone who already accepted various agent-relative norms might not be moved through empathy and sympathy, to feel some special responsibility for relieving the pain. My point is that this would not come through sympathy alone.

[4] Roughly speaking, again, a reason is agent-neutral if it can be formulated without essential reference to the agent (as such); otherwise it is agent-relative. It should also be noted that superficially agent-relative reasons may be grounded more deeply in agent-neutral considerations and values, and/or vice versa. For example, rule-utilitarianism holds that rules of right conduct include agent-relative principles, for example, those defining rights of promise and contract, on grounds of overall agent-neutral value.

[5] Just as might be the case if you were trying to get him to see reasons to believe that you were in pain. A grimace might suffice without your having to presume any authority on the question.

(second-personally). That is just what it is to accept something *as a valid claim or demand*. And if he accepts that you can demand that he move his foot, he must also accept that you will have grounds for complaint or some other form of accountability-seeking response if he does not. Unlike the first reason, this latter is second-personal in the sense that, although the first is conceptually independent of the second-personal address involved in making claims and holding people responsible, the second is not.[6] A *second-personal reason* is thus one whose validity depends upon presupposed authority and accountability relations between persons and, therefore, on the possibility of the reason's being addressed person-to-person. Reasons of this kind simply wouldn't exist but for their role in second-personal address, and their second-personal character explains their agent-relativity. As second-personal reasons always derive from agents' *relations* to one another, they are invariably fundamentally agent-relative.[7]

Of course, there could be agent-relative norms and reasons constraining our conduct toward one another that are not second-personal. Someone might accept an agent-relative norm of conduct without thinking that this is something anyone has any standing to hold him to. For example, we might think of the feet of persons as something like sacred ground and hence that we all have reason, even a supremely authoritative reason, to avoid stepping on, without supposing that this has anything to do with anyone's authority to demand this, even God's. Once, however, we have the idea that there exists a reason to forbear stepping on people's feet in the fact that this is something we can or do reasonably demand of one another, or that we are *accountable* for this forbearance, we have the idea

[6] Here and elsewhere, by 'second-personal,' I refer to thought and speech that explicitly or implicitly addresses claims, requests, demands, and so on, to an addressee and that presumes some authority or standing to do so. Thought and speech that is second-personal in this sense is also first-personal since address must always be from a first-person standpoint, whether singular or plural. Not all first-personal thought and speech involves address in the sense to which I am referring, however. Even some that involves address in a broad sense—'Hey! Stop that!'—may not address a purportedly valid claim and so not be 'second-personal' in the current sense. It may just be an attempt to get someone to do something. It is important also that one can take up a second-personal relation to oneself. I shall argue, for example, that the experience of guilt standardly involves implicitly making a demand of oneself as a member of the moral community.

[7] The formulation of the reason may not always be agent-relative, however. Suppose, for example, that the best way of grounding the Categorical Imperative is, as I believe, in an equal authority to make claims and demands that persons presuppose when they address one another second-personally. It is at least conceivable that what the categorical imperative itself requires is a principle of conduct that can be specified agent-neutrally. R. M. Hare, for example, believes that the categorical imperative can be seen to entail the sort of universal prescriptivism he favors *and* that this entails a form of act-utilitarianism (an agent-neutral theory). See Hare 1993.

of a second-personal reason—a kind of reason that wouldn't have existed but for the possibility of the second-personal address involved in claiming or demanding.

A CIRCLE OF IRREDUCIBLY SECOND-PERSONAL CONCEPTS

Second-personal reasons are conceptually tied to a distinctively second-personal kind of *practical authority*: the authority to make a demand or claim. Conversely, making a claim or putting forward a demand as valid always presupposes the authority to make it and that the duly authorized claim creates a distinctive reason for compliance (a second-personal reason). Moreover, these notions all also involve the idea of responsibility or accountability. The authority to demand implies, not just a reason for the addressee to comply of whatever weight or priority, but also his being responsible *to* the addresser for compliance.[8] Conversely, accountability implies the authority to hold accountable, which implies the authority to claim or demand, which is the standing to address second-personal reasons. These four notions—second-personal authority, valid claim or demand, second-personal reason, and responsibility or accountability *to*—thus comprise an interdefinable circle. Each idea implies the other three. And I contend that there is no way to break into this circle from the outside. Propositions formulated only with normative and evaluative concepts that are not already implicitly second-personal cannot adequately ground propositions formulated with concepts within the circle.

There is, consequently, an important difference between the idea of an authoritative (second-personal) claim or demand, on the one hand, and that of an authoritative or binding norm or normative reason, or even of a normative requirement, on the other. There can be requirements *on* us that no one has any standing *to require of* us. We are under a requirement of reason not to believe propositions that contradict the logical consequences of known premisses, for example. But it is only in certain contexts, say, when you and I are trying to work out what to believe together, that we have any standing to demand that one another reason logically, and even here that authority apparently derives from a moral or quasi-moral aspect: our having undertaken a common goal.[9] Requirements of logical reasoning are, in this

[8] Thus, Michael Dummett remarks that the right to command means that 'the *right* to reproach is an automatic consequence of disobedience' (Dummett 1990: 9).

[9] Of course, these further constraints are frequently in the background, as they are, for example, whenever we do philosophy, say, right now. Because of the relationship

way, fundamentally different from moral requirements. I will argue that it is part of the very idea of moral obligation that moral requirements are what those to whom we are morally responsible have the authority to demand that we do. Clearly this is no part whatsoever of the concept of a demand of logic or requirement of reason.[10] It will follow that the normativity of moral obligation cannot be fully captured in the weight or authority of practical reasons, even if these amount to rational requirements.[11]

THE SECOND-PERSONAL CHARACTER OF MORAL RESPONSIBILITY

In his famous essay, 'Freedom and Resentment,' P. F. Strawson argued influentially against consequentialist accounts of moral responsibility that social desirability cannot provide a justification of 'the right *sort*' for practices of moral responsibility 'as we understand them' (1968: 74). When we seek to hold people accountable, what matters is not whether punishment is desirable, either in a particular case or in general, but whether it is deserved and the authority exists to mete it out. Desirability is a reason of the wrong kind to warrant the attitudes and actions in which holding someone responsible consists in their own terms.

Strawson's point is an instance of a more general phenomenon that can be called the *wrong kind of reason problem* (Rabinowicz and Ronnøw-Rasmussen 2004; Olson 2004; Hieronymi 2005; see also Parfit 2000). For example, there might be pragmatic reasons to believe some proposition, but that doesn't make that proposition *credible*. It doesn't justify believing it in terms of reasons and norms that distinctively apply to belief. Similarly, as D'Arms and Jacobson have pointed out, it is a 'moralistic fallacy' to conclude from the fact that being amused by a certain joke is morally objectionable that the joke is not funny.[12] The former is a reason of the

you and I are currently in, each of us *does* have authority to call one another to account for logical errors, a standing that, without some such context, we lack. But however frequently that or some relevantly similar context obtains, the authority comes, not just from the requirement of reason, but from some other presupposed feature of the context.

[10] I am indebted to Peter Graham for this way of putting the contrast.

[11] On this point, see also Frankfurt 2000.

[12] D'Arms and Jacobson argue that this poses a problem for response-dependent or, as they call them, 'neo-sentimentalist' accounts of various evaluative and normative notions, since it shows that, say, the funny can't be understood in terms of amusement's making sense or being warranted by just *any* reasons. There is a distinction between an emotion or attitude's being 'the right way to feel' and it's 'getting [the relevant value] right.' For an excellent discussion of how what they call 'fitting-attitude' (or 'FA') analyses can

'wrong kind' to justify the claim that a joke does not warrant amusement in the sense that is intrinsically relevant to whether it is funny or not (D'Arms and Jacobson 2000).

To be a reason of the right kind, a consideration must justify the relevant attitude in its own terms. It must be a fact about or feature of some object, appropriate consideration of which could provide someone's reason for a warranted attitude of that kind towards it.[13] It must be something on the basis of which someone could (and appropriately would) come to hold the attitude as a conclusion of a process of considering (deliberating about) *whether* to do so. In considering whether to believe some proposition *p*, for example, it is simply impossible to conclude one's deliberation in a belief that *p* by reflecting on the desirable consequences of believing *p*. That is a reason of the right kind for *desiring* to believe that *p*, but not for believing that *p* (as is shown by the fact that one can come to desire to believe *p* by reflecting on the desirable consequences of believing *p*, but one cannot believe *p* for that reason).[14] The *desirable* concerns norms and reasons that are specific to desire, and the *credible* concerns norms and reasons that are specific to belief.

Similarly, the *(morally) responsible* and the *culpable* concern norms for the distinctive attitudes and actions involved in holding people responsible and blaming them. The desirability—whether moral, social, personal, or otherwise—of holding them responsible or blaming them, or reasons why that would be desirable, are simply reasons of the wrong kind to warrant doing so in the sense that is relevant to whether they *are* morally responsible or blameworthy. The former concerns reasons and norms of desire (even if from the moral point of view), and what is thus desirable is simply a different question from whether we are justified in holding someone responsible or blaming them in the relevant sense. The latter concerns reasons and norms that are distinctively relevant to these latter attitudes.

Strawson dubbed the distinctive attitudes involved in holding people responsible 'reactive attitudes,' with prominent examples being indignation,

deal with the problem of distinguishing reasons of the right from reasons of the wrong kind, see Rabinowicz and Ronnøw-Rasmussen 2004. I am indebted to Julian Darwall for discussion of this general issue and to Joe Mendola for a question that helped me to see that Strawson's point is an instance of it.

[13] Rabinowicz and Ronnøw-Rasmussen put essentially the same point by saying reasons of the right kind also appear in the content of the attitude for which they are reasons: the attitude is toward something 'on account of' these reasons (Rabinowicz and Ronnøw-Rasmussen 2004: 414). As W. D. Falk pointed out, a favoring that is relevant to value is 'by way of true comprehension of what [the object] is like' (Falk 1986: 117).

[14] More accurately, it entails that reasons of the right kind exist for desiring to believe the proposition.

resentment, guilt, blame, and so on. And Strawson himself pointed out what some more recent commentators, notably Gary Watson and R. Jay Wallace, have also since stressed, namely, that reactive attitudes implicitly address *demands*. They invariably involve 'an *expectation of*, and *demand* for' certain conduct from one another (Strawson 1968: 85, emphasis added).[15] To feel a reactive attitude is to feel as though one has a warranted expectation *of* someone. Reactive attitudes must therefore presuppose the standing or authority *to* expect and hold one another responsible for complying with moral obligations (which just are the standards to which we can warrantedly hold each other as members of the moral community). But they also presuppose that those we hold accountable have that standing also. They address another in a way that 'continu[es] to view him as a member of the moral community; only as one who has offended against its demands' (Strawson 1968: 93). It follows that reactive attitudes are second-personal in our sense.

Consider guilt, for example. To feel guilty is to feel as if one is appropriately blamed and held responsible for something one has done.[16] Guilt feels like the appropriate (second-personal) *response* to blame: an *acknowledgment* of one's blameworthiness that recognizes both the grounds of blame and, more importantly for us, the authority to level it (even if only 'to God'). To feel guilt, consequently, is to feel as if one has the requisite capacity and standing to be addressed as responsible. Moreover, guilt's natural expressions are themselves second-personal—confession, apology, making amends, giving future assurances, self-addressed reproach, and so on.

Or consider resentment. Resentment is felt as if in response to a violation of a legitimate claim or expectation, and not simply as *directed* toward the violator, but as implicitly *addressing* her. It is a form of 'holding responsible,' an address of the other as a person with the capacity and standing to be addressed in this way and charged. If it turns out, for example, that someone's foot has been forced on top of yours by the shifting of a heavy package on a careening bus on which you both are traveling, knowing that might not change your desire to get his foot off of yours, but it will lessen your resentment or perhaps redirect it to a new object (the driver).

[15] Gary Watson stresses this in Watson 1987: 263, 264. Note also, R. Jay Wallace: 'there is an essential connection between the reactive attitudes and a distinctive form of evaluation … that I refer to as holding a person to an expectation (or demand)' (Wallace 1994: 19). See also Bennett 1980 and Scanlon (1998: 272–90).

[16] On this point, see Greenspan 1992. For other elements of the contrast between guilt and shame, see Williams 1993: 89–90; also Morris 1976, Wollheim 1984, Rawls 1971, sects. 67, 70–5, and Gibbard 1990, ch. 7. For a contrasting view, see Stocker, forthcoming.

Or finally, think of blame or indignation. Moral indignation involves the feeling that someone is rightly held responsible for some conduct and is itself part of holding him thus accountable. Moreover, and this is part of Strawson's point, indignation differs from its seeming that a sanction would be desirable, or even that some evil's befalling him would make for a more valuable, balanced, or fitting whole (poetic justice, say). The feeling of indignation invariably includes some sense of authoritative demand that may be absent from the feeling that something would be desirable or fitting. Consider what beliefs moderate or undermine indignation. If we come to believe that someone does not deserve blame, say, because he could not possibly have known the true character of what he was doing or because he was under extreme duress, then this will reduce or even defeat our indignation towards him. But if we learn that attempting to hold him accountable would be undesirable, say, because it will provoke him further, this will evidently not undermine indignation. To the contrary. As Strawson points out, we feel indignation and disapprobation when we feel we can demand, as members of the moral community, that people act in certain ways. Indeed, Strawson says, 'the making of the demand *is* the proneness to such attitudes' (1968: 92–3). We address moral demands partly by its being common knowledge that we are prone to second-personal 'demanding' attitudes and to more explicit ways of holding one another responsible.

Holding people responsible and blaming them is a way of relating to them that addresses them second-personally (if only implicitly) and so presupposes the authority to do so. It follows that the idea of culpability and the sense of 'morally responsible' or 'accountable' that is related to it are second-personal concepts. Consequently, the fact that one would be blameworthy if one did something (without excuse) is a second-personal reason not to do it. Strawson's 'wrong kind of reasons' point is thus a specific instance of my claim that there is no way into the circle of interdefinable second-personal concepts from the outside.

I claim that reactive attitudes are always implicitly second-personal, and I shall argue presently that they therefore invariably carry presuppositions of second-personal address about the competence and authority of the individuals who are their targets, as well as about those who have them.[17]

[17] Michelle Mason makes what I take to be a similar claim, saying there is a sense in 'which it is true that all the reactive attitudes are in fact moral attitudes: namely, the sense in which it is true that to regard one as within the scope of the particular reactive attitude is to regard one as answerable to an expectation or demand that forms part of a system of expectations, demands and rights the regulation in accordance with which it is necessary for aspiring to moral community with us' (Mason 2003: 244). Mason makes a persuasive case that *contempt* should also be understood as a reactive attitude.

Personal reactive attitudes are felt as if from the second-person standpoint of a relevant transagent, and impersonal reactive attitudes are felt as from the standpoint of members of the moral community. Even when one blames oneself, as in guilt, one takes a second-person standpoint on oneself; one implicitly addresses oneself as from the perspective of a member of the moral community.

PRESUPPOSING SECOND-PERSONAL COMPETENCE AND AUTHORITY

What gives Strawson's discussion of reactive attitudes its special relevance to the issue of free will is that reactive attitudes invariably address demands, and, as Gary Watson notes, there are 'constraints on moral address' that must be presupposed as felicity conditions of addressing a demand (Watson 1987: 263, 264). 'To be intelligible,' Watson points out, 'demanding requires understanding on the part of the object of the demand' (Watson 1987: 264). The point is not that making a demand is unlikely to be *effective* unless its object has the capacity to understand it. It is rather that reactive attitudes are 'forms of communication' that are simply unintelligible in their own terms without the presupposition that their objects can understand what is being said and act on this understanding.[18] The point is an Austinian one about the felicity conditions of a speech or quasi-speech act. Even if expressing reactive attitudes to those who lack the requisite capacity, like very young children or the insane, causes them to behave desirably, reactive attitudes there 'lose their point as forms of moral address' (Watson 1987: 265). The effectiveness of moral address is a matter of what Austin calls 'perlocutionary force' (consequences brought about by a speech act), whereas addressees' having (and being assumed to have) the capacity to recognize and act on second-personal reasons is, I am claiming, a felicity condition of moral address's having its distinctive 'illocutionary force' (that is, making it the distinctive speech act it is) (Austin 1975).

It is a particularly interesting case, since, as Mason argues, it presupposes a background demand on its object as a person, it may not seem to *address* the demand, since its natural expression is a form of withdrawal. I think, however, that contempt of the sort she is discussing must, if it is to be a reactive attitude as she claims, presuppose that the withdrawal is a way of holding its object accountable and not a non-'reactive' response as, say, disgust, or, indeed, other forms of contempt, seem to be.

[18] Watson remarks, as we noted above, that the communicative (second-personal) character of reactive attitudes does not mean that they are 'usually communicated; very often, in fact, they are not. Rather the most appropriate and direct expression of resentment is to address the other with a complaint and a demand' (1987: 265).

One need not *believe* that someone to whom one addresses a moral demand has the requisite capacity and standing. The point is rather that moral address presupposes these things. Watson is saying that we address others on the assumption that they can understand and be guided by what we are saying. And I am adding that we presuppose this in addressing demands and claims, and hence, second-personal reasons more generally. We presuppose *second-personal competence*, namely, that those we address can guide themselves by a recognition of the second-personal reasons we address and our authority to address them.

If you express resentment to someone for not moving his foot from on top of yours, you implicitly demand that he do so. And any second-personal reason you implicitly address presupposes, first, that he can recognize the validity of your demand and, second, that he can move his foot simply by recognizing a conclusive reason for acting that derives from your authoritative demand. And if I express indignation as a disinterested bystander, I too must make these assumptions. A putatively authoritative demand whose validity someone cannot recognize and act on is guaranteed to be infelicitous. The point is not, again, that such a demand cannot achieve compliance. It may well, but that is a matter of its perlocutionary force. It is that the address is guaranteed to fail in illocutionary terms, that is, as an addressing of an authoritative demand or second-personal reason.[19]

Claiming or demanding is not just calling some claim or demand to someone's attention. It is addressing a distinctively second-personal kind of reason to another person that aims to direct his will but in a way that recognizes his authority and independent practical reasoning. As Strawson emphasizes, to respond to another's conduct with a reactive attitude is 'to view him as a member of the moral community; only as one who offended against its demands' (Strawson 1968: 93).[20] Reactive attitudes are thus unlike critical attitudes of other forms, disdain, for example, that presuppose no authority on the part of their objects. The fact that the object of one's disdain cannot understand its basis or regulate his conduct by it need put no pressure on the disdain; to the contrary, it may seem to confirm it: he is so out of it that he can't even get it. I believe that the role of the second-person stance in mediating (mutual) accountability in Kantian and

[19] At this point, I am not so much arguing for this as claiming it. Part of the argument for the claim, of course, is the claim's role in an overall picture of second-personal address and reasons that I will hope will seem compelling and able to explain significant ethical phenomena.

[20] This is also, I believe, the grain of truth in Hegel's famous idea of a 'right to punishment,' that failure to hold someone accountable can be a failure to respect his dignity as a rational person (Hegel 1991: 126–7).

contractualist ethical conceptions marks a deep difference with the ethical views (frequently ethics of virtue) of thinkers like Plato, Aristotle, Hume, and Nietzsche (to give four prominent examples) for whom evaluation of conduct and character does not take a fundamentally second-personal form.[21] Central to the former conceptions is the idea of *morality as equal accountability*, that is, that morality essentially includes obligations we are responsible to one another for complying with. Of course, these conceptions need not hold that moral obligation is all morality is about, much less that it comprises all of ethics.

A consequence of all this is that we can intelligibly address demands through reactive attitudes only to those we assume able to take the very same attitudes toward themselves.[22] Addressees must be assumed to be able to take a second-person perspective on, and make the same demands of, themselves through acknowledging their validity as, for example, in self-reactive attitudes like guilt, and by appropriately regulating their practical reasoning. To address moral demands, we must presuppose *second-personal competence*.[23]

ACCOUNTABILITY AND THE METAETHICS OF MORAL OBLIGATION

I turn from exhibiting the second-personal character of moral responsibility to an argument that moral obligation is likewise irreducibly second-personal since it is tied to moral responsibility conceptually. It is a curious feature of the contemporary philosophical scene that, although Strawson's critique has been very influential in debates about responsibility and free will within moral psychology and the philosophy of action, its implications for the metaethics of moral obligation and for normative moral theory have been largely overlooked. In this section, I argue that there is a conceptual connection between moral obligation and moral responsibility,

[21] See Korsgaard 1996c for an excellent discussion of this aspect of the Kantian framework.

[22] Note that I am claiming that this is a presupposition of holding someone responsible in the sense of blaming him, finding his conduct *culpable*. We might think someone does wrong, however, even when he is in no position to recognize it, but we do not blame people for wrongs we think they cannot appreciate. We take their incapacity as an excuse. Nevertheless, I shall argue in the next section that there is a conceptual connection between being wrong and being blameworthy *if not adequately excused*.

[23] The moral competence requisite for (equal) membership in the moral community is what Rawls calls a 'range property.' In this sense, people are not more or less competent members of the moral community, since everyone who is within the range is equally within the range. On this point, see Rawls 1971: 508.

one Strawson himself implicitly relies upon. It follows, I argue, that moral obligation has an irreducibly second-personal aspect also and that the fact that something would violate a moral obligation is a second-personal reason not to do it. Moreover, although I do not pursue the point here, I believe that Strawson's influential critique of consequentialist accounts of moral responsibility can be turned into a powerful criticism of consequentialist theories of moral obligation, most obviously, of act-consequentialism, but arguably also of indirect consequentialist approaches such as rule-consequentialism, at least at the most fundamental level. When we reflect on obligation's intrinsic connection to (second-personal) accountability, we see that subserving an external goal is a reason of the wrong kind to justify moral obligation no less than it is to warrant claims of moral responsibility. Like moral responsibility, moral obligation is an irreducibly second-personal notion; to be the right kind of reasons to establish a moral obligation, therefore, considerations must have the requisite force from a second-person point of view.

One way to see this is to note that Strawson specifically includes a 'sense of obligation' as a reactive attitude *and* that he characterizes the skepticism he takes pragmatic approaches to responsibility to be responding to as holding that if determinism is true, 'then the concepts of moral obligation and responsibility really have no application' (Strawson 1968: 86, 71). As Strawson sees it, skepticism about free agency puts pressure on *both* moral responsibility and on moral obligation. Strawson doesn't say why this should be so, but it is clear enough that he must be taking moral obligation and responsibility to be related conceptually and not just as a matter of substantive normative judgment. What we are morally obligated to do, he seems to be thinking, is, as a matter of conceptual necessity, what members of the moral community can appropriately demand that we do, including by responding with blame or other reactive attitudes if we fail to comply without adequate excuse.

Perhaps the best-known invocation of this idea is, ironically enough, by a consequentialist thinker, namely, John Stuart Mill (Mill 1998). In the course of considering in Chapter V of *Utilitarianism* how a utilitarian might account for rights and justice, Mill provides a genealogy of conceptions of justice and concludes that 'the primitive element, in the formation of the notion of justice, was conformity to law' and that this involves the idea of warranted sanctions (Mill 1998: V, §12). Mill then adds that his conceptual analysis applies also to 'moral obligation in general.' And then he famously says:

> We do not call anything wrong, unless we mean to imply that a person ought to be punished in some way or other for doing it; if not by law, by the opinion of

his fellow-creatures; if not by opinion, by the reproaches of his own conscience. This seems the real turning point of the distinction between morality and simple expediency. It is a part of the notion of Duty in every one of its forms, that a person may rightfully be compelled to fulfil it. Duty is a thing which may be exacted from a person, as one exacts a debt. (Mill 1998: V, §14)

Mill seems to be on safe ground in saying that our concept of wrongdoing is essentially related to accountability. We do not impute wrongdoing unless we take ourselves to be in the range of the culpable, that is, the area in which we think the agent is apt for blame or some other form of accountability-seeking reactive attitude if she lacks an adequate excuse.[24]

This aspect of the concept of moral wrong has been stressed by a number of contemporary writers also.[25] Perhaps most striking is the role the connection plays in neo-Nietzschean critiques of morality and moral obligation, most prominently by Bernard Williams.[26] Williams's version of the Nietzschean critique that morality's conceptual system is an enslaving ideology, a form of false consciousness that shackles and sickens, runs through conceptual relations he sees holding between moral obligation, blame, and reasons for acting.[27] Williams evidently assumes that it is a conceptual truth about the morally obligatory that violations are appropriately blamed and that blaming implies the existence of good and

[24] The original meaning of 'impute' is relevant here: 'To bring (a fault or the like) into the reckoning against; to lay to the charge of; to attribute or assign as due or owing *to*.' *Oxford English Dictionary*, on-line edition.

[25] John Skorupski points out that calling an act 'morally wrong … amounts to blaming the agent' and maintains that the idea of moral wrong can't be understood independently of that of blameworthiness. (Skorupski 1999: 29, 142). Allan Gibbard quite explicitly follows Mill's lead in proposing that 'what a person does is *morally wrong* if and only if it is rational for him to feel guilty for having done it, and for others to be angry at him for having done it.' (Gibbard 1990: 42) And we can find versions of this Millian idea in other writers also (Baier 1966; Brandt 1979; Shafer-Landau 2003).

[26] Especially in Williams 1985. For a discussion, see Darwall 1987. See also Williams 1995 and Baier 1993. Nietzsche's diagnosis of morality 'in the pejorative sense' is primarily given in Nietzsche 1994. For useful discussion, see Leiter 1995 and 1997.

It is worth noting that, although Williams is a critic of what we might call the 'internal' aspects of second-personal accountability of the 'morality system,' he does embrace the idea of human rights and the 'external' forms necessary to enforce them. But this means, I believe, that participants in moral practices of enforcement, including all citizens when they participate in public discourse, are unable to accept reasons of the right kind for the second-personal demands through which they seek to enforce their rights. Their justification must consist in something like the desirability of using power, not any *authority* to use it. In my view, Hume's ideas about justice lead to the same result. Both run afoul of the wrong kind of reasons problem. I am indebted here to discussion with Simon Blackburn.

[27] Williams encourages the association with slavery himself by referring to morality as the 'peculiar institution' in the title of Chapter X of Williams 1985.

sufficient reasons to do what someone is blamed for not doing. The idea is not, of course, that normative reasons follow from the fact of someone's being blamed. Rather, *in* blaming, one implies or presupposes that there are such reasons. According to Williams, this presupposition is a bit of false consciousness. What makes it so, according to him, is his famous *internal reasons thesis* that all normative reasons for action must be anchored appropriately in the agent's own 'motivational set' (be an 'internal reason') and his claim that nothing guarantees any connection between what we take to license blame when we attempt to hold agents accountable and their own motivations (Williams 1995).

I believe that Williams is right about these conceptual connections between imputing wrong and blame, and between blame and attributing authoritative reasons. Moral obligation really is conceptually related to standards of minimally decent conduct that moral agents are accountable for complying with. And the forms of moral accountability—blame, guilt, indignation, punishment, and so on—really do imply that agents have conclusive reasons to do what they are morally obligated and accountable for doing, as we shall see presently.[28]

Nothing depends, of course, on whether we use the words 'wrong' and 'moral obligation' in the way Mill and these contemporary thinkers say we do. We could use these words more broadly to include moral ideals or goals. However, if we did, we would still need terms to refer to the idea to which these thinkers point, namely, the part of morality that concerns that for which we appropriately hold one another responsible. And it seems clear enough that, as all these writers agree, this involves a notion of moral demands, that is, of standards of conduct that the moral community has the authority to demand compliance with, including through second-personal forms of accountability of the sort we canvassed earlier. With this understanding, therefore, I shall henceforth use 'wrong' and 'moral obligation' in a Millian way as implying accountability-seeking demands.

Using 'wrong' in this way does not, we should note, require that there be an assignable victim who is wronged (hence, that what Mill regards as a *right* be in play), or even that violations of norms of a community of mutually accountable persons directly threaten the interests of such persons. It is consistent with the idea that wrongdoing is essentially tied to accountability, even accountability to other moral persons, that what we are accountable for can extend, for example, to the treatment of non-rational animals, aspects of the environment, and non-rational human beings.

[28] Again, this is implied or presupposed *in* holding people accountable. It is not implied by the fact that we hold them accountable.

MAKING MORAL OBLIGATION'S SECOND-PERSONALITY EXPLICIT

Debates in moral theory rarely tie moral obligation to second-personal accountability explicitly, but they often implicitly assume such a connection nonetheless. In this section, we shall illustrate this phenomenon in two familiar debates. One takes place within normative ethical theory between consequentialists and their critics over whether act-consequentialism is 'too demanding.' The other concerns morality's authority or, as it is sometimes put, 'Why be moral?' In both cases, analysis of what most deeply underlies the debate reveals an assumption that moral obligation and standards of right and wrong are conceptually related to what the moral community, and we as members, can demand (second-personally).

Take the 'too demanding' criticism first. Act-consequentialism's critics sometimes concede, *arguendo*, that an agent may always do best from the moral point of view by maximizing overall net good, for example, by always investing energy and resources at the margin in combating hunger, disease, and oppression worldwide. But they argue that even if this were so, it wouldn't follow that failing, say, to produce marginal increases in overall value at very large personal cost is wrong. They claim that a theory that would require that we do so is unreasonably demanding.[29] What really underlies this objection?

The following formulation puts the criticism in a way that helps one to see what is going on:

> Perhaps we would admire someone who behaved in this way. But is it plausible to claim that those of us who do not are guilty of wrongdoing; or that we have a moral obligation to devote all our resources to charity.[30]

'Guilty of wrongdoing' is the revealing phrase. Wrongdoing is something one can be *charged* with and, lacking adequate account or excuse, be guilty of, where guilt is a verdict (an Austinian 'verdictive') in some quasi-legal, second-personal form of accountability (Austin 1975).[31]

[29] A particularly good example is Scheffler 1982.

[30] This formulation actually comes from someone who tries to defend consequentialism in the face of the 'demandingness' objection (Mulgan 2001).

[31] Compare Nietzsche's claim that, whereas in the aristocratic ethos, 'good' is the primary notion and 'bad' is defined as not 'good', in morality (in the pejorative sense), 'evil' is the primary notion and 'good' its contradictory (in our terms: 'wrong' and 'right' (not wrong) its contradictory) (Nietzsche 1994).

That one is guilty of wrongdoing is not simply a finding that what one did was less than the best one could have done; it is the judgment that one did less than can be demanded, and that one can (and should) implicitly demand of oneself in a second-personal feeling that acknowledges guilt. What underlies the 'demandingness' objection, therefore, is the worry that act-consequentialism's standard of right goes beyond what we can reasonably demand of one another (second-personally). A moral demand just *is*, *inter alia*, there being warrant to address a second-personal demand to someone as one person among others, 'if not by law, by the opinion of his fellow-creatures; if not by opinion, by the reproaches of his own conscience' (Mill 1998: Ch. V, §14). To make sense of the 'demandingness' objection, therefore, we must see it as resting on the assumption that wrong and moral obligation are conceptually related to holding morally responsible, hence to second-personal demanding as it functions, for example, in the reactive attitude of guilt.

The other debate in which such a conceptual connection is implicitly assumed concerns morality's purported authority, that is, whether moral obligations are categorical imperatives in Foot's sense of always necessarily giving (conclusive) reasons for acting. Why does 'But it would be wrong,' always purport to provide a conclusive reason?[32]

A number of writers, most prominently, again, Bernard Williams, have argued that holding someone accountable for wrongdoing through blame unavoidably carries the implication that she had conclusive reason not to do what she is blamed for doing (Williams 1995: 40–4; see also Gibbard 1990: 299–300; Wallace 1994; Skorupski 1999: 42–3; Shafer-Landau 2003: 181–3).[33] Williams believes that this implication is 'bluff,' a bit of ideology that it is hopeless to try to vindicate or validate, since the only reasons for acting the person we blame can possibly have are internal reasons that are suitably anchored in her desires or other motivational susceptibilities, and nothing we could say on behalf of moral demands could possibly guarantee that.

Again, with the exception of the last bit, this just seems straightforwardly true. Try formulating an expression with which you might address a moral

[32] Note that I say purport to provide conclusive reason. Richard Nixon evidently relied on this implication, according to H. R. Haldeman's testimony, when he said in response to the question of whether hush money could be raised to pay off the Watergate burglars: 'There is no problem in raising a million dollars, we can do that. But it would be wrong.' (The last 'five crucial words' were not confirmed by John Dean's memory of the conversation.) ('Seven Charged, and a Briefcase,' *Time* 103 (March 11, 1974): 10–14).

[33] It is no coincidence that Williams is an original source of the 'too demandingness' objection to consequentialism. (Smart and Williams 1973).

demand to someone. I doubt that you can find one that does not carry the implication that she has conclusive reason to do what you are demanding or not to have done what you are blaming her for. Certainly none of the obvious formulations will work. For example, you can hardly sensibly say, 'You really shouldn't have done that,' and then add 'but you did have, nonetheless, conclusive reasons for doing it.' And if you try to pull your punches, by saying 'You shouldn't have done that, I mean, you know, morally speaking,' although you may end up canceling the implication of conclusive reasons, it's hard to see how you can without also canceling an implication of blame or demand. Or to turn the point around, if someone were actually able to establish that she did have good and sufficient reason for a putative violation of a moral obligation, then it seems she *would* have accounted or answered for herself. When we charge her with wrongdoing, therefore, we must be implying that she can't.[34]

Or recall Philippa Foot's comparison between morality and etiquette (Foot 1972). Norms of etiquette and morality are both categorical in form, and some norms of etiquette can be expressed in no less mandatory terms than can those of morality. One simply must not eat peas with a knife. But even so, we can cancel any implication that a 'must' of manners carries conclusive normative authority without thereby calling into question etiquette's customary normative purport. We can sensibly say that sometimes there is good reason not to do what etiquette requires without any suggestion that we are thereby somehow debunking manners. What explains this difference?

I believe that it is, again, moral obligation's essential tie to second-personal accountability. It is part of the very idea of a moral demand that we are accountable for complying with it. But such accountability seems no part whatsoever of the concept of a requirement of etiquette. That doesn't mean that manners are not, to some extent, 'morals writ small,' or that etiquette cannot be an important part of or supplement to equal respect, or, even more obviously, that some violations of etiquette are not morally wrong, like, for example, a rude joke at a funeral.[35] The point is that accountability is no part of the *concept* of etiquette in the way it is of moral obligation. To the contrary, what etiquette customarily calls for when its norms are violated is not accountability, but something more like distracting attention from an otherwise embarrassing reciprocal recognition of a gaff or, perhaps, third-personal disdain. Calling someone to account for bad manners is frequently bad manners itself.

[34] I am indebted to Christine Korsgaard for this way of putting it.
[35] I am indebted to a reader for the Press for this example.

MORAL OBLIGATION'S NORMATIVITY AND SECOND-PERSONAL REASONS

The concepts of wrong and moral obligation are, therefore, intrinsically related to the forms of second-personal address that, as we saw in the first section, help constitute moral accountability. It follows that the fact that an action is wrong, or that it violates a moral obligation, has an irreducibly second-personal aspect, that is, that it must itself be or entail a second-personal reason (or reasons). There can be no such thing as moral obligation and wrongdoing without the normative standing to demand and hold agents accountable for compliance. Of course, many of the considerations that ground claims of wrong and obligation are not themselves second-personal. That an action would cause severe harm, or even pain to your bunions, is a reason for someone not to do it, whether or not there is such a thing as a normative standing to demand that. But the action cannot violate a moral obligation unless such a standing exists, so the reason that derives from the moral obligation must be second-personal. Consequently, if moral obligations purport to provide conclusive normative reasons, other reasons to the contrary notwithstanding, then this must derive somehow from their second-personal character.[36]

As we noted at the outset, the projects of analyzing and vindicating morality's distinctive purported authority are generally framed in terms of 'categoricality' and the normative weight or priority of reasons to be moral. We are now in a position to appreciate why I there added that any attempt to account for moral obligation's distinctive bindingness must *also* explicate its distinctive tie to accountability. An adequate analysis of the concept of moral obligation must account for its conceptual connection to warranted (second-personal) demands. Even if, consequently, it were possible to account otherwise either for moral obligations' invariably purporting to provide superior normative reasons, or its actually doing so, it would still be impossible to explicate the distinctive hold or bindingness that moral obligations purport to have in non-second-personal times.

This means that the project of analyzing and accounting for moral obligation's normativity is seriously incomplete as it has traditionally been conceived. I believe, moreover, that the significance of the fact that moral

[36] In Chapter 12 of Darwall 2006, I argue that a promising way of accounting for specific moral obligations, and consequently, for why, for example, causing severe harm is wrong-making, is within a contractualist framework that is itself grounded in equal second-personal authority.

obligation's distinctive reasons are second-personal goes beyond even this. I have argued as well that appreciating moral obligation's tie to accountability also provides the best explanation of morality's purported normativity *as traditionally conceived*. Addressing moral demands second-personally presupposes that the reasons we address are supremely authoritative, on pain of our addressee's being otherwise able to justify his wrongdoing to us. Furthermore, I believe that this normative purport can also be vindicated from a second-person perspective. Again, I am not arguing for that bold thesis here.[37] Were it true, however, it would follow that appreciating the second-personal character of moral obligation is essential both to understanding its normative purport (both its putative normative weight as traditionally conceived and its distinctive *to*-ness) and to the most promising way of backing this hefty promissory note.

REFERENCES

Austin, J. L. (1975) *How to Do Things With Words* (Cambridge, Mass.: Harvard University Press).

Baier, Annette (1993) 'Moralism and Cruelty: Reflections on Hume and Kant,' *Ethics*, 103: 436–57.

Baier, Kurt (1966) 'Moral Obligation,' *American Philosophical Quarterly*, 3: 210–26.

Bennett, Jonathan (1980) 'Accountability,' in Zak Van Stratten (ed.), *Philosophical Subjects* (Oxford: Clarendon Press).

Bond, E. J. (1983) *Reason and Value* (Cambridge: Cambridge University Press).

Brandt, Richard (1979) *A Theory of the Good and the Right* (Oxford: Oxford University Press).

Dancy, Jonathan (2000) *Practical Reality* (Oxford: Oxford University Press).

D'Arms, Justin, and Jacobson, Daniel (2000) 'The Moralistic Fallacy: On the Appropriateness of the Emotions', *Philosophy and Phenomenological Research*, 61: 65–90.

Darwall, Stephen (1983) *Impartial Reason* (Ithaca, NY: Cornell University Press).

____ (1987) 'Abolishing Morality,' *Synthese*, 72: 71–89.

____ (2002) *Welfare and Rational Care* (Princeton, NJ: Princeton University Press).

____ (2006) *The Second-Person Standpoint: Morality, Respect, and Accountability* (Cambridge, Mass: Harvard University Press).

Dummett, Michael (1990) 'The Source of the Concept of Truth' in George Boolos (ed.), *Meaning and Method: Essays in Honour of Hilary Putnam* (Cambridge: Cambridge University Press).

Falk, W. D. (1986) 'Fact, Value, and Nonnatural Predication,' in *Ought, Reasons, and Morality* (Ithaca, NY: Cornell University Press).

Foot, Philippa (1972) 'Morality as a System of Hypothetical Imperatives', *Philosophical Review*, 81: 305–16.

[37] I do in Darwall 2006.

Frankfurt, Harry (2000) 'Rationalism in Ethics,' in Monika Betzler (ed.), *Autonomes Handeln: Beiträge zur Philosophie von Harry G. Frankfurt* (Berlin: Akademie Verlag).
Gibbard, Allan (1990) *Wise Choices, Apt Feelings* (Cambridge, Mass.: Harvard University Press).
Greenspan, P. S. (1992) 'Subjective Guilt and Responsibility,' *Mind*, 101: 287–303.
Hampton, Jean (1998) *The Authority of Reason* (Cambridge: Cambridge University Press).
Hare, R. M. (1993) 'Could Kant Have Been a Utilitarian?,' in *Kant and Critique: New Essays in Honor of W. H. Werkmeister* (Dordrecht: Kluwer Academic Publishers).
Hegel, Georg Wilhelm Friedrich (1991) *Elements of the Philosophy of Right*, ed. Allen W. Wood and Hugh B. Nisbett (Cambridge: Cambridge University Press).
Hieronymi, Pamela (2005) 'The Wrong Kind of Reason,' *Journal of Philosophy*, 102: 437–57.
Korsgaard, Christine (1996a) 'Two Distinctions in Goodness,' *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press).
____ (1996b) *The Sources of Normativity* (Cambridge: Cambridge University Press).
____ (1996c) 'Creating the Kingdom of Ends,' in *Creating the Kingdom of Ends* (Cambridge: Cambridge University Press).
Leiter, Brian (1995) 'Morality in the Pejorative Sense: On the Logic of Nietzsche's Critique of Morality,' *British Journal for the History of Philosophy*, 3: 113–45.
____ (1997) 'Nietzsche and the Morality Critics,' *Ethics*, 107: 250–85.
McNaughton, David, and Rawlings, Piers (1991) 'Agent-Relativity and the Doing-Happening Distinction,' *Philosophical Studies*, 63: 167–85.
Mason, Michelle (2003) 'Contempt as a Moral Attitude,' *Ethics*, 113: 234–72.
Mill, John Stuart (1998) *Utilitarianism*, ed. Roger Crisp (Oxford: Oxford University Press).
Moore, G. E. (1993) *Principia Ethica*, rev. edn. with the preface to the (projected) 2nd edn and other papers, ed. with an introd. Thomas Baldwin (Cambridge: Cambridge University Press).
Morris, Herbert (1976) 'Guilt and Shame,' in *Guilt and Innocence* (Berkeley: University of California Press).
Mulgan, Timothy (2001) *The Demands of Consequentialism* (Oxford: Oxford University Press).
Nagel, Thomas (1970) *The Possibility of Altruism* (Oxford: Clarendon Press).
____ (1986) *The View from Nowhere* (Oxford: Oxford University Press).
Nietzsche, Friedrich (1994) *On the Genealogy of Morals*, ed. Keith Ansell-Pearson and Carol Diethe (Cambridge: Cambridge University Press).
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).
____ (2000) 'Rationality and Reasons,' in Dan Egonsson, Björn Petersson, Jonas Josefsson, and Toni Ronnøw-Rasmussen (eds.), *Exploring Practical Philosophy: From Action to Values* (Aldershot: Ashgate Press).
Pettit, Philip, and Smith, Michael (1990) 'Backgrounding Desire,' *Philosophical Review*, 99: 565–92.

Quinn, Warren (1991) 'Putting Rationality in Its Place,' in *Morality and Action* (Cambridge: Cambridge University Press).
Rabinowicz, Wlodek, and Toni Ronnøw-Rasmussen (2004) 'The Strike of the Demon: On Fitting Pro-Attitudes and Value,' *Ethics*, 114: 391–423.
Rawls, John (1971) *A Theory of Justice* (Cambridge, Mass.: Harvard University Press).
Scanlon, T. M. (1998) *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press).
Scheffler, Samuel (1982) *The Rejection of Consequentialism* (Oxford: Clarendon Press).
Shafer-Landau, Russ (2003) *Moral Realism: A Defense* (New York: Oxford University Press).
Skorupski, John (1999) *Ethical Explorations* (Oxford: Oxford University Press).
Smart, J. J. C., and Williams, Bernard (1973) *Utilitarianism: For and Against* (Cambridge: Cambridge University Press).
Stocker, Michael (forthcoming) 'Shame, Guilt, and Pathological Guilt: A Discussion of Bernard Williams.'
Strawson, P. F. (1968) 'Freedom and Resentment,' in *Studies in the Philosophy of Thought and Action* (London: Oxford University Press).
Wallace, R. Jay (1994) *Responsibility and the Moral Sentiments* (Cambridge, Mass.: Harvard University Press).
Watson, Gary (1987) 'Responsibility and the Limits of Evil: Variations on a Strawsonian Theme,' in F. D. Schoeman (ed.), *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (Cambridge: Cambridge University Press).
Williams, Bernard (1985) *Ethics and the Limits of Philosophy* (Cambridge, Mass.: Harvard University Press).
____ (1993) *Shame and Necessity* (Berkeley: University of California Press).
____ (1995) 'Internal Reasons and the Obscurity of Blame,' in *Making Sense of Humanity* (Cambridge: Cambridge University Press).
Wollheim, Richard (1984) *The Thread of Life* (New Haven: Yale University Press).

6

Value and Autonomy in Kantian Ethics

Robert N. Johnson

Kantian ethics can at times appear to defend the position that there is a unique sort of value that plays a foundational role in morality. For instance, Kant's most well-known work in ethics, the *Groundwork of the Metaphysic of Morals*, begins by trying to establish that a good will is good 'without qualification' and then ends with a first statement of the fundamental principle that divides right from wrong, the Categorical Imperative.[1] This presentation can make it seem as if Kant believes the authority carried by the Categorical Imperative is somehow supposed to be grounded in the value of a good will. Again, the humanity formulation of the Categorical Imperative, the formulation that tells us we must respect the humanity in ourselves and others by treating it as an end in itself, appears to allude to a special value possessed by some feature of persons, their humanity, and then explain the authority of moral obligation by way of that value.[2] This extolling of the value of humanity and the dramatic refrain about the unique value of a good will both appear to portray Kant as telling us that moral reasoning consists of taking notice of the peculiar value that they possess which demands that we adjust our deliberation and actions in light of them. We appear to be told that the good will and humanity are bits of metaphysical glitter, jewels carrying their value around with them, and that this unique glitter is source of the authority of moral obligation.

I say that Kantian ethics can *appear* to say such things because I believe that it is in fact just an appearance. The value of humanity or of a good will does not in fact underwrite the authority of moral obligation. As I

Thanks to Peter Vallentyne, Ben Bradley, Jon Kvanvig, Richard Dean, Matt McGrath, those at the 2nd Annual Metaethics Conference at University of Wisconsin, Madison and at the University of Tennessee-Knoxville, and anonymous referees for helpful comments and suggestions.

[1] Kant (1996: 4.393–405). [2] Ibid., 4.428–9.

shall eventually argue, that authority must come solely from the fact that it is a demand of our own reason. If this last claim sounds familiar, it should; it is at the core of Kantian orthodoxy. But as familiar as the claim is, many have recently gone on record taking issue with this orthodoxy. In their view, while the orthodoxy holds that in fact Kantian ethics does not ground moral requirements in a value of some sort, in their view the value attributed to our humanity or to a good will is indeed the source of the authority of moral obligation. In what follows, I defend the orthodoxy against this attack. I argue that their position is in fact utterly incompatible with central doctrines of Kantian ethical theory, in particular, with the view that the human will is self-legislating or autonomous. The distinctive reason we have to conform to moral obligation, in Kantian ethics, stems simply from the autonomous nature of rational agency itself, not from its value or the value of anything else. Thus, if the human will is to be autonomous, then not even the value of a good will or of the humanity in persons can be the source of the authoritative reason Kantian ethics claims there is to conform to moral obligations. In short, there is no room and no need for metaphysical glitter in Kantian ethical theory.

In the first section, I detail the claim about value in Kant's ethics that a number of Kantians have recently defended. In the next, I explain why I think the Kantian doctrine of the autonomy of the will is incompatible with this claim. In the final section, I discuss how we should understand the relationship between Kantian views on value and the authority of moral obligation. Here, I bring elements of the Kantian theory of value to bear on the familiar discussion of the nature of the value of the good will and humanity.

1. THE ANTI-DEONTOLOGICAL KANTIANS

What I believe to be incompatible with the doctrine of the autonomy of the will is the claim that an appeal to some (however unique) value is required to explain the authority of moral obligation. By the 'authority' of moral obligation, I mean both the reason there is to comply with our obligations and the particular force or standing that this reason has. Thus, the claim goes, in order to explain how there could be a reason to conform to obligations that can often require significant personal sacrifice, we must appeal to some special value that is realized in or achieved by these actions. This is the way teleological views such as utilitarianism explain the authority of moral obligation. In the case of utilitarianism, that explanation is in terms of the nature of what she claims is of value, the general happiness. The reason to conform to moral obligations, in her view, is that such actions promote the overall good, which she identifies as general

happiness. According to such a view, there is a good, achieving the general happiness, which provides reason to comply with moral obligations. And the fact that this reason is found in promoting the general good, and not just one's own, is the source of its special status. Likewise, some think that there must be a value provided by Kantian ethics. To be sure, the Kantians who believe that there is some sort of grounding value in Kant's ethics do not think of that value as a utilitarian does, as an *outcome* of conformity to moral obligation. Instead, they simply believe that some good is realized by, or is in some other way connected to, performing one's obligations. And that value, they say, is contained in the value possessed by our humanity, our freedom, or the good will.

As a first example of such a view, consider Paul Guyer's position. He holds that for Kant 'freedom has an "inner value, i.e., dignity" ' and that this 'is the fundamental normative fact' that is 'the premise on which Kant's' moral philosophy rests.[3] Freedom gives us such 'an extraordinary sort of dignity, a dignity which so overwhelms the perceived value of satisfying any particular needs or inclinations that it can be immediately recognized to fulfill the expectations of unconditional value raised by our ordinary conceptions of morality and duty'.[4]

Thus, Guyer's position is that Kant's moral philosophy finds a special value, dignity, possessed by the freedom of a rational will, and that this value is the source of the unique reason to comply with the Categorical Imperative. For that value 'overwhelms the perceived value' of satisfying our own interests, thus showing that there is some good in complying, albeit perhaps not our own personal good. That, in Guyer's view, is the way in which we explain the authority of moral obligation.

Barbara Herman also defends a Kantian position that appeals to a distinctive value. She argues that 'Kant's project in ethics is to provide a correct analysis of "the Good" understood as the ultimate determining ground of all action' and so it is false that his view implies that 'no moral principles subtended from a concept of value can explain obligation'.[5] She defends the view that the principles of practical reason to which a good will conforms *themselves* constitute 'a conception of value'. Her defense is meant to counter the criticism that there is no such value in Kantian ethics. The criticism turns on the idea that unless there is a value that provides reason to conform to moral obligation, 'the rationale for moral constraint is a mystery'.[6] Thus, Herman as well holds that a unique value, the practical principles characterizing 'the good will', provides a reason to conform to moral obligation in Kantian ethics.

[3] Guyer (1998: 33).
[4] Ibid. 28
[5] Herman (1993: 209–10).
[6] Ibid. 210.

Another Kantian, Allen Wood, interprets Kant as holding that the value of humanity provides such a reason. He argues first that our own rational nature is itself 'the source of the fact of [other things'] goodness—indeed of the fact that anything at all is objectively good.'[7] The value of our own nature is the source of reason-giving value in much the way that the source of the authority of someone's recommendations to us is our respect for that person's authority on the question. The value of the recommendations depends on the value of her authority: advice is only as good as the expertise of the advisor. That means that the choices of a rational will 'can confer objective value on other things only if it is presupposed that it has objective value'.[8] Hence,

> if rational nature is in this way the prescriptive source of all objective goodness, then it must be the most fundamental object of respect or esteem, since if it is not respected as objectively good, then nothing else can be treated as objectively good … rational nature is presented as the only thing that could answer to the concept of an objective end or an end in itself.[9]

Wood's position is that while every other value is derived from the fact that they are chosen or pursued by a rational will, the rational will itself must have a unique intrinsic value, a value not derived from anything else. Our rational nature, our humanity, is a self-standing value. And that self-standing value is the source of the universally binding reasons to comply with moral obligations.

Wood develops his position in response to Christine Korsgaard's reconstruction of Kant's argument for the value of humanity. Korsgaard herself appears to be of two minds on this issue. In defense of the Kantian view of value, she argues, as does Wood, that

> good things are good in the way that Ross describes as relational, because of attitudes taken up towards them or because of other physical or psychological conditions that make them important to us. Only one thing—the good will itself—is assigned an intrinsic value or inner worth, and even the argument for that is not ontological. If we regard ourselves as having the power to justify our ends, the argument says, we must regard ourselves as having an inner worth—and we must treat others who can also place value on their ends in virtue of their humanity as having the same inner worth.[10]

She thus regards the good will has having a non-relational value, a value that is the source of our capacity to justify our ends. This position seems to be in line with Wood and the rest, the position that the unconditional value of the good will is a grounding source of reasons for moral demands.

[7] Wood (1999: 130); see also pp. 157–8.
[8] Ibid. 130
[9] Ibid.
[10] Korsgaard (1998: 272–3).

But Korsgaard has also defended a view she dubs 'procedural realism', which appears to reject this position.[11] Procedural realism denies that there are any substantive normative entities or properties to which our practical terms refer. Those who assert that there are such entities she dubs 'substantive realists'. The problem with substantive realism, she argues, is that 'the appeal to the existence of objective values cannot be used to support our confidence' that we have moral obligations, since it is that very confidence that supports our conviction that such values exist.[12] The only genuine grounds for asserting the existence of objective values is just this confidence, and in her view we have this confidence quite apart from proving their existence. An obvious way of construing this position that objective values cannot be used to support our confidence in moral obligation is as a denial of the claim that it is the objective value of humanity that is the source of the authority of moral obligation. That is, Korsgaard appears to mean that no objective values are sources of reasons to comply with moral obligations.

Another reason for thinking that Korsgaard would deny that a value grounds the authority of obligation is her view that 'normative concepts like right, good, obligation, reason are our names for the solutions to normative problems.'[13] I take it what she means to say is that value terms do not apply to value entities, in particular, value entities that could generate reasons for action. Of course, normative problems clearly do have solutions that refer to things that appear to give us reasons for actions. 'What ought we to do for the homeless?' is a normative problem with a solution—the offers of food, shelter, and medical aid, for instance—that are valuable. And it seems their value is what gives us a reason to make these offers. But, for Korsgaard, in fact there is no substantive value, such as the value of an offer of food, shelter and medical aid, which generates a reason to do what we ought and make that offer. If anything would be the source of a reason on this view, it would be that rational agents are disposed to agree on giving such an offer to the homeless.

While these positions seem to support the view that Korsgaard holds that no value is the source of reasons to conform to moral obligations, they fall short of explaining everything that needs explaining. It is in how this further part of the explanation is filled out that will answer the question whether she accepts the view that I've attributed to Guyer, Wood, and Herman. In fact, what is left to explain is what Wood's view adds to Korsgaard's, namely,

[11] For an excellent discussion of Korsgaard's metaethical views, see Hussain and Shah (2006).

[12] Korsgaard (1996: 40).

[13] Ibid. 47. I assume that she didn't mean to say that concepts are names.

the explanation for the fact that we often decide what to do by appealing to values we regard as in some way supporting our actions, and the fact that, when reasoning together, take ourselves to be trying to converge on values that will provide reasons for acting. What justifies following the rules or procedures of reasoning that produce this convergence if it is not the value of the rational agency of the deliberators who converge on them? One could say that the reasons supporting a choice flow from the value of what that choice aims at, and procedures would be justified only if they yield such values. But value facts or properties that could explain the rationality of such choices are the very things that Korsgaard denies exist. So we are left to explain the reason we have to comply with moral obligation as in some way coming from the value of humanity in ourselves and in others.[14] Korsgaard does not explicitly say so. But, often enough, she treats certain values as being exceptions, saying, for instance, that some things, a good will for instance, 'must obviously carry its own value with it'. [15] Although the evidence is equivocal, I shall assume that she would hold, as do the others, that while the value of everything else is nothing more than its possessing a relational property, of being the object of a rational will or being the point of agreement between deliberating rational agents, the good will and humanity possess values that are 'carried around with' them, and that she thinks these values provide reasons to comply with moral obligations.

Thomas Nagel will be my final example of a philosopher who holds that Kant's ethics views some unique value to be the source of reasons. Kant's ethics, in his view, implies that 'freedom can be pursued and approached only through the achievement of objective and ultimately ethical values of some kind'.[16] The 'achievement' of such values does this by allowing 'the will to expand at least some of the way along the path of transcendence possible for the understanding' and attain a practical kind of 'view from nowhere'.[17] The will expands to this view, not by attaining the sort of 'external' view available from the point of view of the sciences. That sort of external point of view is inconsistent with action, which always requires to some extent taking one's own internal point of view as agent. Yet free willing is willing not merely what is of value to you, from your own point of view, but what is of value from an objective point of view. So being guided by objective values is supposed to provide whatever degree of freedom we think we enjoy in action, by providing a certain sort of reason for action, a reason that supports conforming to deontological constraints. This well-known line of reasoning from Nagel's work seems to me to show

[14] See also Korsgaard (1997: 218).
[15] Korsgaard (1998: 257).
[16] Nagel (1986: 137).
[17] Ibid. 136.

that he too defends the claim, in the name of Kantian ethics, that some objective value is the source of the authority of moral obligation.

2. AUTONOMY AND REASONS

So far, I have outlined a trend I see among Kantians to portray value, such as the value of freedom, humanity, or the good will, as a source of a distinctive reason to conform to moral obligations. In this section, I explain why I think that they must, to remain consistent, deny that the value of these things provides such a reason. In the following section I will go on to argue that this is nevertheless consistent with holding that their value is unique.

It is instructive to contrast Kant's views, at least as I portray them, with that of others who also deny that some objective value gives us a reason to conform to moral obligation. These others reject this position simply because they think something subjective provides such a reason. For them, it is, for instance, satisfying our desires, not realizing some objective value, which constitutes whatever reason there is to comply with moral obligations. The Kantian view is different from such desire or interest-based explanations of the authority of obligation in holding that what we call 'good' or 'valuable' is whatever is an object of *practical reason*, and it is our own practical reason that is the source of the authority of obligation, not our desires nor the value of the objects of practical reason. According to this view, the 'practical' in 'practical reason' concerns the will conceived of as a faculty consisting of something distinct from the sorts of empirical psychological states collected under the term 'desire'.[18] Details about the faculty of practical reason are not important. The key idea, deployed by the Kantians I have been discussing, is that things, or at least most things, are valuable because they are objects of rational (in a Kantian sense) choice. What is at issue is whether, in addition to things that acquire value by being related in the right way to rational choice, there is another sort of value, the value supposedly residing in the good will or humanity, that does not get its value in this way, is the condition or source of the value of the objects of rational choice, and is ultimately what provides a reason for conforming to moral obligation.

In my view, Kantian ethics must deny that there is such a value. It must deny this because it is inconsistent with claims that are central to Kantianism, namely, that (a) there are requirements that are absolutely binding on rational agents, (b) there are such requirements only if rational

[18] Kant (1996: 5.60).

agents are wholly self-legislating or are autonomous, and (c) rational agents are wholly self-legislating only if no value explains their authority over us. That (a) is a core Kantian doctrine should be familiar enough. This claim just comes to the idea that there are categorical imperatives, or commands that apply to us no matter what our personal contingent reasons might be for or against complying. Admittedly, the existence of such requirements is controversial. That they in turn require the autonomy of rational agency, as (b) states, is more controversial, though still a familiar aspect of Kantian ethical theory.[19] It is (c) that most concerns the relationship between autonomy and value.

First, let me give you an intuitive sense of why autonomy commits Kantian ethics to denying that value is a source of reasons. Consider a parallel example: the Divine Command theory's resolution of a Euthyphro-style dilemma. That dilemma begins with the assertion that God loves (or responds in some appropriate way to) all and only good things. This raises the question, Why? Is it because their value provides a reason for God to give them their due, His love? But if God loves a good thing only because its goodness gives him a reason to love it, then its goodness explains the appropriateness of God's love for it, and this is incompatible with God's omnipotence. The value that provides a reason for God to love it would be a constraint on God's love in the sense that God must respond to the reasons provided by the value of things or else fail to have the requisite response. The alternative is to say that those things are good because of God's love for them. God's love explains their value. But then goodness looks like an arbitrarily distributed shadow cast by God's attention. For if God loved some entirely different set of things, then those other things would have been good. The Divine Command theorist opts for the second horn, and then is saddled with the problem of explaining why goodness isn't arbitrarily distributed after all.

Kantians must resolve a similar Euthyphro-style dilemma in the same way as the Divine Command theory. What possess value on the Kantian view are all and only the objects of rational agency. Now if value is the source of the reasons for the pursuits of rational agents, then the authority governing rational agency is external to that agency itself, in the value of the things that are its objects. But on the Kantian view, rational agency must be autonomous, in the sense that the requirements binding it are wholly self-generated and self-imposed. The autonomy of reason, the central guiding idea behind Kantian moral theory, is thus the very foundation of the case

[19] This has been discussed under the heading of 'the Reciprocity Thesis' recently by Henry Allison in his (1986).

against the claim that there is some value that provides reason to conform to moral obligation. Autonomy requires that value not be a source of reasons.

Rational agency must be autonomous because, as I mentioned above, if there is to be an absolutely authoritative rational requirement such as the Categorical Imperative, then it must be self-legislated and gain its authority from this self-legislation. The line of reasoning is this: A practical requirement is binding only if the agent can voluntarily comply with it. But an agent can comply with a requirement only if there are reasons for her to comply. Now if an agent is bound absolutely to comply with a requirement, then she must always comply with it. But then if she must always comply, there must always be a reason for her to comply. But for any requirement, there will not always be personal reasons for an agent to comply—reasons that come from her own contingent circumstances. In a circumstance in which there are no contingent personal reasons, then, there must be some other reasons. But the only reasons that will always be present for complying with an absolutely binding requirement will just be whatever reasons there were for imposing that absolutely binding requirement on the agent in the first place. So, if the agent is bound absolutely to a requirement, then, given that the ability to comply requires reasons, it must be possible for her reasons to be those on the basis of which the requirement was imposed on her (or legislated).

Now if it is possible for her reasons to be those on the basis of which the requirement was legislated, then she must be able to appreciate fully why the absolutely binding requirement came to be legislated. And if she is able to appreciate this, then she must be able to engage fully in the deliberation that led to its legislation. But she will be able to engage in that reasoning herself only if she herself possesses whatever capacities and hence authority a legislator of such a requirement can claim is sufficient to generate and enact it. So an agent who is bound absolutely by a requirement must be no different from its legislator, the source of its authority over her.[20]

A rational agent will therefore be wholly self-legislating only if the authority of the principles governing her agency comes from the fact that she herself is the author of those principles. The reason she must conform to these principles is thus that she gave them to herself—not because some good comes from or is realized by following them. And rational agents give themselves the laws that they do because it is in the nature of their rationality to do so. As it happens, on Kant's own view, rational agency operates on the basis of laws valid for all rational agents, and it is essential to (and not *analytic of*) rational agency that it do so. This is because rational agency is essentially a kind of causation. Since any causation in

[20] I here draw on Andrews Reath's reconstruction of this argument in his (1994).

Kant's view brings with it universal laws, it is essential to rational agency that it govern itself by *universal law*, laws valid for all rational agents. This is how Kant proposes to show that being governed by the Categorical Imperative is essential to rational agency. My argument, however, does not require this last step connecting autonomy with the universal law version of the Categorical Imperative through the idea of causation. What autonomy of the will requires is only that the explanation of the authority of the principles governing the will comes from the fact that the will is the source of those principles. And if the reason for you to conform to a law is the fact that you gave that law to yourself, then the reason does not derive from any value, such as the value of your will or your humanity. To be sure, your humanity is, in Kantian ethics, composed in part of your capacity to lay down practical laws for yourself. And that humanity is of unique value. But the reason you must conform to moral requirements does not derive from the unique value of this capacity, in you or in anyone else. It derives from your exercise of that capacity alone. So the thesis that the will is autonomous is not compatible with the claim that the source of the reasons there are to conform to moral obligation is a value of some sort.

There is, then, no room for the metaphysical glitter of a special value in Kantian ethics. But isn't there yet a need for it? For if no value is realized, achieved, respected, or otherwise brought about by conforming to moral obligations, then what reason is there to do so? Kantian ethics, one might object, is satisfied giving no reason at all. This objection is right insofar as Kantian ethics does not appeal to the value of something to give such a reason. But the assumption of the criticism is that a reason requires something of value, and the Kantian position I have laid out above contests just this assumption. There is a reason to conform to moral obligations, but that reason is that you demand it of yourself. Kantian ethics is at bottom the idea that the reason you should conform to moral requirements is the fact that you imposed them on yourself, and you do this simply because you are a rational agent. Quite apart from whether the fundamental moral principle is the Categorical Imperative, it must, according to an ethical theory in which rational agents are autonomous, be constitutive of rational agency that the authority of its principles be grounded simply in the fact that it is their author. And that, in turn, means that the fact that you gave yourself this principle is the only reason that exists in every circumstance to conform to it.

To be sure, the autonomy of rational agency remains a controversial aspect of Kantian ethical theory. My point is that *given* Kantians are committed to it, they simply must reject the idea of an authoritative good—as it might feature in the value of a good will or the value of humanity as an end in itself—a value that explains the authority of moral requirements. But, as

I hinted at the outset, if, as I think, Kantians are committed to rejecting an authoritative good of any kind, this raises questions about views that also seem central to Kantian ethics, in particular, views about the value of the good will and humanity. How can humanity and the good will possess the unique sorts of value that they supposedly do yet not provide reason to conform to moral obligations, such as the demand that we respect humanity in ourselves and others? How could its value *not* be the reason we are to treat humanity with respect? I will turn to this last question in the next section to explain how I think they should be answered.

3. THE VALUE OF A GOOD WILL AND HUMANITY

Let's return for a moment to Wood's argument for the value of humanity and its centrality to explaining the authority of moral obligation. Wood's view is that since rational choice is the source of all value it must be objectively good. This, for instance, is the sort of reasoning Kant seems to engage in here:

> Nothing can have a value other than that determined for it by the law. But the law-making which determines all value must for this reason have a dignity—that is, an unconditioned and incomparable worth—for the appreciation of which, as necessarily given by a rational being, the word '*reverence*' is the only becoming expression. (1996: 4.436)

One can easily read this passage as asking us to accept the claim that humanity has a special value, and it is just this value that is the source of the authority of morality. Indeed, Kant appears to be saying that there is a difference between values, those that are authoritative and those that are not. The dignity and worth possessed by a rational will, he appears to say, is quite unlike other sorts of value, having 'incomparable worth'. That incomparable worth explains the value of every other thing, and because of this, deserves our reverence. And that it deserves reverence is the reason to treat it in the various ways demanded by the Categorical Imperative. Wood appears to be reading such passages from Kant in this way, as saying, in summary, that the authoritative value of humanity is the source of reasons to conform to moral obligations.

Nevertheless, I think that this tempting reading is entirely wrong. We should read such passages about the value of humanity and the value of a good will in a different way. We should, in other words, take Kant at his word when he says 'nothing can have a value other than that determined for it by the law'. 'Nothing' means just that, nothing. So this must apply to 'the lawmaking that determines all value' as well everything else. If it—that

is, rational willing—is of value, it is because it too is an object of a rational will. This sounds dark and tautological, but it is neither. Let me explain.

Assume that to possess a good will is, very roughly, to be deeply committed to the principle designated as Categorical Imperative. By 'commitment' I mean a disposition to affirm the Categorical Imperative as the ultimate standard for one's behavior (although not formulated necessarily in the precise words used by Kant himself), to try (absent irrationality) to conform to it and to disapprove of oneself and others when deliberately or negligently failing to do so. Assume further that a disposition to be committed to this principle is that element of our humanity that gives humanity its special status in Kantian ethics. This would be a higher-order disposition, a disposition to acquire the dispositions I have said characterize the commitment that is a good will. That second-order disposition, given it is characteristic of every person's will, is present in every circumstance in which we act. So the value of humanity will be the value of having this second-order disposition to acquire a commitment to the Categorical Imperative, and the value of the good will is the value of the realization of this second-order disposition. [21] I shall ignore for present purposes any differences there might be between, on the one hand, humanity and a good will, and on the other, freedom, 'rational nature' and other things regarded as of special value by the Kantians I've so far discussed.

Recall that the official Kantian theory of value *in general* (that is, without specifying the kind or modification of value, such as moral or non-moral value) is that to be good is to be an object of a rational will, where 'rational' is understood in terms of the Kantian theory of rationality.[22] In short, what has value is whatever is an object of a rational choice. The Kantian views I have been discussing so far hold that the value of every good thing *except* the values of a good will and our humanity comes from being their objects of rational choice. I believe that this exception is a mistake. The value of the good will and humanity also comes from their being the objects of rational choice, and not the other way around. To understand how this is so, we must discuss the variety of ways in which something can be an object of choice.

There are two quite different kinds of objects of choice for Kantians, means and ends, as is implied by the Kantian slogan 'who wills the end wills the means'. We choose our means and, according to Kantians, our ends, making each a distinct kind of object of choice. Further, something's being an end of one's choice does not preclude its being a means in another and vice versa. My means of attaining the end of getting ice cream was a trip to

[21] Thanks to Judy Thomson for pressing me to clarify this.
[22] Kant (1996: 5.57–66).

the parlor, but getting the ice cream might in turn have been a means of satisfying my hunger. Moreover, something's being an end of one's choice does not preclude it from being a means in that selfsame choice. If my end is to engage in some activity, such as golfing, choosing to golf makes golf both the end I'm aiming at in my choice and the means of fulfilling it. Finally, not only, according to the Kantian view, *can* our ends be objects of choice, but they *must* be, since every object of will must contain an end, and all of a rational agent's ends must be chosen.[23]

When means are objects of the will, their value is only conditional and extrinsic—they are good only on the condition that the ends they serve are objects of our rational choice. It is the property of realizing or producing those ends that makes them the objects of our choice. Humanity, however, is supposed to be 'an end in itself', never to be treated as a mere means to our personal ends, and this is supposed to mean that it is intrinsically valuable. Now there are different senses in which something can be an end. In one sense, an end is simply whatever we choose to produce or bring about in the world by some means. For instance, if having some ice cream is my end, then having some ice cream is an event in the world I set myself to bring about. Adopting these kinds of ends guides actions in that when I set myself to pursue them, I then must find actions that will be means of producing them. Choosing or willing an end is in this way a source of a law for my action: Willing the end dictates that rational agents do something, namely, act to bring about that end.

Humanity is not an end in this sense, but a good will can be such an end. For we cannot choose to have the second-order disposition of the sort we have identified as our humanity; that is simply something we have in virtue of being rational agents of the sort that we are, not something we can produce. So humanity is not valuable because we rationally will to produce it. But a good will can be said to be an object of choice in this sense, since a good will is the realization of the second-order disposition of acquiring a commitment to the Categorical Imperative of the sort I described above. Such a thing can be brought about, indeed can only be brought about, through choosing to acquire that commitment. As such, the value of a good will comes from its being the object of a fully rational choice, as being the end of that choice. That is to say, a choice to acquire a good will is a choice that is completely determined by Kantian rational principles, including the Categorical Imperative. Thus, the good will can be an object of rational choice and this is all that its goodness amounts to.

[23] Kant (1996: 6.380, 385). I argued in Johnson (2002) that this doctrine is incompatible with another doctrine Kant holds, that we have our own happiness as our end by natural necessity (e.g. see Kant (1996: 4.415–16, 6.382)).

However, being the product of a choice is not the only sense in which something can be an end. When I shop, one of my ends is to economize.[24] But this is not something I set myself to produce. It is better understood as something that *prevents* me from acting in a variety of ways. For instance, choosing to economize prevents me from buying name brands instead of generic brands, from buying the first item I see, and so on. Some things, that is, are ends in the sense that they are that *against which* I may not act while pursuing other ends. Humanity is clearly meant to be an end of this sort, and hence can be an object of choice in this sense. Humanity is an object of my choice in the sense that economizing is an object of my choice: It is a limit on my other ends, and so is a 'negative' end in the way that economizing limits what I may purchase.

Economizing, of course, is not an end every rational being must have. Some make economizing an end, others do not, and although it often limits the brands I myself buy, my health is more important than economizing, limiting how much I will rein in my spending. But my rationality does not depend on whether I have chosen economizing as my end, or on whether economizing limits my health or vice versa. By contrast, humanity is supposed to be an *objective* end, an end that all rational beings must have, no matter what other ends they have and must limit every other pursuit, while no other pursuit may limit it. Hence, it limits what I may do—if I am to be fully rational—when I pursue my positive *and* subjective negative ends.

Although it is not something produced by my actions, humanity is also supposed to be in a different sense also a *positive* end. Sometimes an end is neither an outcome nor a limit, but an activity. Speaking a language and playing a game or musical instrument are ends of this kind. They are not produced by actions, but realized in them. When my end is speaking German, my actions do not, or at least not simply, produce 'speaking German'; they constitute or realize it. Humanity is also an end of this kind, a *positive* end, that is, something to be realized in various activities. Realization in this sense is making actual what is potential, so we can think of humanity being our (rational) end as making actual whatever potentialities humanity is composed of. And one such potentiality is the second-order disposition to acquire a commitment to the Categorical Imperative.

To summarize, then, our humanity and the disposition of which the good will consists can be thought of as objects of rational (in a Kantian sense) choice in the sense that they are in several, though not all, senses rational ends (which is of course compatible with humanity also being a

[24] I borrow this example and the explanation of this distinction in ends from Barbara Herman in Herman (1993: 14).

means of which we choose to make use). Our humanity—especially in the sense of its including a second-order disposition to acquire the commitment to the Categorical Imperative—and good will can be something to realize in our activities and can be limits on our other ends and activities. If this is so, then we can quite easily make sense of the idea that they are good on the official Kantian theory of value in general. Necessarily for every agent who is rational in the full-blooded Kantian sense, and good will and the humanity in herself and in others are objects of her will in the sense that they are rationally necessary limits on every other end she pursues in every circumstance, as well as are something to be realized and furthered in her actions. They are thus valuable because they are objects of a rational will—because they are related in the right way to a rational will. It is not that their value stems from their being related in the right way to rational choice that distinguishes things of ordinary value from the special value of the good will and humanity. It is the nature of the relation in each case. Humanity and the good will are necessarily and universally the objects of a choice rational in the Kantian sense. Other things are only contingently so. Their value thus need not be seen as a kind of metaphysical glitter.

If I am right that humanity contains the idea of good will (at least understood as the disposition to acquire a commitment any minimally rational will has to the moral law), then we now understand how the good will can be under every circumstance an object of rational choice, and is the condition of which anything else is an object of choice. It is a kind of limit on every other object of choice. Put simply, to say a good will is unconditionally good is to say that I rationally may not choose in a way that will degrade or thwart the development of a commitment to the Categorical Imperative, no matter what other ends I pursue.

REFERENCES

Allison, Henry (1986) 'Morality and Freedom: Kant's Reciprocity Thesis', *Philosophical Review*, 95: 393–425.

Guyer, Paul (1998) 'The Value of Reason and the Value of Freedom', *Ethics,* 109: 22–35.

Herman, Barabara (1993) *The Practice of Moral Judment* (Cambridge, Mass.: Harvard University Press).

Hussain, Nadeem, and Shah, Nishi (2006) 'Misunderstanding Metaethics: Korsgaard's Rejection of Realism' in *Oxford Studies in Metaethics* Vol. 1 (New York: Oxford University Press).

Johnson, Robert (2002) 'Happiness as a Natural End' in *Kant's Metaphysics of Morals*, ed. M. Timmons (Oxford: Oxford University Press), 317–30.

Kant, Immanuel (1996) *Practical Philosophy*, trans. M. Gregor (New York: Cambridge University Press). All page numbers refer to the numbering in the edition

of Kant's collected writings, edited by the Königlich Preussischen Akademie der Wissennschaften (Berlin, Reimer 1900–), vols. 1–9.

Korsgaard, Christine (1996) *The Sources of Normativity* (New York: Cambridge University Press, 1996).

——— (1997) 'The Normativity of Instrumental Reason', in *Ethics and Practical Reason*, ed. G. Cullity and B. Gaut (New York: Oxford University Press), 215–54.

——— *Creating the Kingdom of Ends* (New York: Cambridge University Press, 1998).

Nagel, Thomas (1986) *The View from Nowhere* (New York: Oxford University Press).

Reath, Andrews (1994) 'Legislating the Moral Law', *Nous*, 28: 436–64.

Wood, Allen (1999) *Kant's Ethical Thought* (New York: Cambridge University Press).

7

Where the Laws Are

Mark N. Lance and Margaret Olivia Little

1. MORAL CONTEXTUALISM

A number of theorists have recently urged that the moral principles so prized by many are in fact strewn with exceptions.[1] Lying is always wrong-making—well, not when playing the game Diplomacy, in which lying is the point of the game, or again when confronted with the Nazi concentration camp guards, to whom the truth is not owed. Pleasure is always good-making—well, not when it is the pain enjoyed by the sadist, delighting in his victim's pain.[2] The claim proffered is not simply that the wrongness of the lie or goodness of the pleasure are in these instances *outweighed* by other considerations, or again that the exceptions can be expunged if only we refine our propositions carefully enough. The claim, instead, is that the 'moral valence' these features carry to their respective situations have themselves switched from their more familiar mode, and in ways that cannot be helpfully codified. The fact that something is a lie does not always count against it; the fact that something would bring pleasure is not always a count in favor; and there is no specifying in genuinely explanatory terms the conditions under which they do. Moral considerations, on this view, are *radically* context-dependent.

Of course, much debate has ensued on whether this is the right picture of morality.[3] But a separate question is what would follow if it were. According to many, the answer—for better or for worse—is moral particularism. As its name implies, moral particularism is a view that stands opposed to certain

[1] Dancy (1993), (2000), (2004), Little (2000), Murdoch (1970), McNaughton (1988).

[2] The first example is David McNaughton's; the pleasure example is Jonathan Dancy's.

[3] See, for instance, the essays in Hooker and Little (2000); for a recent defense of generalism, see McKeever and Ridge (2006).

roles for generalizations. More specifically, it is the view that explanatory or theoretical moral generalizations play no essential role in moral understanding.[4] According to particularists, moral understanding is the exclusive province of particular judgments, perceptions, skills, and inarticulate syntheses of individual considerations. While moral generalizations may still stand as useful rules of thumb or helpful heuristics, they do not provide true explanatory generalizations illuminating the structure of morality, for morality cannot be thusly illuminated. In the view of many, it turns out, to accept radical moral contextualism is to believe that morality—and our understanding of it—is not a domain governed by laws.

We disagree. In a series of recent papers,[5] we have argued that moral contextualism is true, while moral particularism, thus defined, is false. Indeed, our primary purpose has been to diagnose why the two disputes are so often thought to be one. In our view, a central culprit is the widespread assumption that generalizations must be exceptionless if they are to do genuine and fundamental explanatory theoretical work. Such an assumption can arise from a number of different sources—from conceptions of the nature of reasons, explanation, theory, or laws, or again from broad metaphysical assumptions; our goal has been to challenge it nonetheless.

The wedge that we have employed in driving distance between contextualism and particularism has been the notion of a 'defeasible generalization.' Our claim is that there is an important kind of generalization that is both fundamentally explanatory and fundamentally porous—shot through with holes. Our primary goal has been to provide an account of the semantics and epistemology of defeasible generalizations in order to show that it is possible for them to play explanatory roles without being reducible to, replaceable by, or ultimately beholden to exceptionless generalizations. Such a view makes room, we believe, for moral contextualists to accept plausibly necessary ties between reasons, explanations, concepts, and generalizations, to embrace moral theory as a significant enterprise, and to recover much more natural accounts of moral dispute and moral learning.

In this paper, we take specific aim at the objection that exception-filled generalizations cannot function as *laws*—the most fundamental sort of theoretical generalization, meant to undergird more everyday ones and, crucially, to underwrite the nature of *kinds*. We want to argue that defeasible generalizations, properly understood, are capable of functioning as genuine moral laws. Reflection on the function of laws helps us to recover an approach to laws that is not threatened by the right sort of exception. Determining just which domains admit of the 'right sort' of exception

[4] See, for instance, Dancy (2004).
[5] Lance and Little (2004), (2006a) (2006b).

can, further, help us to understand the sort of objectivity that is likely for domains governed by such porous—that is, defeasible—laws.

2. DEFEASIBLE GENERALIZATIONS

Exceptions pepper generalizations in all sorts of disciplines, from discourse on artifacts to economics, biology to semantics, aesthetics to epistemology, and, most especially, ethics.

- Defeasibly, matches light when struck.
- Subject to provisos, an increase in supply leads to a drop in price.
- Ceteris paribus, sheep reproduce only with other sheep.
- Other things being equal, fish eggs develop into fish.
- Defeasibly, consistent usage of a term across a linguistic community is consistent with the proper meaning of the term in that community.
- Paradigmatically, chairs are for sitting in.
- Typically, lack of originality is aesthetically negative.
- In normal conditions, appearances are epistemically trustworthy.
- As a rule, the future is like the past.
- Defeasibly, lying is wrong-making.
- For the most part, pain is morally bad-making.
- In normal conditions, people should be taken at their word in expressing their own desires.

According to many philosophical views, the exceptions that are seemingly sanctioned by provisos such as 'defeasible,' 'ceteris paribus,' and the like are ultimately antithetical to genuine explanation. Their presence calls for us to purify the accused generalization in one of several ways—to refine its claim, or delimit its scope, to weaken its quantifier into statistical form, or mark it as merely useful shorthand. At best, it requires grounding the generalization's theoretical bona fides on the truth of some further generalization, located at a 'deeper' level, which is finally free from exception. If all this fails, then—so a standard story goes—we have a sign that the theory within which the generalization functions is deeply defective.[6]

We ourselves are convinced that, very often, neither an enthymematic strategy nor a replacement by statistical generalization succeeds in capturing the intended force of these hedged generalizations. Rather than arguing for this here, we aim to undercut motivation for the claim that there

[6] See, for example, Earman and Roberts (1999), which argues that no legitimate empirical theory can make use of unreducibly defeasible generalizations. We address this argument in our (2004).

must be an underlying invariantist layer by showing how to take defeasible generalizations seriously. With this understanding in place, we suggest that the onus of proof is upon the totalitarian invariantist to show why, in any particular case, the subject matter in question ultimately requires exceptionless generalizations.

What then is a defeasible explanatory generalization? The key to understanding such generalizations, we argue, is to see them as made true by a complex normative structure that demarcates some conditions as theoretically *privileged* in one way or another. Such generalizations tell us both what happens in conditions that are thus privileged *and* what compensatory moves are required by the ways in which one's situation may stand in distance from the privileged ones. This is not to say that one can rotely translate 'Defeasibly *P*' as 'In privileged conditions, *P*;' it is to say, rather, that with a suitably articulated variety of privileging notions, along with a range of semantic and epistemic devices familiar from other contexts, one can make sense of any genuine (i.e. irreducibly) defeasible generalization.

Let's look at a simple example: defeasibly, matches light when struck. Only defeasibly, for there are all manner of conditions—wet, overly cold, overly hot, overly lacking in oxygen, etc.—in which matches do not light when struck, and no thought that we could specify in finite form the list of suppressed premisses. Nor is it a merely statistical generalization: in certain circumstances—say, for those who live in watery Atlantis—the exceptional cases are far more locally common and far more salient than the non-exceptional. That is, Atlantans will most certainly not conclude from the generalization that if they were to strike this match here, it is likely to light. But even in Atlantis, it is true that defeasibly matches light when struck, for that is just what it is for something to be of the artifactual kind *match* rather than *red-phosphorous-and-crushed-glass-tipped stick*. Atlantans may well not have such a concept, since the artifact it governs would be singularly useless for them; nonetheless, if they do have the concept, they are committed to the generalization.

Very roughly put, to understand the defeasible connection between striking and lighting that governs the concept match (to practically understand matches) is to know the various ways in which conditions can vary from the privileged ones and the differences those deviations make to the behavior of matches. It is to understand, for instance, that matches don't light when wet, unless again they are in the presence of a particularly heavy concentration of oxygen, but even then not if the temperature is near to absolute zero, and on and on. Such a generalization is an explanatory or theoretical one: it captures what it is one comes to understand when one comes to learn to use matches—and concomitantly to learn to master the concept of 'match':

namely, the distinction between privileged and non-privileged conditions for lighting matches and the differences those differences make. 'Privileged,' we emphasize, for the understanding of *matches*. Our claim is not that there is some one official set of worlds that counts as the once-and-for-all privileged set for all defeasible generalizations. Privilege is relativized by theoretical domains (such as biology), by concepts (such as match), or by clusters of generalizations (rules of etiquette).

Defeasible generalizations of this sort, in short, involve a subjunctive. We can read 'defeasibly, matches light when struck' as 'in privileged conditions (for the exhibition of the nature of matches) if any match were to be struck, it would light.' To understand the function of matches requires a practical grasp of a particular similarity relation among worlds—a grasp, that is, of the modal geography of worlds near to the privileged one. Our practical understanding of privileged conditions functions, then, to give us a baseline set of conditions in which the generalization holds directly, to allow us to single out as salient the ways in which another situation may be non-privileged, and, finally, to understand what compensatory adjustments are asked for by the ways in which we there depart from privileged conditions. If, then, one has an adequate understanding of the relevant notion of privileging and of the relevant notion of nearness of world to understand 'in privileged conditions, if any match were to be struck, it would light,' one will also be in a position to know that, in those worlds nearest to privileged in which a match is struck when wet, it will not light.

Let's move now from matches to morality. We follow Kant in thinking it essential to both morality and our conception of a person that, in privileged conditions, lying is at least wrong-*making:* that is, in privileged conditions, the fact that something counts as a lie counts against doing it, even if that count is ultimately outweighed by other exigencies. Differences from Kant quickly appear, though. For not only do we believe that there is no possibility of strengthening this claim to an invariant prohibition on lying, we believe there is no possibility of strengthening it to a declaration that lying invariantly carries wrong-making import: there are circumstances in which its very valence switches. David McNaughton gives the example of lying while playing the game Diplomacy: hardly wrong-making, lying is the very point of the game. For an example closer to home, suppose that a strengthened version of the Patriot Act is passed, and one finds oneself confronted by a government death squad agent asking where one's activist daughter is staying. By virtue of active collaboration with an oppressive state apparatus, and in virtue of the structurally non-paradigmatic system of human relations within which such an interaction would be embedded, this person, we would argue, is simply not worthy of the truth. It is not wrong to lie in such a situation; more than that, we want to claim, it

is not even a negative feature of one's act that it involves lying to such a one.

In the latter case, note, the deviation from privileged conditions indicates that one occupies a morally *defective* situation. It is a bad-making feature of the situation we are in that lying is here not wrong-making: would that the world did not have rational creatures in it to whom the truth is not owed. The valence-shift in playing Diplomacy, in contrast, counts as merely *deviant*. What is important to both is the fact that situations in which it *is* morally wrong-making to lie are privileged in a deeply conceptual manner. Part of what it means to take something to be a *person*, we would argue, is to understand the creature as belonging to a kind that defeasibly has a claim on our honesty. Situations in which one takes something to be a person but not worthy of honesty are inherently riffs, as it were, on the standard theme of person. We cannot, then, fully understand the moral situation of the death squad agent's demand merely by understanding that lying there has its 'thumbs-up' valence; rather, we must understand that the situation is deviant (here, indeed, both in being conceptually derivative and in being morally objectionable) and that lying has here the status that it does precisely *because* of that deviation.

Further, the specifics of what is here wrong-making have to do with the specifics of *how* this situation departs from the privileged sort of situation. Whether it is permissible to lie to the agent in such a way that he is lured to his own death, for instance, may well depend on whether he is an officer and leader of the fascist movement or a coerced conscript. Deviations from privileged conditions lead to other compensatory deviations; understanding in a context of defeasibility comes, not from expunging exceptions, but from learning to appreciate the difference they make to the entire theoretical fabric.

Now some will accept the above as far as it goes—as a way, perhaps, of recovering Aristotle's lesson that judgment is an irreducible part of moral epistemology—but argue that its lessons remain superficial in an important sense. The presence of defeasible moral principles is perfectly acceptable, it is admitted, but only if their explanatory force is redeemable by ascending (or descending!) to a more fundamental level of explanation: namely, to the level of moral laws, where exceptions must finally be firmly expunged.[7] The presence of exception, in short, signals that we are not at the level of ultimate explanation. We want to argue, though, that to truly appreciate the role of defeasible generalizations is to appreciate that, on the right picture of laws and in the right sorts of disciplines, ultimate laws themselves can be defeasible.

[7] This is Roger Crisp's position in his essay, 'Particularizing Particularism' (2000).

3. LAWS AND LAWLIKENESS

Very broadly put, there are two fundamentally different approaches to thinking about laws, approaches we will label 'metaphysical' and 'pragmatist.' On the metaphysical approach, the central question about laws is what special aspects of reality they capture: answers have ranged from relations of ideas, to structures of social constraint, to relations between objective universals. On the pragmatist approach, in contrast, one begins with the question of lawful *purport*, as Marc Lange puts it; one begins with what special epistemic function lawlike generalizations serve.[8] While one can then go on to ask of such generalizations what aspects of reality they capture (a question to which there may or may not be a non-trivial answer), the central issue that demarcates something as a law instead of some other sort of theoretical generalization is given in terms of the role the claim plays in the functional structure of epistemology. On the first approach, metaphysics constrains the practice of theory. On the pragmatic approach, the epistemic role served by law-claims places constraints on what answers can defensibly be given to metaphysical claims: reflective epistemology constrains metaphysics.

Now those attracted to the metaphysical approach will—and should—find the idea of defeasible laws, if not strictly impossible, deeply suspect. Take David Armstrong's view of laws, for example, according to which laws are grounded in identities between universals. To accept the idea that laws could nonetheless be defeasible, one would be forced to adopt a very strange sort of contingent-identity view of universals; and while this might be possible, it is hard to see the motivation for such technical gymnastics. The universals, one would think, being what they are irrespective of contingency and context, are either identical once and for all, or not. Hence the laws that describe such metaphysical relations, one would also think, are themselves either absolute or non-existent. On the metaphysical approach to laws, in short, defeasibility looks suspicious indeed.

Not surprisingly, we are no fans of this approach to laws—and not just, as it turns out, for its implications regarding defeasibility, though we can't here defend such an audacious claim. Here our purpose is to point out that the metaphysical approach is not the only view of laws available, and that those who accept a pragmatist approach to laws can—and should—accept the possibility of defeasible laws.

To explain, we turn to the best and most thoroughly developed pragmatist account of laws, that which is provided by Marc Lange. Lawlike

[8] Lange (2000).

generalizations are, of course, a kind of explanatory or theoretical generalization; but they're a special such kind. As Lange reminds us, there are two concrete marks of 'lawlikeness' beyond plain explanatoriness. The first feature is counterfactual robustness. To give an example, the mere fact that every philosopher whose work is substantially involved in the current defeasibility project has initials 'ML' does not in any way imply that such a generalization would obtain in any other possible situation. Acceptance of this generalization in no way commits us to the claim that Wayne Davis could not substantially contribute tomorrow, or that if he did he would change his name. On the other hand, a garden-variety law—say, 'plants require light to grow'—most certainly does have counterfactual implications. Such a law implies that if Wayne had planted an additional rose bush in his yard, it would have needed light to grow. Laws are, as we typically say, at least necessary.

The second mark of lawlikeness, for sufficiently empirical domains, is a particularly forceful kind of inductive confirmability. Lawlike statements are the kinds of generalizations each instance of which receives some degree of inductive confirmation from any confirming instance. Compare, for instance, the lawlike generalization 'all samples of salt dissolve in water' with the non-lawlike generalization 'all the samples of salt belonging to Martha are in her dining room.' If we take any sample of salt, add it to water, and observe that it dissolves, we have provided some confirmation, regarding each instance of salt, that it will dissolve: our rational confidence in each instance is raised by the observation of any one case. In the non-lawlike case, however, no such confirmation is built in. Our observing a bit of salt in a shaker on the table in Martha's dining room does not provide any evidence whatsoever that the box from which the shaker was filled will be there.

So much all can agree to, including those who take the metaphysical approach. In a pragmatist account, though, what it *is* for a given generalization to be a law just *is* for it to be a true generalization capable of serving the specialized inferential functions thus demarcated. Lange fills in this notion by arguing that theories begin with a postulation of an 'inductive strategy'—'a mode of reasoning by which to justify believing in the reliability of a given inference rule.'[9] Part and parcel of what it is to work within physiological biology, for instance, is to take the behavior of one egg from a given animal to be indicative of the behavior of all of them—in similar circumstances, of course. To do so is both to treat generalizations about the functioning of, say, fish eggs, as a law, and to treat fish eggs as forming a kind within the relevant field of study. Broad inductive strategies stand

[9] Lange (2000: 207); the general view is laid out in section 3.1, pp. 207–11.

or fall with the success of the entire theoretical enterprise, and particular laws stand or fall with their ability to be inductively confirmed, given the contingent data, within the framework of a flourishing inductive strategy.

In anchoring lawlikeness to inferential strategies implicit to theoretical enterprises, Lange points out that disciplines exhibit important practical autonomy. Intersecting with strategies in other realms, to be sure, they nonetheless form wholes that answer to their own animating questions, concerns, and aspirations. This point has important implications for the notion of necessity inherent to laws.

In particular, it appears that each theoretically articulated subject-matter determines a range of possible situations that constitute the scope of the necessity operator of that field: the interests and broad explanatory strategy of a field of study mark off certain worlds as relevant and others as, well, 'don't care' or 'off-stage' worlds. Biology, for example, doesn't care about worlds in which all entities are created in a laboratory by robot scientists: its laws do not purport to cover such entities.

As Lange further notes, it is often assumed that the ranges of necessities relevant to the various subject-matters form a nice ordered sequence: psychological necessity requires counterfactual robustness across one range of worlds, biological necessity across a strictly broader range, physical necessity a still broader range, and logical or conceptual or alethic necessity the broadest of all. As Lange explains, though, this is a mistaken assumption. There are often no neat set-inclusion relations between the worlds of interest to a pair of fields. To give one comparison offered by Lange, island biogeography includes among its principles the 'area law': 'that the equilibrium number S of species on a given island is an increasing function of the island's area A, ceteris paribus.' Now there are many physically possible worlds—indeed, many actual situations—which are in the 'don't care' category for this theory: say, islands of frozen methane in a sea of ammonia on a gas giant. At the same time, there are physically *impossible* worlds that *are* relevant to the theories of island biogeography—that is, physically impossible worlds that are nonetheless in the scope of the laws of such theories. If, for instance, the claims of the law turned out not to hold just because instantaneous information transfer were possible between species in distant solar systems, this would implicate the success of the theory, by the lights of its own ambitions.

Note that exactly the same phenomenon applies to the relation between moral and physical theory: the scope of counterfactual robustness of neither is a subset of the other. The same examples will do. Moral theory, we would argue, should disregard thought-experiments about what occurs in all manner of physically possible worlds the inhabitants of which depart sufficiently from humans (creatures incapable of second-personal speech

acts, say). At the same time, if a moral theory's claims on, say, lying could not hold up under postulation of real-time communication across solar systems, it would seem to be a genuine problem for that theory, contra-legal as the case might be to the laws of physics. The laws of moral theory (if any there be, of course) must be counterfactually robust across certain physically impossible worlds. This is a point of relevance, note, to those who worry that morality's metaethical bone fides must be held hostage to recovering a deep isomorphism between moral and natural properties. On a pragmatist understanding of laws, we shouldn't expect moral laws, or moral kinds, to map onto natural ones, because the modal shape of such realms are autonomous.

Before leaving the marks of lawlikeness, we mention one further point that goes rather without saying in Lange's work: laws do not present themselves individually. A theory is always a structure of interanimating laws that can be used together to explain and predict; so, too, inductive strategies must involve postulation of a whole range of inductive inferential proprieties. On the confirmation side, it is only such structures of laws that can be confirmed (one cannot confirm one of Newton's laws of motion in full abstraction from the others, for no predictions follow from one law alone). In general, it is not even possible to *understand* what a law means except in the context of the other laws that, together with it, form a complete theory.

4. DEFEASIBLE LAWS

Can laws be genuinely defeasible on a pragmatist approach? Yes, just so long as one can redeem the ability of defeasible generalizations to serve the key inferential functions marking lawlikeness.

The first worry raised about their ability to do so is one that traditionally haunts deployment of hedged generalizations. Both the counterfactual and inductive marks of lawlikeness require that we be able to distinguish between instances that support and those that contradict a claimed law; how, though, can porous generalizations do this, if they agree from the start that their claimed connection does not always hold? How in the world are we to tell the difference between a counterinstance and a sanctioned exception; how could an instance confirm—and counterinstance disconfirm—a theory's claim if the theory has built into it admission that the connection does not always hold? The worry, in short, is the traditional gripe about irreducible provisos: the generalizations in which they appear seem simply to say that *x* does *y* except where it doesn't.[10]

[10] See Earman and Roberts (1999).

But this, we want to argue, betrays a misconception of what it is one understands in understanding defeasible laws. An epistemic grasp of the operator 'in privileged conditions' is not simply a matter of possessing some list of worlds. (What would the list contain? Names of worlds? Complete—that is, infinite—descriptions of worlds?) Rather, an understanding of the distinction between privileged and deviant emerges out of a substantive ability (often more practical than explicit) to appreciate the upshot of contextual change. To understand a set of privileged conditions is to have a practical understanding of the ways things need to be in order to be privileged, and why this is so. One must understand the *sense* in which the worlds are privileged—that is, the type of explanatory, conceptual, or justificatory priority these situations enjoy over the non-privileged, and, hence, the type of explanatory, conceptual, or justificatory demands that departures from them place on us. Such an understanding cannot be explicated in finite terms—that is the point of saying the enthymematic strategy is here bankrupt. But explicability is not a necessary condition on graspability. Indeed, even the most hard-line contemporary moral generalist will recognize the need for non-explicit judgment in such things as moral perception, conceptual understanding, and conceptual application. Once such elements of Aristotelian skill are on board, there is no reason not to appeal to them in accounting for our grasp of the space of privileging as well.

Once we do so, we are in a position to vindicate the distinction between disconfirming instances and instances that occur in non-privileged conditions, or again between modal counterexamples and countenanced exceptions. We recognize that skillful understanding is required to make out such distinctions. Without an ability to demarcate these cases, there is no way to make use of empirical (dis)confirmation or to judge counterfactual import; but, again, commitment to aspects of reality apprehendable only with skill is already accepted by anyone who takes on Aristotle's point about the irreducible need for judgment in morality.

Indeed, and more to the point for the present discussion, our earlier discussion helps to demonstrate that skillful judgment is needed for appreciation of *any* law, even exceptionless ones. In all cases of lawlike generalizations, we noted, there is counterfactual import, which in turn means that 'thought-experiments' (uses of modal imagination), and not just actual experiments, can serve as counterinstances to laws. If, for instance, we have good reason to believe that metals would often fail to conduct electricity in conditions not realized in reality—say, in a universe with a higher than extant density of galaxies—then the generalization that metals are universally conductive cannot count as a physical law. But given Lange's point that all theories have a circumscribed range of worlds that they answer to, such outcomes always present us with a *choice*: is a universe with a higher

density of galaxies really significant to the animating purposes of physics as an inquiry, or is it a mere logical possibility? The answer to this sort of question can only be given by a substantive and skillful grasp of those animating concerns and purposes.

In the case of exceptionless generalizations, then, it's true that the on-stage worlds are just 'those in which the laws are all true;' but this can't be our full epistemic grip on the theory, or else falsifying empirical evidence would, by definition, just indicate that the condition it came from was off-stage. Put differently, the distinction between on- and off-stage worlds must be substantive, on pain of rendering necessity claims trivial by rendering the scope of worlds defining the necessity operator as those in which the generalization holds true. The mere fact, then, that a thought experiment provides a logical counterinstance to a law does not yet entail that there is a defect in the claimed law, for we must determine whether the instance is in a 'don't-care' world.

Whether laws are exceptionless or defeasible, then, we need skill to be able to distinguish relevant from non-relevant worlds. Defeasible laws, to be sure, add *another* distinction, another key partitioning of possibility space—namely, the distinction between privileged and non-privileged conditions. But skills are already irreducibly needed to make out the first distinction, so one cannot cry foul with their necessity for the second. In both cases we require an antecedent understanding of the distinction; in neither case can this understanding be captured as a grasp of a fully explicit part of the theory.

How, more specifically, then, do defeasible explanatory generalizations meet the two marks of lawlikeness? Consider first counterfactual robustness. Laws, we remember, must be able to guide us across counterfactual situations. If we know that 'all *A*s are *B*s' is a law, then we know that *A*s would be *B*s were things to differ in various ways: *A*s would still be *B*s across any of the changes in background situations that are in the realm of significance determined by the subject-matter at hand. Lawlike defeasible generalizations (more precisely, our understanding of them) yield precisely such counterfactual guidance; but the counterfactual robustness comes in layers. The most immediate import concerns what happens in privileged conditions: if a fish egg were to exist in privileged conditions, it would become a fish. To understand the counterfactual implications of such a law in non-privileged conditions, one will need to understand the theoretically relevant ways in which things can depart from privileged conditions. Those who have a full grip of the kind of theoretical backdrop needed to understand a given defeasible generalization in the first place will thereby have a grip of the counterfactual implications it has in non-privileged as well as privileged cases.

What about inductive confirmability?[11] Given the kind of counterfactual import had by defeasible laws—viz., 'In privileged conditions, if *A* were to be the case then *B* would be the case,' the question of what counts as an instance of a law is a bit more complicated than in the case of a non-defeasible law. In the case of a non-defeasible law–say, 'all metals conduct electricity'—any instance of a metal thus conducting in the actual world (as well as any such instance within the range of physically relevant worlds) counts as an instance of the law; and any non-conducting instance scores as a counterinstance. In the case of a defeasible generalization such as 'defeasibly, fish eggs turn into fish,' the actual *instances* of the generalization are the cases in privileged conditions; but the whole carries testable, if indirect, implications for deviant worlds as well.

Suppose, for instance, we find a fish egg that doesn't turn into a fish. Is this a disconfirming instance of the purported law? It depends. If it's an egg that progresses in the 'normal' expression of its genetic structure, from a normal progenitor pair of fish, and ends up as, say, a reptile, then our law is refuted: such evidence would show us that there is something seriously wrong in our biological conception of fish. But if the egg is progressing in the standard manner until it is eaten by another fish, progressing into ever simpler proteins in the digestive system, this is not a disconfirming instance, since the failure merely marks a departure from privileged conditions. Indeed, the right sort of departure can even be confirming of the broad theory of which the defeasible generalization is a part.

Given that we can find confirming and disconfirming instances of a defeasible generalization, can we retain the idea that a confirming instance provides some confirmation of all instances? We see no reason why not. Observing a fish egg maturing into a fish confirms that this is what they all do—all in proper conditions, that is. Similarly, observing that a lie damages one's human relationships, or that a public insult is degrading, will serve as evidence, regarding each explanatorily central context, that other instances would do likewise. The fact that there are non-central conditions—Nazis to whom the truth is not owed, situations in which we are playing the game Diplomacy, Comic Roasts etc.—in no way interferes with this.

[11] A question whose relevance to morality depends on one's metaethics. Some substantive views of moral facts preclude the idea of empirical confirmation altogether: various sorts of rationalism, constructivism, etc., imply that moral facts are not the sorts of things that can be taken up receptively from the world. If that is right, then of course no moral generalization can be confirmed empirically. Our goal here is not to argue against such views; what we want to claim is that, for those metaethical views amenable to the idea of moral inductive confirmability, defeasible laws can do the trick as well as exceptionless ones.

Defeasible generalizations, then, can play laws' characteristic theoretical role of charting the relations and connections between situations. While an exceptionless law ties together facts across a range of subject-matter-relevant worlds in the most straightforward way—namely, by stating something true in each, defeasible laws chart a more complicated function across worlds. *P* is true in the privileged group; there is then a comprehensible function from deviations from privileged worlds to changes in *P*. In both sorts of laws, understanding comes from appreciating the fact that the generalization in question is essential to the course one charts across possibility space. Defeasible principles are as fit for lawlikeness as are non-defeasible ones.

Indeed, one paradigm of legal respectability—idealization laws—are, we would argue, simply a special case of defeasibility. Consider ideal gas laws. The principle $pv = nrt$ is not literally true of any actually existing gas. This is not, of course, because the actual world is of no interest to us in physical theory (it is, indeed, paradigmatically interesting); but the law is only literally true of a certain sort of privileged condition: one in which gases are made of perfectly elastic, perfectly spherical particles. As in the case of all defeasible laws, to know how to apply this law is to be able to understand the salient ways in which a given condition departs from the 'ideal' and to work out the difference that such differences make. Of course, in idealizations, the relation between the ideal privileged condition and the deviant conditions of the actual and near-possible worlds is one of *approximation*. Idealizations posit some sort of monotonic function from similarity of situation to nearness of approximation: as the situation gets more and more similar to the ideal one, the law must provide a closer and closer approximation to actual fact.

But if approximation is the most familiar of the relations between privileged and deviant worlds, it is a mistake to think it the only, or the most important. In morality, for instance, idealization need not be the order of the day. When thrown into the game of Diplomacy or a nightmare authoritarian world, the moral behavior prescribed by our laws need not *approximate* that of privileged conditions. Nor need closeness of situation relate to closeness of moral import in a monotonic way. Sometimes, getting more like normal conditions will make the morally appropriate actions more distant. As in the famous problem of 'second best,' adding a *further* difference sometimes allows us to better approximate appropriate privileged moral behavior. The function from situation to moral implication must be epistemically tractable; but approximational functions are by no means the only sort of graspable such function.

What, finally, about theoretical interanimation? Nicely enough, it turns out, defeasible laws enjoy richer such animation than do exceptionless ones. Theories in general, we remember, do not involve just a single law;

nor are laws confirmed in isolation. Newton's laws of motion come as a package; they jointly imply observable predictions and together demarcate a set of physically possible situations. With exceptionless laws, there is an important sense in which their interanimation is quite simple: they give us a set of exceptionless generalizations that are claimed true across a particular range of worlds—namely, the worlds of interest to the subject-matter of the theory. With defeasible laws, in contrast, the very ways in which the laws interanimate are themselves subject to emendations across deviations to privileged worlds. Even though the structural interrelations between the laws may be revised in the process of revising the generalizations they relate, these interrelations still place substantive constraints on an acceptable application of a moral theory to a non-privileged situation.

To illustrate, let's consider one of the (rather more provocative) cases we've discussed in earlier work, namely, the sexual practices of S&M. In 'standard' (non-S&M) conditions, it is plausibly a law of moral behavior that one should take others' statements about their own desires at face value: 'No' means no! Standardly, when a sexual partner tells you to stop doing something, or says that he wants you to stop, it would be an assault on his autonomy to continue. In the practice of S&M, though, 'Please stop, I don't like that!' is rightly (in both the epistemic and the moral sense) taken to indicate that one's partner enjoys what is happening. The valence of *not* taking someone at their word, note, displays a justificatory dependence on its paradigmatic valence: it is only *because* we have willingly consented to be engaged in such a practice that it is possible for it to be morally good to treat 'No, please stop!' as an indication of desire to continue; and this consent must itself be understood as the stating of one's desire given in a context in which the *normal* valence of taking others at their word holds. (If someone were to say, 'I don't want to engage in S&M play' and you took this as meaning that they do want thus to engage, this would not be a morally acceptable hermeneutic.[12])

The move to the S&M context, then, switches valences of certain morally significant features of acts. But crucially, the practice of S&M is a complex one.[13] In particular, while it is indeed constitutive of the practice that

[12] More precisely, it could be. One could be engaged in a consensual context in which one debates what context to be in with altered signals. But that too is only acceptable if one has consented. There can be any number of layers of nested non-privileged conditions of this sort, but they need to bottom out in consent in a privileged condition else the valence of interpretation switches in none of them. Another point here is that consent, of course, need not imply explicit statement of rules.

[13] The practice of S&M is not some simple variant in which one finds pleasurable something that most people find painful; the practice involves a complex erotics of control and domination, of giving and withholding of pleasure, and—only within

the usual interpretations of kindness and respect for autonomy change, they do not, for all that, go right out the window. Take, for instance, the practice of 'safe words' and other limits. While one can in standard S&M practice take another to really want the opposite of what she says, one should stop one's actions when the specified code word is used, and one shouldn't assume that a partner wants something done that would cause permanent injury (at least in the most privileged versions of S&M, there being variants upon variants upon variants). Similarly, though one might well believe that domination is here an expression of love and respect for the other's autonomy—as opposed to contravening autonomy, as would be the case in privileged conditions—this, too, has its limits. One dominates within the bounds of the practice, for purposes of mutual pleasure, and insofar as it plays out a role that the other willingly takes on. And though one might well cause pain—genuine pain, mind you, not merely pleasure in a sensation that others find painful—this does not remove the obligation of kindness. Indeed, in the S&M room, some givings of pain are reportedly seen as expressions of kindness—another example of explanatory valence-switching.

More specifically, then, here we have two defeasible moral principles interacting: defeasibly, kindness is a good-making feature of acts, and defeasibly, respect for autonomy is a good-making feature of acts. In privileged conditions, the morally wise agent knows how to balance, refine, or specify these principles in relation to one another: she knows how to respect autonomy in light of the value of kindness and many other principles, how to express kindness in the way least hostile to another's autonomy, how to make use of both ideas in an effort to construct a sophisticated understanding of a situation from which they can perceive what to do. In moving to the valence-switching context of the non-standard S&M practice, she knows that many things change—including what counts as respect for autonomy or an expression of kindness, and whether in various instances kindness or respect for autonomy are good-making; but neither norm is dropped, and their interanimation remains. One can still, in the practice of S&M, be immorally disrespectful or unkind; and the quite different ways in which these virtues express themselves remain linked. There are still ways in which the nature of kindness is tied intimately to the ways that we are reacting to another's autonomy.

Interanimation, in short, is particularly rich in systems of defeasible generalizations. Not only should our theoretical understanding of the

this—a complexification of the relation between pleasure and pain. An extensive bibliography of the (highly uneven) scientific literature on BDS&M and related practices can be found at **www.datenschlag.org/english/bisam/index.html**

relevant moral terrain grant us an understanding of the moral differences that various departures from privileged conditions make, it should see these changes as evolving holistically. One does not proceed in some lock-step fashion, first seeing that we have departed from privileged conditions in a certain way, then figuring out what difference this makes to the norms surrounding kindness, then going back to see what difference this in turn makes to riffs on the norms surrounding autonomy. Rather, one begins with a complete, theoretically robust understanding of how to behave in privileged conditions and works out in a uniform way the deeply interconnected differences that ramify through the theoretical structure in response to a given departure from privilege.

5. DEFEASIBLE KINDS

Kinds are coordinate with laws. Kinds, that is, are the subject-matters of laws and determine the extension to which the laws apply. If the presence of genuine laws in a moral theory implies the presence of genuine moral kinds, the possibility of defeasible laws introduces the possibility of *defeasible kinds*.

In the case of a non-defeasible law, counterfactual robustness entails that there is a sense of necessity such that, for any law linking kind K and property F, it is necessarily the case that all Ks are F. It thus makes sense to speak of this connection as being *of the essence* of kind K.[14] In the case of defeasible laws—say, defeasibly, all Ks are F—the counterfactual implication is not that all Ks in all situations of interest to the subject matter are F, but merely that all Ks in all privileged conditions are F. Thus being-F is not of the essence of the kind K—but, we want to insist, it is perfectly intelligible to say that being-defeasibly-F is.

On a pragmatist approach, we explain kind concepts, as we do the lawlikeness of generalizations, in terms of the role they play in broad theories. Kinds, that is, are simply whatever in the world answers to the concepts playing that role, just as laws are simply the true lawlike generalizations. If we take then a generalization of the form

- $P(\forall x)(Kx > Fx)$, where P is the relevant privileged conditions operator, and $>$ the subjunctive conditional,

[14] Once this fairly innocuous version of essence talk is on the table, there are others who will want to up the ante and begin talking of individual essences as well, features that this very individual must have in order to exist, simply qua individual. We have nothing to say on this topic at present.

then any K, whether in privileged conditions or not, is of a kind that is constitutively such that in privileged conditions it is F.

Precisely what this implies about instances of K that are in non-privileged conditions depends on the sort of privilege at issue with a given defeasible generalization. K may be what we have called a 'paradigm-riff' concept—such as 'chair,' whose paradigm cases are to be sat upon, and whose non-paradigms, such as frail, artsy chairs in the museum display, are understood as chairs in virtue of the way they function as a riff on privileged chairs. Here, the implication is that the kind is a functional one: what it is to say that Ks are essentially defeasibly F is to say that what it is to be a K is to be the sort of thing which functions either as F, or as something that serves, in context, as a suitable variant on the kind of thing that is in a privileged context and is F.

If, instead, the defeasible law reflects an explanatory dependence, the relevant role of the kind is an explanatory one. Take the example of pain. Defeasibly, pain is bad-making; defeasibly, only, since it can shift valence, as in Elijah Milgram's nice example of pain in the context of athletic challenge. To understand pain's nature is to understand not just that it is sometimes not-bad, but to understand that there is an explanatory asymmetry between cases in which it is bad and cases in which it is not: it is only because pain is paradigmatically bad-making that athletic challenges come to have the meaning they do, and hence provide the kind of rich backdrop against which instances of pain can emerge as not-bad-making, as not always and everywhere to-be-avoided. Our understanding of pain's switched valence, as we might put it, is ultimately redeemed by reference to its normal valence.[15]

In saying of an un-bad case of pain that it is nonetheless a case of pain, we identify it, in short, as having a kind of character that is bad in those situations which are explanatorily basic for the determination of the import of the kind in any situation. Conceptual deployment of this pragmatic role is to commit oneself to the space of privileging characteristic of the defeasible connection between pain and bad-making; the essentiality of the connection between the law and the kind lies in the fact that such a commitment is involved in the attribution of the kind-concept.

[15] Indeed, and once again, the *way* in which such a situation differs from the privileged conditions—namely, that it is a context in which we are engaged in a practice one of the central points of which is to strive to endure beyond normal human abilities—tells us *how* the defeasible bad-makingness of pain plays out there. Thus pain caused by a vicious and illegal tackle that takes someone out of the game is not nobility inducing, while the pain of pushing oneself on to achieve in the face of exhaustion is. Once again, someone who knew only that 'pain was bad-making in various privileged conditions but not in sport' would not have understanding.

6. DEFEASIBLE THEORIES AND OBJECTIVITY

Suppose we have convinced you that a theory can comprise defeasible laws, and that these are even compatible with kinds having defeasible essences. Our final question is whether there is anything general we can say about the sort of theory in which such defeasible laws are to be expected. The presence of defeasible laws, we have said, implies the presence of a normative structure of privilege and coordinate revision among possible states of affairs. For a field of study to produce a theory with defeasible laws just *is* for it to see the space of relevant possible situations to be structured in such a normative manner. So when would it be plausible to think of situations as being thus structured? When, for example, would it make sense for us to think that a kind is defined asymmetrically by its manifestation in one sort of circumstance? Most plausibly, we want to suggest, only in contexts that are at least—and at most—subtly dependent on human interest.

Consider two extremes. Let's take a game that is 'shallowly conventional'—that is, a game comprising rules we simply make up, with no reference to any deeper point or *telos*. One could easily imagine coming up with a set of *exceptionless* rules or 'laws' here—say, queen always trumps king; but it would be strange indeed to think of the queen as ruled by a defeasible rule or law, for it would be hard to understand by what lights we would regard certain conditions as privileged, on what grounds we would determine the import of various sorts of non-privilege. Such rules would, we suspect, end up having the content: 'Queen trumps king except … sometimes.'

At the other extreme, imagine some aspect of reality that is altogether independent of human interests (assuming, as we don't, that there is such a thing). The kinds present in this aspect of reality just 'are what they are.' Once again, one could imagine coming up with an exceptionless law describing such aspects, but it's odd to think of such kinds as governed by defeasible laws, structured around notions of privileging. Perhaps there can be some sort of explanatory or other normative structure just out there in the fabric of possibility-space-in-itself; but we'll admit to being drawn towards 'incredulous stares.' To say that there is such a normative structure to possibility, after all, would be to say that there are instances of a kind K in world w' which lack some of the features had by all instances of K in w, and yet which are of the same kind. Why would that be? Apparently, because w is a privileged world and instances of kind K in w' are essentially riffs on the instances of K in w. But what would constitute

the difference between *this* metaphysical kind-identity, and a structure in which these things simply aren't instances of K in w' in virtue of their lacking the features consistently had in w? It seems to us that it would be a queer property indeed that was simply a brute feature of possible worlds—independent of interest or socially instituted *telê*—in virtue of which some worlds were privileged for defining various kinds. We find such brute metaphysics—even more, anyone's claimed ability to know its contents—baffling.

Our bafflement is alleviated, however, once we begin to see kinds as arising within fields constitutively yet robustly structured around human interests. Return, first, to a game, but this time one structured around some overriding point—some discernible *telos* that motivates both a common identification of certain things as the same game across contextual changes and also particular determinations of the differences various changes make. In this case, we might well begin to determine a privileging structure inherent in our practice of instituting the kind. Our purpose in playing soccer is to play a sport with a particular sort of athletic difficulty, with a particular esthetic, with a particular kind of intelligible tie to a history of soccer, etc.; together, these constitute a point or *telos* we value enough to continue to play soccer even if we move to an area, say, with no level fields. Soccer in a bumpy or sloped field is not paradigmatic soccer, and compensatory adjustments are made. (The fact that one side has an advantage of running downhill, for example, is a reason to give them a narrower goal to shoot at since part of the *telos* of soccer is that the sides have an equal chance of scoring.) Thus, the presence of a *telos* within a socially arbitrary practice begins to motivate treating one sort of circumstance as privileged and understanding others in terms of that.

Such *telê* are present, we would argue, not just in games, but in theoretical enterprises of inquiry such as biology. Much of our attempt to understand the world is oriented around postulation of specific explanatory strategies. In biology, for instance, we come to understand organisms in terms of a privileged progression through a course of life, and then to account for deviations from this course in terms of specific differences. There is much one could say about why we employ this explanatory strategy, about why we take up this stance towards happenings in our world; but take it up we do. This does not make biology facts 'socially constructed,' of course: there is still a world that must cooperate with our theories; things turn out to fit our accounts or not, quite independently of our desires. But the sorts of kinds we employ are nonetheless motivated and defined by their function within a structure of explanatory interest. We can here make sense, then, of defeasible laws (such as those about fish eggs), for the enterprise is governed by an animating *telos* which itself can structure, in

a non-trivial way, a partitioning of possibilities and exceptions in robustly guiding ways.

A similar story can be told of our idealized gas laws. As we said, we use idealized laws when we find that we can get approximate predictions in actual situations by looking to their similarities to idealized situations. So the concept of a gas—one that includes ideal gases as privileged exemplars and actual gases as non-privileged—is a concept that is motivated by no more idiosyncratic an interest than our interest in prediction. Indeed, we believe, it is *only* in virtue of this interest that demarcating into the kind makes sense. Imagine that ideal gas laws turned out to be the only good ways to predict the behavior of gases in our world. If there really were no better theory possible, would we have a case of privilege-in-itself? We doubt it.

Imagine two different accounts. On one account, there is a kind 'gas' which includes both privileged ideal gases and non-privileged actual gases. They are structured around a space of privileging in which proximity to ideal worlds guarantees that the laws of ideal gases approximately capture the behavior of actual gases. Thus, we take the actual gases to be bound by the same defeasible laws—*def* ($pv = nrt$)—that govern the ideal.

On another account, the following holds true. In some worlds, there are things called 'ideal gases;' these obey the ideal gas law in its exceptionless form. In other worlds, there are things that are gas-like but that don't obey those exceptionless variants. We could call them 'non-ideal gases;' but, since they are governed by different laws, they are not the same kind of thing as ideal gases. While there are no laws that accurately predict the way these behave, it turns out that, by engaging in certain creative fictions, we can pretend that they are ideal gases and have a theory that, while false, is instrumentally useful.[16]

It is hard to see why we should believe in a purely objective fact of the matter—a fact, that is, independent of any relation to human interest—which could constitute the truth of one or the other of these accounts. Against the background *telos* of prediction, though—if we define the practice within which laws are postulated as a predictive one—then, given the superiority of the idealization theory, we can see the motivation for believing in a category of gases. It is easy to see that there may be

[16] It seems to us that this last characterization is the one that Nancy Cartwright opts for in her *How the Laws of Physics Lie* (1983). She claims that exceptionless laws are never literally true for actual entities, except in very specialized cases of laboratories. So, they are merely instrumentally useful and not true. The possibility that the natural kinds of physics are projected against the background of a human interest in prediction, an interest that institutes a privileging structure on possibility space and thereby makes intelligible a defeasibility reading of the laws, is one she seems not to consider.

'predictive-kinds,' or conceptual functions within an essentially predictive practice, that are verifiably governed by the defeasible gas laws; abstracting from this interest, though, no such kind emerges. An omniscient observer with no interest in prediction—Boethius's God looking at the whole temporally extended world as a unity—would, we think, see no reason to choose between the two interpretations of possibility space.

We want to suggest, then, that defeasibility emerges only in the context of practices and fields of study constitutively governed by *telê* that are neither brutely independent of human interests and values nor shallowly dependent upon them. Such a claim, crucially, does not deny objectivity in any of the senses worth caring about: it does not preclude the existence of empirically inductively verifiable, explanatory, theoretically interanimating, counterfactually robust laws, even laws that determine kinds.

If this is right, then deep moral contextualism, it turns out, is consistent with deep moral theory. Indeed, even if *all* explanatory generalizations in morality turn out to be ineliminably exception-laden, morality can still be a terrain governed by theoretical generalizations and, indeed, laws. To be sure, such a commitment requires us to see morality as ultimately and essentially structured around human interest, but this does not render morality the poor second cousin to reputable enterprises such as biology; it is, we venture, something best taken on board in any event.

REFERENCES

Cartwright, Nancy (1983) *How the Laws of Physics Lie* (Oxford: Clarendon Press).

Crisp, Roger (2000) 'Particularizing Particularism,' in Brad Hooker and Margaret Olivia Little (eds.), *Moral Particularism* (Oxford: Clarendon Press).

Dancy, Jonathan (1993) *Moral Reasons* (Oxford: Blackwell Publishers).

____ (2000) *Practical Reality* (Oxford: Clarendon Press).

____ (2004) *Ethics without Principles* (Oxford: Clarendon Press).

Earman, John, and Roberts, John (1999) ' "Ceteris Paribus," There is No Problem of Provisos,' *Synthese*, 118/3: 49–78l.

Hooker, Brad, and Little, Margaret (2000) *Moral Particularism* (Oxford: Clarendon Press).

Lance, Mark, and Little, Margaret (2004) 'Defeasibility and the Normative Grasp of Context,' *Erkenntnis*, 61/2–3: 435–55.

____ ____ (2006a) 'Particularism and Antitheory,' in David Copp (ed.), *Oxford Handbook of Ethical Theory* (New York: Oxford University Press), 567–94.

____ ____ (2006b) 'Defending Moral Particularism,' in James Dreier (ed.), *Contemporary Debates in Moral Philosophy* (Oxford: Blackwell Publishers).

Lange, Marc (2000) *Natural Laws in Scientific Practice* (New York: Oxford University Press).

Little, Margaret (2000) 'Moral Generalities Revisited', in Hooker and Little (2000), 276–304.
McKeever, Sean and Ridge, Michael (2006) *Principled Ethics: Generalism as a Regulative Ideal* (Oxford: Clarendon Press).
McNaughton, David (1988) *Moral Vision* (Oxford: Blackwell Publishers).
Murdoch, Iris (1970) *The Sovereignty of Good* (Oxford: Blackwell).

8

Practical Reasons and Moral 'Ought'

Patricia Greenspan

Morality is a source of reasons for action, what philosophers call practical reasons. Kantians say that it 'gives' reasons to everyone. We can even think of moral requirements as *amounting to* particularly strong or stringent reasons, in an effort to demystify deontological views like Kant's, with its insistence on inescapable or 'binding' moral requirements or 'oughts.'[1] When we say that someone morally ought not to harm others, perhaps all we are saying is that he has a certain kind of reason not to, one that wins out against any opposing reasons such as those touting benefits to him of ignoring others' concerns.

Philosophers may feel the need for a deeper understanding of reasons, but interpreted essentially as facts relating acts and agents—considerations counting in favor of or against someone's performing a certain act—moral reasons at any rate would not seem to involve any intrinsic moral properties of acts, of the sort that people used to worry about even for less extreme examples than Kant's of a deontological approach to ethics. We need to refer to reasons in any case to understand ordinary non-moral cases of rational deliberation and action. So it is now common, for instance, to substitute for Ross's notion of prima facie duties talk of pro tanto reasons, reasons counting in favor of or against some act as far as they go, but capable of being defeated by opposing reasons.

The explanation of moral 'ought' in terms of practical reasons might seem to lend support, though, to contemporary Kantian arguments that practical rationality is all one needs to supply the impulse to be moral, with Nagel's *The Possibility of Altruism* as a primary source.[2] Even granting that an agent might be rational and yet not fully aware of the reasons bearing

[1] Because the term 'obligation' has some implications that do not pertain to 'ought,' let me deviate from English idiom and use the verb 'ought' as a noun. An ought is what we have when the verb applies, i.e. when we ought to do something. Henceforth I use quotes only when referring to the word or concept.

[2] See Nagel (1970).

on a particular act (he might, for instance, be unaware that a certain act would cause others harm), if he is aware of a reason, how could he possibly justify a failure to be moved by it, except by appeal to opposing reasons he considers just as strong?

Some recent work on practical reasons weakens the force of a reason, in effect, by defending a subclass of *optional* reasons—reasons such that knowingly failing to act on them, without any equally strong opposing reasons, is compatible with practical rationality. In Joseph Raz's terms, reasons as such do not *require* action but merely render it 'eligible' for choice.[3] From the standpoint of rationality, then, not all undefeated reasons are compelling reasons. Some authors go further and assign a lesser normative force to certain reasons: what Jonathan Dancy marks off as merely 'enticing' (as opposed to 'peremptory') reasons, and Joshua Gert calls 'justifying' (as opposed to 'requiring') reasons.[4] If moral reasons, or even just some of them, are rationally compelling—inescapable in the sense of demanding obedience of all rational agents, as well as *applying* to all of them—we need to do more than insist on their status or strength as *reasons* to explain why.

We might just insist that their status as *moral* reasons is enough to make them compelling. But left in such general terms, this strikes me as mere table-pounding that at best is a last resort. Instead, I would hope that an account of what is involved in rationally *discounting* certain reasons would enable us to pinpoint the fundamental error (as opposed to the irrationality) of someone who recognizes moral reasons but is not motivated by them—what I call a 'reasons-amoralist.' I have a somewhat different way of making out optional versus compelling reasons—in terms of a conception of practical reasons as offering or answering criticism—that will support such an account. It should still allow us to use the notion of a reason to capture binding moral oughts, on a deontological view more or less in the spirit of Kant, but without any claim that an agent who deliberately flouts a moral ought must be irrational.

In short, then, my aim here is to defend the interpretation of strong or binding moral 'ought' in terms of practical reasons within an appropriately loose general conception of practical reasons. My strategy is, first, to sketch the main lines of a 'critical' conception of practical reasons that allows one to recognize some consideration as a reason while turning it down as

[3] See Raz (1999: e.g. p. 65).

[4] See Dancy (2004) and Gert (2004: chs. 2–6). My own view overlaps with both authors', particularly Gert's. However, Gert understands reasons in terms of rationality and takes the latter notion as ruling out mistakes about one's reasons (cf., e.g. his treatment of Scanlon on irrationality vs. mistake on p. 215). I discuss differences from Dancy in Greenspan (2005).

a motive. Then, in the central argument of this paper, I show how the view can handle the reasons-amoralist. I go on to answer a different but related challenge to the attempt to understand 'ought' in terms of reasons, as suggested by recent work on undetachable conditional oughts.

REASONS, DISCOUNTING, AND ENTITLEMENT

My central argument amounts to a defense of externalism—in several varieties, which I attempt to sort out in my next section, but in the first instance, reasons externalism, since it lets a rational agent simply reject some acknowledged reasons as motives. Bernard Williams's defense of reasons internalism ultimately turned on insistence that the notion of a practical reason made sense only as a potential motivator.[5] But there is an alternative conception of practical reasons that loosens the tie to motivation, even granting that the usual point of acknowledging a reason is indeed to motivate—to guide or influence action, one's own or others'. What is essential to a practical reason on the critical conception is instead a relation to criticism: a practical reason serves either to offer a criticism—meaning a potential criticism, not necessarily one that is put to the agent—or to answer one, by citing some valuable feature of the act or other practical option in question. The normative role of a reason is thus either critical or defensive—or some combination of the two. This is in contrast to a common conception of practical reasons as essentially action-guiding, which I think Williams assumes. More generally, the critical conception represents an alternative to understanding reasons in terms of ends, whether an agent's actual ends or some independent notion of what has value.

The critical conception instead emphasizes disvalue by shifting the normative spotlight to negative reasons—reasons against, or one might say 'cons.' It makes out the normative function of positive reasons (reasons in favor, or 'pros') in terms of what negative reasons supply, namely criticism. Though discussions of practical reasons usually focus on examples of positive reasons, this shift fits well with ought-based approaches to morality, since requirements, though expressed in positive form, have to be explained in terms of negative reasons, considerations counting against alternatives to the acts they require.

To illustrate with a non-moral example what I have in mind by a negative reason, consider the reason commonly cited against smoking: that it causes

[5] See Williams (1981: esp. pp. 108–9; cf. 1995: 39). Korsgaard (1986) formulates internalism to require only that reasons have the *capacity* to motivate, but she interprets this as making an exception only for irrational cases.

cancer. Note that what is said to be negative here is neither the content of the reason (the relation of smoking to cancer is not naturally expressed by a negation) nor the act that the reason is cited against (smoking is not a mere omission), but rather the *bearing* of the reason on the act: that its relation to cancer does or would count *against* smoking. On the kind of reasons externalist view I favor, it would have that bearing even without having action-guiding force for a given agent—if she were indifferent to the prospect of contracting cancer, say, and lacked any other desires that would be frustrated by it, or at any rate weighted them less heavily than the advantages of smoking—as long as ill health could still be said to frustrate her interests.

On the basis of the same sorts of considerations, I also have a reason to get a certain amount of exercise; this is stated in positive form, but it is negative in my sense insofar as it counts against some alternative that excludes it, such as leading too sedentary a life. Indeed, any reason capable of generating a practical requirement has to be seen as negative in this sense. To apply this to a moral example: an altruistic reason in a given case has to count against acting solely in self-interest, if it is to yield anything stronger than a *recommendation* of altruistic action. To grant this is not to deny, though, that reasons that simply cite valuable features of acts or other practical options might play an important role in morality, particularly in relation to the virtues, as ideals of human behavior. In motivational terms, as incentives to action, they may be at least as important as negative reasons. I think of them as 'purely' positive reasons, counting in favor of an option but without implying significant criticism of alternatives. I assign them a secondary role, though, in moral or other normative systems meant to supply a standard of correctness in action, not just a scale of better and worse. In the first instance, on my account, purely positive reasons serve to ground permissions, defending the favored option against whatever criticisms it might be subject to itself, and supporting recommendations insofar as some options are more defensible than others.

I limit attention to ought-based ethics in the discussion that follows, though my comments should apply to any version of ethics that can generate practical requirements. If we allow for optional reasons, eudaimonism and similar views that might be thought to be based on purely positive reasons would have to allow for an important, if implicit, element of negativity in my sense—to count a life lacking in *eudaimonia* as deficient and thus to be avoided—if they are to generate anything stronger than practical recommendations. My concern in this paper is just to handle a problem posed by optional reasons for views that attempt to make reasons the basis for strong moral 'ought.'

To minimize verbal complexity I also make a number of other simplifying assumptions in what follows. For instance, though I am working from the

standard view of practical reasons as objective—as amounting to facts independent of the agent's beliefs about them—I sometimes follow our natural way of speaking and refer to reason-judgments or -statements as reasons (not always spelling out 'practical' reasons), on the assumption that they fit the facts. My talk of reasons on the critical conception as 'offering' or 'answering' criticism is a case in point: more strictly I should say that reasons can be cited as part of a criticism or in answer to a criticism, or that they ground or are based on or amount to criticisms or answers to criticism, but these longer-winded formulations are clumsier and less perspicuous.

Elsewhere, focusing on non-moral cases, I introduce the critical conception as a general view of practical reasons and begin to answer some of the many likely objections to the distinction I derive from it, between purely positive and negative reasons.[6] The distinction is easily misunderstood, in part because these terms might seem to make it out as a distinction in surface form. Though I introduce it as a distinction between reasons in favor and reasons against, my treatment of requirements should make it clear that some reasons naturally stated in positive form really imply negative reasons and hence are not *purely* positive in my intended sense. Indeed, the logic of reasons would let us restate even purely positive reasons in negative form, since a reason *for* something implies at least a trivial reason *against* something else, namely omitting it.

For a simple example of reason that would count as purely positive in my sense—later I introduce a meatier case and focus discussion on variants of it—consider my choice between two blazers in my closet that differ only in color. Supposing that green happens to be my most flattering color, this counts as a reason in favor of choosing the green; and trivially, of course, it yields a reason against *not* choosing the green, which counts against choosing any other blazer, if we rule out wearing two blazers at once. However, the blue blazer also looks perfectly fine on me, so on a day when I have no particular reason for looking my absolute best, it would seem to do just as well. The fact that the green would look better does not yield any significant criticism of choosing the blue, of the sort that would keep it from counting as a purely positive reason on my understanding of the notion.

While recognizing problems with this semi-technical use of common terms, I think I do need something of the sort to convey the distinction I have in mind, and the only alternatives I can think of seem to be either no less technical than 'positive/negative' or more seriously misleading in application to moral cases. But since some readers might find a less formal way of representing the distinction helpful to keep in mind, let me mention two other possibilities. We might, for instance, recast the distinction in

[6] See Greenspan (2005), (unpublished-a), and (unpublished-b).

terms drawn directly from the critical conception of reasons and contrast 'defensive' with 'critical' reasons. However, I think these terms for different sorts of normative force would be misleading in application to cases insofar as they ignore the motivational aspect of reasons. (The recommended motive for altruistic action, say, is not to defend oneself against moral criticism.) Instead, we might modify Raz's talk of (positive) reasons as rendering options eligible for choice and distinguish between 'qualifying' and 'disqualifying' reasons. But that would be misleading in some ways too, since part of my point is that reasons counting against an option *tend* to disqualify it—to rule it out as unworthy of choice—but would not actually succeed in doing so in cases where the agent legitimately discounts them. So instead of switching terminology, let me stay with 'positive' versus 'negative' and invite the reader to fill in either of these alternative formulations, if it seems to convey more.

I call a reason purely positive, then, in cases where it tends to qualify an option for choice without disqualifying any competing options. This presupposes a threshold of adequate value, so that competing options may still be accepted as worth choosing, where they exceed the threshold, even though the reason in question does not apply to them. An example I use elsewhere involves a choice between staying on the Riviera, where I now am enjoying a long-planned vacation, and traveling on to Rome, which I would enjoy even more. A unique advantage of Rome—that the coliseum, which I have yet to see, is there, say—gives a reason in favor of traveling on, but assuming that my current vacation is working out well enough, either choice would be within reason.[7] In representing a certain option as choiceworthy in some respect, a purely positive reason does not represent alternatives as objectionable or problematic and hence does not yield a significant criticism of them; the fact that it fails to apply to them can be said to amount to a reason against them, but only a trivial reason.

In this paper, however, I want to make relatively short shrift of the issues surrounding purely positive reasons in order to focus on metaethical issues raised by negative reasons—reasons of the sort that, if not discounted, would yield requirements. I want to say that such reasons may be *rendered* optional in a given case by the agent's appeal to higher-order reasons to discount them. This is in contrast to simply recognizing a reason as optional in virtue of the sort of bearing it has on action. In the present section I discuss discounting in general terms, moving in my next section to the question of its application to moral reasons.

A purely positive reason—a reason that serves just to answer (potential) criticism of an act or other practical option, without implying significant

[7] See Greenspan (unpublished-b) and (unpublished-c).

criticism of alternatives—is discountable (legitimate to discount) at will. We can think of it as offering an opportunity rather than imposing a requirement, even on the assumption (which I make for all cases of optional reasons discussed here) that it defeats any opposing reasons. In other words, it cites a valuable enough feature of action to answer any applicable criticism, leaving the agent a choice as to whether to act in light of the valuable feature or instead in light of the criticism. So a mere appeal to preference on a particular occasion will be enough to explain a decision not to act on it.

To illustrate this, consider my reasoning a few years ago in response to an administrator who tried to supply a pure incentive for service on extradepartmental committees by citing the possibility of thereby gaining power in the University. I would not deny that the administrator offered me a reason to serve on a committee, insofar as power would be a benefit to me. But citing my lack of interest in power seems to be enough to rebuff his appeal—assuming it does not really mask appeal to something negative, a stick lurking behind the carrot, such as some likely bad consequence of my failure to gain power. This would be so even if we suppose that I have enough time and energy during a given term to add committee service to my other obligations and priorities.

By contrast, discounting a negative reason, as involved in a requirement, needs defense in terms of further, higher-order reasons. Bartleby's line, 'I prefer not to,' will not be adequate, if the aim is to back up the rationality of deciding not to.[8] But one can sometimes give a higher-order reason for 'bracketing' a certain class of reasons. In his early work on reasons and the law Raz explains 'exclusionary' reasons as reasons for excluding certain first-order reasons from consideration.[9] The fact that the law requires something is supposed to block us from placing deliberative weight on reasons that would otherwise count against it. We still recognize them as reasons, that is, but exclude them from deliberation.

An exclusionary reason does not outweigh first-order reasons but rather essentially outranks them (though it might itself be countered by competing second-order reasons). Raz's notion is introduced as explaining the sense in which legal reasons are authoritative, but it also is meant to help clarify various concepts extending to individual practical reasoning. Raz makes out a decision, for instance, in terms of both first- and second-order reasons, or what we may think of as reasons on two levels: at the lower level, a first-order reason in favor of carrying out the decision, and above it, a second-order reason excluding any competing first-order reasons from consideration. Appealing to a decision one has made to discount certain first-order reasons, then, would not necessarily mean ascribing greater

[8] See Melville (1853). [9] See Raz (1990: 37–45).

strength to the competing first-order reason stemming from the decision. What gives a decision 'binding' force is instead the higher level of the second-order, exclusionary reason.

Given Raz's focus on the law, one might think of exclusionary reasons as buttressing the authority of certain first-order reasons. But they do so only by undermining the authority of others—including some we might think authoritative insofar as they otherwise would yield requirements. An example of this is provided by Scanlon's recent suggestion of a 'structural' account of reasons whereby, instead of comparing reasons in terms of strength of desire, we bracket some reasons as inappropriate to a given context—discounting personal concerns, say, such as regard for an opponent's hurt feelings, in playing a competitive game.[10] However, Scanlon's examples of second-order reasons seem to involve *disallowing* action on the first-order reasons in question, rather than making it optional, which would depend on also taking the second-order reasons as optional. His examples also suggest that discounting a reason means denying it the status of a reason—declaring it irrelevant to the choice at hand. On the account I am taking from Raz, all the agent denies to a reason in discounting it is a role in his deliberation, which I treat as tantamount to denying it motivational force. The discounted reason still is acknowledged as justifying action—if the agent should choose to act in light of it, after all.

The critical conception affords a way of granting an agent multiple levels of optional reasons without threat of regress. Consider a modified version of the power case involving negative first-order reasons. Suppose I do need to serve on a University committee this term in order to correct a deficit in my current level of power. How can it still be rational—meaning 'within reason,' whether or not the most prudent thing to do—for me to turn down that option? By hypothesis, I am not in a position to cite equally weighty first-order reasons against it, such as those I might have for instead completing a paper by a deadline this term. Whatever benefits I stand to accrue from completing the paper on time would be less than those of committee service, say. But in turning down the administrator's appeal, it would seem to be enough for me just to cite a decision I have made to stress intellectual aims over political. That would not necessarily satisfy the administrator, but if defense of my rationality is what is in question, I think it is all I need to say, at least assuming that the consequences of my power deficit will not be *dire*. I have a certain leeway, that is, to discount some harms to myself—remaining without input into matters that concern me, such as class size, for instance—in favor of aims I choose to stress.

[10] See Scanlon (1998: 50–5).

What we have here is an intrapersonal analogue of self-sacrifice for others, which comes under the category of supererogation in the moral sphere. I am sacrificing some of my interests to a purpose of my own, and am within my rights to do so, rationally speaking. Raz in fact has a subcategory of exclusionary reasons, 'exclusionary permissions,' that he interprets as entitlements and applies to supererogation.[11] What I would call the source of entitlement in the power-deficit case is a *non-stringent personal ideal*—an ideal that the agent does not take to rule out an occasional deviation, though it provides him with a second-order reason to avoid deviating: a purely positive reason and therefore optional. In a case of supererogation, similarly, an agent need not be committed to sacrificing himself to others as a general rule in order to be doing so with adequate reason in a given case. So I would also be within my rights to accept the administrator's appeal and serve on an occasional committee to gain some power.

My account of the power-deficit case is somewhat complex—with a different defense of optionality for the different levels of reasons in play—but I think the complexity is needed to capture it as a case of genuine options that does not threaten a regress of appeals to higher-order reasons. My treatment of the case is meant to show that we can have optional reasons all the way up, but without going on forever, even where what is in question at the initial level is the sort of reason that without discounting would yield a requirement. Moreover, my reasons remain optional even if we suppose that my options in the case are commensurable, in contrast to the cases that Raz in his later work takes to involve optional reasons.[12] I also have the option of accepting the administrator's appeal. With a purely positive reason at the upper level, I need no further reason to justify discounting *it*.

There are other cases of optionality where one discounts a first-order reason by setting a threshold of practical attention, rather than priorities—dismissing certain harms or benefits to oneself (such as the cancer risks of 'red dye #7' and similar food ingredients, or small increases in the length or quality of one's life above a reasonable level) as too minor to have to bother with in deliberation, though not so trivial that paying some attention to them would be irrational. A feature of all my cases of discounted negative reasons is that they involve decisions, as sources of further reasons that might be said to be 'enacted' by the agent, rather than

[11] See Raz (1990: 89–90). The addition of Raz's terminology to my own may make my discussion rather cumbersome at times, but the tie to a well-known systematic account of reasons seems to me worth exploiting. Let me stress that my focus on cases of self-sacrifice is not meant to suggest that these exhaust the category of supererogation.

[12] Cf. Raz (1990: 96–105 ff.).

simply existing external to his will, as considerations to which he can or must respond in specified ways. An agent essentially gives himself a reason by setting a threshold, or setting priorities, for practical attention. In the words of a familiar 'self-help' affirmation: I 'give myself permission to say no' to the administrator who appeals to considerations of power in order to get me to serve on committees.

If we build in Raz's account of decision as yielding both higher- and lower-order reasons, a non-arbitrary decision would take us to third-order reasons, but as long as we have purely positive reasons at some level, we should be able to allow for optional first-order reasons without regress.[13] I do not mean to suggest, of course, that an agent explicitly makes appeal even to two levels of reasons as part of ordinary deliberation, but just that he is aware of reasons on several levels as available to justify what he does. But now the question before us is why someone could not just similarly give himself permission to discount moral requirements, or reasons of the sort that would otherwise yield requirements? I approach this question by considering a variant of a familiar figure in metaethics: the amoralist.

PINPOINTING THE REASONS-AMORALIST'S ERROR

The standard figure of the amoralist featured in contemporary discussion is someone who accepts moral judgments and yet, without irrationality, fails to be motivated by them. The possibility of such an agent is called into question by what is now distinguished as 'judgment internalism'—or more precisely, what we might call 'moral judgment internalism': the view that motivational force is internal to the meaning of a moral judgment, so that there could not conceivably be a rational agent who accepted a moral judgment but was not at all motivated to act accordingly. Elsewhere I have defended the possibility of such an agent on the basis of an understanding of the institution of moral language as dependent on a general tie to motivation that allows for exceptions in individual cases; I took myself to be defending a version of externalism, on the usual conception of internalism as a doctrine about the meaning of any particular moral judgment, though I noted that the view also allows for a general version of internalism.[14]

[13] Raz's account would also elude the argument against giving oneself a reason simply by forming a first-order intention in Broome (2001). I discuss further limitations of Broome's arguments in my final section.

[14] See Greenspan (1995: esp. pp. 70–1; cf. pp. 121–2 for early discussion of the distinction between positive and negative reasons, but in application specifically to

'Reasons internalism' is the term now used to distinguish from judgment internalism Williams's view that an agent's reasons can include only considerations capable of being brought to bear on his existing desires or other motivations by rational (meaning rationally unobjectionable) deliberation. I noted above that the critical conception of practical reasons allows for at least some reasons that do not fit this view. But the figure I now want to discuss under the heading of the 'reasons-amoralist' connects more directly to a version of judgment internalism that departs from the usual version, moral judgment internalism, in that the judgments in question are judgments that one has a reason—with the term 'moral' taken as qualifying the reason, rather than the judgments. In the first instance, what fails to motivate the reasons-amoralist is the judgment that he has a moral reason to do something. A denial of the possibility of the reasons-amoralist (on the standard conception of an amoralist as a rational agent) might be spelled out as 'moral reasons-judgment internalism.' Here, too, I would make out my own view as externalist, but as possibly fitting within a broader notion of internalism. I discuss this issue toward the end of the present section, with a suggestion in hand as to where one sort of reasons-amoralist may be going wrong.

The reasons-amoralist, then, is a rational agent who does recognize moral reasons but discounts them as factors affecting his choice of action. He does not think they are defeated by other first-order reasons; rather, he thinks he has adequate higher-order reasons for discounting them. In the kind of case I have in mind, rather than simply making an arbitrary exception of himself, he appeals to a non-stringent version of a Nietzschean ideal of freedom from moral constraints. This might be thought of as a case of 'principled' discounting, discounting by appeal to a further reason (citing the value of achievement or creativity, say), in contrast to 'preferential' discounting, discounting simply at will, of the sort I explained as applying to purely positive reasons. However, the agent's principles do not *require* him to violate morality (as on what I assume would be Nietzsche's own view); they simply permit or entitle him to do so.

In thus arrogating authority to himself to discount moral requirements, the reasons-amoralist is of course doing something morally wrong. He is acting on the basis of an objectionable moral view, and in that sense making a mistake in normative ethics. But on at least some versions of the case, I think we can say more than this, something metaethical, while retaining the assumption of rationality. The point is not to convince him to change his ways, but just to spell out something objectionable in non-moral

motivating states). I see that Blackburn (1998: pp. 61 ff.), simply incorporates this view into internalism; cf. also Gert (2002: esp. 299).

terms about the way he treats his reasons. I take it that an error may be extreme or deep enough to count as a kind of delusion, even though it falls short of irrationality—at any rate in the narrower sense that distinguishes irrationality from mistakes about one's reasons.[15] According to the general account I have given, unresponsiveness to first-order reasons can sometimes be justified by appeal to higher-order reasons, but I want to say that one would have to be in some way deluded—albeit perhaps willfully (and perhaps even strategically) so—to apply this to moral reasons.

We might think of the reasons-amoralist as a kind of moral megalomaniac—extending the term a bit, in the manner of the popular use of 'paranoid,' to cases that are not pathological but involve such inflation of reality (in the present case, self-inflation) that it seems an odd understatement just to call the agents in question mistaken. Instead, we naturally think of such agents as deluded (in cases of self-inflation, grandiose), but again canceling out implications of pathology and taking these terms to apply to something like a pattern of serious distortion.

A megalomaniac in the usual sense has fantasies of unlimited power, among other things; here what is in question is authority. One possibility would be to attribute to the reasons-amoralist some sort of bizarre metaphysical view, on the model of Nagel's charge of practical solipsism, leveled against an agent who does not see others' good as directly providing him with reasons. But the sort of amoralist who fits my account of optional reasons does see others' good as providing him with reasons. The problem is that he thinks he is entitled to discount those reasons, presumably on the basis of features he has that others lack.

Of course, the reasons-amoralist might just be mistaken about his own abilities or prospects of achievement or the like. Even if such delusion involves bias in the gathering and assessment of evidence, and hence a kind of *theoretical* irrationality, it need not be seen as practically irrational, any more than the less extreme evidential biases that apparently result in a somewhat inflated self-image on the part of successful individuals.[16]

[15] See McDowell (1978) and Scanlon (1998: 25 ff.); Scanlon later refers to his 'narrow' sense of irrationality (on which one can count as rational even if deeply confused) as the 'structural' sense; see Scanlon (forthcoming). These authors of course have in mind a mistake consisting in simple failure to recognize some reason, whereas I add a further kind of mistake about the nature of moral reasons in what follows.

[16] Cf. Mele (2001), for a defense of self-deception in terms of bias in gathering and assessing evidence; Mele cites Gilovich (1991: 77), on the tendency toward self-inflation. In popular venues I have also read of studies establishing a correlation between the tendency to overoptimism and high achievement—suggesting that certain kinds of evidential bias can even be *ideally* rational in practical terms, albeit theoretically irrational. Cf. my own account of strategic self-trickery in generating emotions in Greenspan (2000).

More fundamentally, even if the reasons-amoralist is right about the facts, he inflates his appropriate role in relation to moral reasons by failing to appreciate fully their social basis. We can use the critical conception of practical reasons to charge him with a *normative* delusion, about where he stands in relation to the sources of moral reasons, rather than either a metaphysical or a factual delusion about his own or others' existence or nature. For at bottom what he fails to see, or to take in properly, is that he is in no position to waive the criticism supporting a moral reason, understood as a criticism lodged by others on their own behalf.

In the case of discounting that I defended above as an intrapersonal analogue to self-sacrifice for others, the agent chose to sacrifice some interests of her own to aims she preferred to stress. The underlying assumption was that whatever negative reason was in play offered a criticism that was essentially her own. It represented a certain action as in some way problematic or objectionable from her standpoint. So it is appropriate for her also to waive the right to issue it, given that an agent has authority to commit her future self. While it is always possible that she will later change her mind and regret not acting on the reason, in discounting it she commits herself to withholding the relevant criticism.

The contrast is to reasons whose underlying criticism has its source in another agent's standpoint.[17] It does not make any clear sense—rather than just being morally questionable—to claim authority to commit others to withholding criticism. So the reasons-amoralist, while he accepts others' good as providing reasons, and is unconfused about their first-order bearing on his action, shows by his second-order discounting of them that he fails to understand that what ultimately makes them moral reasons—or more specifically, 'core' moral reasons, the sort that ground altruistic requirements. In that sense, he is making a mistake *about* his reasons, since

[17] For a discussion of 'bipolar' reasons see Thompson (2004). I was led to Thompson's article by Wallace (unpublished), which was delivered at the 2005 University of Maryland Conference on Practical Rationality. Wallace uses the notion of bipolar reasons in a contractualist defense of a variant of the optional/compelling reasons distinction applicable to moral rather than rational requirement. His argument equates bipolar reasons with moral reasons, at least in the sense that Scanlon marks off as 'what we owe to others.' Thompson himself thinks of them as a subset of moral reasons, which he identifies as reasons of justice; but perhaps this is meant in the broader sense of classical philosophy. At any rate, I take the reasons in question here to include those commonly referred to as altruistic, following Nagel (1970). In Greenspan (1995), chs. 3 and 6, and Greenspan (1998) I sketch a noncontractualist way of making out socially based ethics, in terms of virtues of social groups, in effect appealing to an interpersonal standpoint that an individual agent would not be in a position to discount. Perhaps a consequentialist might use the distinction between 'agent-relative' and 'agent-neutral,' as formulated for reasons and value in Nagel (1986: 164–75), to limit discounting to reasons based on criticism relative to the standpoint of the agent.

he fails to appreciate fully what moral reasons amount to. Besides reflecting how his acts affect others' welfare, as he recognizes, moral reasons accord a certain authority to others over what counts as acceptable influence. A justification has to be addressed to them.

On this account, the reasons-amoralist would seem to be making a *metaethical* error—in a sense of 'metaethical' somewhat broader than the traditional sense, in which metaethics was limited to questions about the meanings of moral terms, concepts, or judgments. I think most people working in metaethics now construe the subject more broadly, to include metaphysical, epistemological, logical, and moral-psychological as well as semantical questions about ethics, though the narrower conception is still widely assumed by philosophers working in other areas (who often are averse to metaethics on grounds that depend on it). I take it that, in discounting moral reasons, the reasons-amoralist is not deluded about the semantics of moral reason-judgments, but rather about their practical implications, since he misunderstands the sources of moral reasons. Besides recognizing the reasons in question in objective terms—recognizing the facts that constitute them—he recognizes them as moral in the sense that they concern others' welfare, which I think is the common view.

This sort of rough-and-ready characterization of morality is enough for us to allow that the reasons-amoralist uses the term 'moral reasons' with the same *meaning* as most of us, though he exhibits a deficient understanding of the term, in a sense of 'understanding' that includes more than meaning. He does acknowledge its social reference, but he thinks that can be adequately handled without going beyond his own deliberative standpoint. He is not just using the term in an 'inverted commas' sense, if that means attributing it to common usage or a figure of speech, without endorsement.[18] He has the usual concept of a moral reason, we might say, but he exhibits a deficient conception of moral reasons (or more precisely, of core moral reasons in ought-based ethics, of the sort in question here), when he fails to acknowledge their basis in criticism from standpoints other than his own. Similarly, I have the concept of a quark and *mean* the same thing as a scientist does when I use the term, though my conception of a quark no doubt omits much that a scientist would say is essential to understanding the nature of quarks and possibly contains some errors. I can use 'quark' meaningfully without really knowing what quarks are.

Allowing for the reasons-amoralist as a rational (though metaethically deluded) agent who recognizes moral reasons but is not motivated by them involves rejecting internalism, understood as a view about motivation and

[18] Cf. Hare (1965: 189–90). Let me thank Michael Smith for raising this issue in discussion.

meaning—a view that takes motivation as 'internal to' the meaning of a moral judgment (which I take to include a moral reasons-judgment), so that anyone who sincerely makes the judgment must be motivated by it. However, one might suggest that internalism should now be understood more broadly, in line with the broader notion of metaethics: we should interpret 'meaning' to include a full understanding of the moral reason-judgments in question, which would require acknowledging the social basis of (core) morality, as the reasons-amoralist fails to do.

In order to rule out all varieties of amoralism, though, internalists in this extended sense would seem to have to incorporate into meaning the answers to many disputed metaethical questions. There is disagreement, for instance, about whether moral reasons can ever be overridden by non-moral considerations; presumably those who deny this would also deny that non-moral reasons can ever be higher-order than moral reasons. If they are right, and we accept internalism in the extended sense, agents who claim to be acting appropriately in acting against a moral reason would be dismissed as not really meaning *moral*—along with any metaethical theorists who take the opposing view. The result would tend to trivialize or undermine metaethical debate—it reminds me a bit of redefining 'God' in such broad terms that no one can call himself a non-believer—so I resist broadening the notion of internalism and continue to call myself an externalist.

My remarks here have focused on one sort of reasons-amoralist (which is all we need to defend the possibility of such an agent)—not just the sort who happens to be indifferent to moral reasons, but rather someone who discounts them in a principled fashion, by appeal to a higher-order reason, but a reason appealing to a merely personal ideal, by analogy to my power-deficit case in the preceding section. I think the common picture of a Nietzschean 'free spirit,' though, would be of someone who assigns his ideal an impersonal value and hence sees his pursuit of it as indeed answering criticisms from other standpoints. Perhaps he thinks that the value of achievement or creativity should be recognized by all agents as outweighing any independent forms of disvalue, such as harms his promotion of those ideals might inflict on agents incapable of pursuing them as well as he. This kind of case would seem to involve an objectionable *normative* assessment of moral reasons, rather than a metaethical misconception of them. Where the agent in question is not deluded about the facts, about his own prospects of achievement or creativity, all we could charge him with would be moral error. But optional reasons are not in question in this normative ethical version of the case. Here the agent appeals only to higher-order negative reasons, reasons that would rule out assigning greater weight to first-order reasons against inflicting certain harms.

The point of lodging the charge of metaethical error against my own version of the reasons-amoralist was to keep my defense of optional reasons from undermining binding moral 'ought.' The only relevant version of the reasons-amoralist I can think of whose error might seem to lie within normative ethics would be an agent who subscribes to one of the views invented to contrast with ethical egoism in introductory ethics texts, 'first-person egoism.' A first-person egoist thinks that everyone morally ought to promote *his* (the egoist's) good. But while this is a normative *view*, it might still be based on metaethical error, such as an error about the point and purposes of morality and about what sort of conception of practical reasons could support it. For as thus described, a first-person egoist is someone who accepts a certain moral view, not just an agent who characteristically acts in accordance with it. The view implies that his reasons always outweigh those (if he recognizes any) based on criticism from other standpoints—for no particular further reason beyond the fact that his reasons rest on criticisms from *his* standpoint. At the very least, this is out of line with the function of morality (or of core morality on an ought-based account) as yielding a viable code of social behavior, one that a group could be motivated to abide by. One might also question whether its underlying conception of practical reasons can be made coherent.

Let me acknowledge that there are cases of discounting moral reasons that involve *no* error—namely, cases of imperfect duties.[19] A duty to give aid to those in need presumably rests on criticisms from each of the standpoints of needy individuals, rather than just from some general standpoint, but a moral agent does have authority to discount some indefinite set of them. We can get this result within the framework outlined above by taking a decision to give aid to some needy individuals as the source of an exclusionary permission, a permission to discount reasons based on the criticisms of others appealing for aid.

As I have set it up, the reasons-amoralist's error is at bottom a theoretical error, about the nature of moral reasons. It results in faulty practical reasoning, but possibly in the service of the agent's ends, on the model of cases of promoting success by inflating one's own abilities or achievements. So I would not call it practically irrational. Moreover, it occurs at such a sophisticated level that I think the reasons-amoralist is clearly no fool—except perhaps in Hobbes's sense, of the fool who has 'said in his heart' there is no justice. We can think of him as deluded, though, insofar as his error involves a grandiose sense of himself as authorized to speak for others. Instead of simply inflating his abilities in the manner of a common-variety

[19] I owe thanks to Stephen Darwall for bringing up this issue.

megalomaniac, the reasons-amoralist inflates his role in relation to core moral reasons, those based on criticism from other standpoints.

REASONS AND WIDE-SCOPE OUGHTS

My preceding argument was part of an attempt to defend the view that moral 'ought' can be understood in terms of reasons, even though reasons as such may be optional. I now want to respond to a different sort of challenge that might be suggested by John Broome's recent defense of a distinction between reasons and undetachable 'wide-scope' oughts, of the sort Broome calls 'normative requirements.'[20] I think I can do so relatively briefly by referring to some earlier work of my own on conditional oughts.[21]

The oughts in question have the form *O*(if *p*, then *q*) and include, most notably, the Kantian hypothetical imperative, which requires that, if you will an end, you also will the means to it. As Broome points out, they also cover rules of theoretical reasoning, such as the requirement that, if you believe the premises of a valid argument, you also believe the conclusion. It would be natural to take wide-scope oughts to cover moral rules as well, such as the rule requiring that, if you make a promise, you keep it. Surely these count as normative requirements.

Broome tells us that wide-scope oughts do not admit of detachment; that is, we cannot apply modus ponens to *O*(if *p*, then *q*) to derive *Oq* if we simply grant that *p* is true, since making *p* false represents an alternative way for the agent to satisfy the requirement, even if it is an option he in fact turns down. In the case of the hypothetical imperative, he does not necessarily have to take the means to what in fact is his end; he has the option of repudiating the end instead. But Broome distinguishes between normative requirements and reasons in that reasons are pro tanto and need to be weighed against competitors. So a rational agent can act against a reason by appeal to countervailing considerations. By contrast, if an agent neither repudiates his end nor takes the means to it, that is enough to

[20] See Broome (2004). Broome interprets 'ought' in a relatively weak sense, common in ordinary language, as conveying recommendation rather than requirement—in contrast to the usual interpretation of *moral* 'ought' as having the force of a command. I of course assume the stronger interpretation here, but I should think it also fits the wide-scope oughts concerning logical rules that Broome has in mind.

[21] See Greenspan (1975). For discussion of surrounding issues about 'ought,' 'obligation,' and deontic logic (but on the basis of a picture of oughts as essentially action-guiding); cf. Greenspan (1972). From email correspondence I gather that Broome would accept at least some of the limitations on his argument that I argue for below.

warrant a judgment of irrationality.[22] While the fact that something is a means to our ends might give us *reason* to will it, then, we could not derive a narrow-scope requirement to that effect.

We might be tempted to conclude from this argument that no set of reasons, however structured and qualified, could possibly add up to a moral ought. But if we look more closely at the application of wide-scope oughts to moral requirements like promise-keeping, I think we will want to qualify both Broome's claims about detachment and his distinction between reasons and normative requirements. An ought conditional on something, such as the making of a promise, that is already settled by the time assigned to the act it conditionally requires does admit of detachment. Once the antecedent is no longer capable of being falsified—the promise already has been made—one of the agent's two options for satisfying the conditional will be closed off. The only way of satisfying it, then, will be to do the act in question—keep the promise.

Of course, a full representation of the conditional ought relevant to promise-keeping will be more complex than this, with further conditions specifying that one has not been released from the promise, among other things. But the agent cannot simply falsify such further conditions at will, in the way he normally can repudiate an end (or belief in the premisses of an argument). So here we have what my argument in earlier work referred to as a 'time-bound' ought, as distinct from the timeless instances of logical rules that are the basis for Broome's argument. In Broome's cases of normative requirement—for which '*logical* requirement' might be a better term (with the hypothetical imperative seen as a logical requirement of practical reasoning)—it is my *current* ends and beliefs, not what I wanted or believed at some earlier time, that are assumed to dictate what I should do or believe now.

To return to the case of promise-keeping: this also seems to involve a wide-scope ought that is subject to comparison of strengths with conflicting oughts that might defeat it, on the model Broome apparently restricts to reasons. Since the term 'pro tanto' applies more naturally to reasons, we might revert to Ross's terminology for duties and refer to these as 'prima facie' oughts. For instance, in Plato's case (*Republic* 331c5–9) of the agent who has to decide whether to return borrowed weapons to someone who has gone berserk, O(if he promises to return weapons, then he returns them) would seem to be defeated by a competing prima facie ought: $O\sim$(he gives

[22] There might be cases of 'rational irrationality,' though, where that instance of irrationality is in his long-term interests—perhaps because someone has offered him a large reward for violating the hypothetical imperative. Broome (2004: 43–5) discusses a parallel point for belief made by Andrew Reisner.

weapons to a lunatic). For that matter, it makes perfect sense to consider reasons for and against having or accepting rules like promise-keeping—for or against taking them *as* normative requirements—in contrast to logical requirements, for which Broome has a point in dismissing the notion of weighing reasons.

There are other distinctions to be drawn between reasons and oughts, some of them relevant to the critical conception of reasons, as defended in this paper. First, note that my argument above for narrowing 'ought' in time-bound cases to options still open to the agent applies only to present-tense statements of 'ought-to-do'—as opposed to statements about what one ought to have done, what ought to be the case, or the like. It depends on taking time-bound 'ought' as indeed essentially action-guiding, in contrast to reasons generally, on the critical conception. There is a corresponding subset of reasons, of course—reasons *to do* something—but according to the critical conception the class of reasons bearing on a given act extends wider. There is no time-limit on assessing an act in light of criticism.

Further, a reason seems to be detachable even in cases where the corresponding narrow-scope ought is not, since the condition on the normative requirement in question is not yet satisfied.[23] My pursuit of a certain end, most notably, does give me a *reason* to take the necessary means to it, even though I also have the option of repudiating the end instead. If I did repudiate it, I would no longer have the reason, but the mere fact that I might repudiate it leaves the reason in force—as a pro tanto consideration against failing to take the means. Still, framing the hypothetical imperative as an action-guiding wide-scope 'ought' is useful in defense of the critical conception of reasons, since it explains how a reason to achieve one's ends can be seen as a response to criticism.[24] The relevant criticism is directed toward a conjunction—pursuing the end without taking the means—from which we can detach a reason, though not a requirement.

I should also note that there is also another form of detachment applicable to wide-scope oughts, besides the one modeled on modus ponens, that I call deontic (as opposed to factual) detachment.[25] The basic idea here is that an ought-statement *requiring* (rather than asserting) the truth of the antecedent of a conditional or wide-scope ought would also allow us to detach an ought-statement of the consequent. From an instance of the hypothetical imperative, *O*(if I want to provide for my old age, I save some money), we need only grant that I *ought* to want to provide for my old

[23] See the argument for this point in Raz (2005: 12–13).

[24] Let me thank Gunnar Björnsson for pressing this point in comments on Greenspan (unpublished-b).

[25] See Greenspan (1975: 260).

age, whether or not I actually want to, in order to detach a requirement that I save some money. Broome (in personal correspondence) poses the following as a counterexample to deontic detachment: *O*(if you go running, you wear your running shoes) and *O*(you go running) would let us detach *O*(you wear your running shoes)—which is implausible if you are not going running and have no intention of doing so. But I think this can be handled if we recognize oughts as time-bound in the way I sketched above.

To make contact with my deontic detachment principle, Broome's ought-statements here need to be restated a bit, so that it is your *intention* of going running (or your wanting to, having that end, or the like) that requires putting on the shoes. Once we do this, however, I think that the claim that you ought to wear your running shoes will not seem so implausible as he suggests. Really, we would say, what you ought to do is decide to go running and put on the shoes before you go.[26] While it is still possible for you to put on the shoes and go running in them, we can detach a requirement that you put them on, as your first step toward acting appropriately on the intention you ought to have to go running in the shoes. If we know that in fact you are not going to form the intention to go running, we might also say that you should not put on the shoes, but I would take this as short for a conditional ought—*O*(if you do not want to go running, you do not wear your running shoes)—that does not allow detachment, since the antecedent is neither required nor settled (even if true) *at times when you can still satisfy the conditional*. So we can apply deontic detachment to Broome's case, as long as we recognize appropriate limits on factual detachment.

In application to moral oughts, then, Broome's argument from normative requirements as wide-scope oughts shows only that we could not get a moral requirement from a wide-scope ought conditional on ends that the agent can *and may* still repudiate at the time assigned to action. Besides past acts, such as promises, any necessary features of human nature or of agency or the like, including ends, would let us detach narrow-scope or unconditional oughts by factual detachment. A Kantian approach to ethics might be seen as working from an a priori version of this model of factual necessity. However, a deontic detachment model, making out the basis for morality as normative through-and-through, would provide

[26] I defend a claim of this sort in Greenspan (1978); note that the trees on pp. 78–9 are reversed. It sounds odd to apply 'ought' directly to forming an intention, so I apply it here to making a decision—and we also apply it to adopting an end, making a plan, and the like—as a way of generating an intention, though typically at some distance in time from what it is an intention to do. Note that the intention in question in the running case is to not to run immediately, but to do something else first (put on the shoes).

an alternative interpretation of Kant's talk of morality as holding for rational agents as such, on the assumption that certain ends are required by rationality.

There are general ends such as interpersonal coordination that would seem to be required to facilitate fulfillment of whatever other, more specific ends an agent should happen to adopt.[27] For that matter, the Strawsonian ideal of mutual recognition in a community of persons, as suggested by Kant's 'kingdom of ends,' is invoked in recent work by Scanlon, among others, as something valuable in itself.[28] Selected agents like the reasons-amoralist may be able to do well for themselves without adhering to some such ideal, but relying on this ability is risky, at best. However, if oughts are to be understood in terms of practical reasons, and practical reasons are interpreted in accordance with the critical conception, this or some similar basis for ethics could not be described solely in the language of positive value, as talk of ideals might suggest, but would also have to refer to something negative: respect for persons *as* sources of *criticism.*

What displaced attention to moral 'ought' in recent years was the move on the part of a number of philosophers back to virtue ethics, with its preference for the language of positive value.[29] There was Anscombe's well-known dismissal of 'ought' as empty without belief in a divine lawgiver.[30] Perhaps relatedly, some philosophers thought of notions of moral duty or obligation as motivating only by way of some sort of extrinsic threat—of divine or legal punishment, social censure or emotional guilt—that compromised the value of the moral motive. However, by interpreting 'ought' in terms of practical reasons, understood as referring to criticism from other persons' standpoints, we can both bring the notion down to earth and connect it to a sanction that being morally motivated just *means* wanting to avoid.

[27] Cf. Bratman (2001: 207), for a defense of cross-temporal consistency and other elements of planning agency as a 'universal means' (though not particularly in reference to ethics). Something similar would seem to fit ideals of identity or integrity, of the sort proposed as a Kantian basis for morality in Korsgaard (1996: 101 ff.). But while all agents necessarily have some ends or other, it is not clear that all accept ideals of identity or integrity. To get by Broome's arguments and allow for detachment, then, we would apparently need to treat such ideals as ends everyone *ought* to have—perhaps rationally, but not just as a consequence of the hypothetical imperative plus agents' actual ends.

[28] See Scanlon (1998: esp. p. 163); cf. Strawson (1959: ch. 3) and (1962).

[29] But cf. Thomson (this volume), for what seems to amount to a negative version of virtue ethics—Thomson calls it 'vice ethics'—that is set up to generate oughts. In discussion at the Wisconsin Metaethics Workshop, however, it turned out that a single ought-violation would be enough to make one's character defective in Thomson's intended sense; so I think the approach might instead be seen as a version of duty ethics that hinges in a serious way on virtue-ethical notions.

[30] See Anscombe (1981: 26–42, p. 37).

REFERENCES

Anscombe, G. E. M. (1981) *Collected Philosophical Papers*, Vol. 3, *Ethics, Religion and Politics* (Minneapolis: University of Minnesota).

Blackburn, Simon (1998) *Ruling Passions: A Theory of Practical Reasoning* (Oxford: Clarendon Press).

Bratman, Michael (2001) 'Taking Plans Seriously,' in E. Millgram (ed.), *Varieties of Practical Reasoning* (Cambridge, Mass.: MIT Press), 203–21.

Broome, John (2001) 'Are Intentions Reasons? And How Should We Cope with Incommensurable Values?', in C. W. Morris and A. Ripstein (eds.), *Practical Rationality and Preference: Essays for David Gauthier* (Cambridge: Cambridge University Press), 98–120.

____ (2004) 'Reasons,' in R. J. Wallace, P. Pettit, S. Scheffler, and M. Smith (eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (Oxford: Oxford University Press), 28–55.

Dancy, Jonathan (2004) 'Enticing Reasons,' in R. J. Wallace, P. Pettit, S. Scheffler, and M. Smith (eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (Oxford: Oxford University Press), 91–118.

Gert, Joshua (2002) 'Expressivism and Language Learning,' *Ethics*, 112: 292–314.

____ (2004) *Brute Rationality*: *Normativity and Human Action* (Cambridge: Cambridge University Press).

Gilovich, Thomas (1991) *How We Know What Isn't So* (New York: Macmillan).

Greenspan, Patricia S. (1972) *Derived Obligation: Some Paradoxes Escaped* (unpublished Ph.D. dissertation: Harvard University).

____ (1975) 'Conditional Oughts and Hypothetical Imperatives,' *Journal of Philosophy*, 72: 259–76.

____ (1978) 'Oughts and Determinism: A Response to Goldman,' *Philosophical Review*, 87: 77–83.

____ (1995) *Practical Guilt: Moral Dilemmas, Emotions, and Social Norms* (New York: Oxford University Press).

____ (1998) 'Moral Responses and Moral Theory: Socially-Based Externalist Ethics,' *Journal of Ethics*, 2: 103–22.

____ (2000) 'Emotional Strategies and Rationality,' *Ethics*, 110: 469–87.

____ (2005) 'Asymmetrical Practical Reasons,' in M. E. Reicher and J. C. Marek (eds.), *Experience and Analysis: Proceedings of the 27th International Wittgenstein Symposium* (Vienna: oebv&hpt), 115–22.

____ (unpublished-a) 'Adequate Reason.'

____ (unpublished-b) 'Reconceiving Practical Reasons.'

____ (unpublished-c) 'Sensible Satisficing.'

Hare, R. M. (1965) *Freedom and Reason* (New York: Oxford University Press).

Korsgaard, Christine M. (1986) 'Skepticism about Practical Reason,' *Journal of Philosophy*, 83.

____ (1996) *The Sources of Normativity* (Cambridge: Cambridge University Press).

McDowell, John (1978) 'Are Moral Reasons Hypothetical Imperatives?' *Proceedings of the Aristotelian Society*, 52: 13–29.

Mele, Alfred R. (2001) *Self-Deception Unmasked* (Princeton, NJ: Princeton University Press).

Melville, Herman (1853) 'Bartleby the Scrivener: A Story of Wall Street,' in *Putnam's Monthly. A Magazine of Literature, Science, and Art*, Vols. 1–2 (New York: G. P. Putnam & Co.).

Nagel, Thomas (1970) *The Possibility of Altruism* (Oxford: Clarendon Press).

——(1986) *The View from Nowhere* (New York: Oxford University Press).

Raz, Joseph (1990) *Practical Reason and Norms* (Princeton, NJ: Princeton University Press).

——(1999) *Engaging Reason: On the Theory of Value and Action* (Oxford: Oxford University Press).

——(2005) 'The Myth of Instrumental Rationality,' *Journal of Ethics and Social Philosophy*, 1: 2–28.

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

——(forthcoming) 'Structural Irrationality,' in G. Brennan, R. Goodin, and M. Smith (eds.), *Common Minds* (Oxford: Oxford University Press).

Strawson, P. F. (1959) *Individuals: An Essay in Descriptive Metaphysics* (London: Methuen).

——(1962), 'Freedom and Resentment,' *Proceedings of the British Academy*, 48: 1–25.

Thompson, Michael (2004) 'What is it to Wrong Someone? A Puzzle about Justice,' in R. J. Wallace, P. Pettit, S. Scheffler, and M. Smith (eds.), *Reason and Value: Themes from the Moral Philosophy of Joseph Raz* (Oxford: Oxford University Press), 333–84.

Wallace, R. Jay (unpublished) 'The Deontic Structure of Morality.'

Williams, Bernard (1981) 'Internal and External Reasons,' in *Moral Luck: Philosophical Papers 1973–1980* (Cambridge: Cambridge University Press), 101–13.

——(1995) 'Internal Reasons and the Obscurity of Blame,' in *Making Sense of Humanity and Other Philosophical Papers, 1982–1993* (Cambridge: Cambridge University Press), 35–45.

9

The Humean Theory of Reasons

Mark Schroeder

This paper offers a simple and novel motivation for the Humean Theory of Reasons. According to the Humean Theory of Reasons, all reasons must be explained by some psychological state of the agent for whom they are reasons, such as a desire. This view is commonly thought[1] to be motivated by a substantive theory about the power of reasons to motivate known as *reason internalism*, and a substantive theory about the possibility of being motivated without a desire known as the *Humean Theory of Motivation*. Such a motivation would place substantial constraints on what form the Humean Theory of Reasons might take, and incur substantial commitments in metaethics and moral psychology. The argument offered here, on the other hand, is based entirely on relatively uncontroversial methodological considerations of perfectly broad applicability, and on the commonplace observation that while some reasons are reasons for anyone, others are reasons for only some. The argument is a highly defeasible one, but is supposed to give us a direct insight into what is philosophically deep about the puzzles raised for ethical theory by the Humean Theory of Reasons. I claim that it should renew our interest in the relationship between these two kinds of reason, and in particular in the explanation of reasons which seem to depend on desires or other psychological states.

1.1 THE HUMEAN THEORY OF REASONS: WHAT

Consider a case like that of Ronnie and Bradley. Ronnie likes to dance, but Bradley can't stand even being around dancing. So the fact that there will be dancing at the party tonight is a reason for Ronnie to go there, but not for Bradley to go there—it is a reason for Bradley to stay away. Ronnie

[1] See, for example, Williams (1981), Bond (1983), Darwall (1983), Korsgaard (1986), Hooker (1987), Hubin (1999), and others.

and Bradley's reasons therefore differ—something is a reason for one to do something, but not for the other to do it. And this difference between their reasons seems obviously to have something to do with their psychologies. It may not be ultimately explained by the difference in what they *like*, of course—the explanation may ultimately derive from a difference in what they *value*, or what they *care* about, what they *desire*, *desire to desire*, what they take or would take *pleasure* in, or what they *believe to be of value*. I'm not claiming that it is uncontroversial that one rather than another of these kinds of psychological states is what really explains the difference between Ronnie and Bradley—after all, many of these psychological characteristics often go hand in hand, and even moderately sophisticated views can make them hard to distinguish simply by considering cases. All I'm claiming is that it should be pretty close to uncontroversial that there are at least some reasons like Ronnie's, in that they are explained by *some* psychological feature.[2]

The *Broad* Humean Theory of Reasons says that all reasons are explained in the same way as Ronnie's—by the same kind of psychological feature:

Broad Humean Theory Every reason is explained[3] by the kind of psychological feature that explains Ronnie's reason in the same way as Ronnie's is.

The Broad Humean Theory of Reasons is really too broad to sound familiar to most readers familiar with the philosophical literature on reasons. That literature is full of references to, and attacks on, a familiar view that is

[2] Allow me to head off a possible distraction. There is a sense in which what reasons one has depends on what one *believes*. In this sense, though there will be dancing at the party and Ronnie and Freddie both like to dance, if Freddie is aware of this but Ronnie is not, then we might say that Freddie has this reason but Ronnie does not. This is the *subjective* sense of 'reason'. When I say that it is uncontroversial that at least some reasons depend on psychological states, this is not what I intend. What I mean, is that it is uncontroversial that at least some reasons *in the objective sense* depend on psychological states.

[3] A qualifying note about how to understand this talk about *explanation*. The fact that there will be dancing at the party tonight is a reason for Ronnie to go there, in part *because* Ronnie likes to dance. That must be part of *why* it is a reason for Ronnie to go there, because it is not a reason for Bradley to go there, and liking to dance is precisely what distinguishes Ronnie from Bradley. The Humean Theory of Reasons is a generalization of *this* claim. It is the claim that whenever R is a reason for X to do A, that is in part *because* of something about X's psychology—that this is part of *why* R is a reason for X to do A. I'm using the term 'explained by' to cover these kinds of claims about what is so *because* something else is so, and what is part of *why* it is so. This is not intended to import epistemic or pragmatic ideas about what *agents* might be doing when they engage in the behavior of *explaining* things to one another. In my sense, X explains Y iff Y is the case *because* X is the case, or X is part of *why* Y is the case. The explanation is the *content* of the answer to a 'why?' question—not the answer itself, nor the process of giving it.

more narrow than the Broad Humean Theory. This view is a *version* of the Broad Humean Theory because it agrees that all reasons must be explained by the same kind of psychological feature as explains Ronnie's. But it is more specific than the Broad Theory, because it takes a view about what kind of psychological state does explain the difference between Ronnie's and Bradley's reasons. It says that it is a *desire*, in the traditional philosophical sense:

Narrow Humean Theory Every reason is explained by a desire in the same way as Ronnie's is.

Even the Narrow Humean Theory of Reasons, of course, is only loosely called 'Humean'; there is an excellent case to be made that Hume himself was not a Humean in either sense. Both theories are associated with Hume's name primarily because their proponents have typically been loosely inspired by Hume.[4]

So allow me to reveal my hand. I believe that a version of the Narrow Humean Theory of Reasons is true, and I have defended such a theory elsewhere.[5] But in this paper I will not be arguing for the Narrow Humean Theory. The argument of this paper is only a motivation for the Broad Humean Theory. It is my *view* that there are good arguments from the Broad Humean Theory to the Narrow Humean Theory, but I will not advance those arguments in this paper. Indeed, I think that for most of the philosophical reasons for which philosophers have been interested in whether the Humean Theory of Reasons is true, whether the Humean Theory is Narrow or not is beside the point. In the next subsection I will explain why.

1.2 THE HUMEAN THEORY OF REASONS AND MORAL SKEPTICISM

The Broad Humean Theory of Reasons takes no stand on what kind of psychological state it is that explains the difference between Ronnie and Bradley. It only claims that whatever it is, it is also needed to explain every other reason. But this does not water the Humean Theory down so much as to make it of little interest. On the contrary, it is exactly the right specificity of view that we should be worried about, for exactly the reasons

[4] So it's not worth quoting Hume for the purpose of refuting either view. Compare Korsgaard (1997). See also Setiya (2004) for an excellent discussion of how to understand Hume's commitments about practical reason.

[5] Schroeder (2004), (forthcoming-b), (forthcoming-c).

that philosophers have been worried about the Narrow Humean Theory of Reasons all along.

The principal philosophical interest of the Narrow Humean Theory of Reasons, after all, is that it is supposed to play a special role in motivating certain kinds of skepticism about the universality or objectivity of morality. The problem is that according to the Humean Theory, every reason must be explained by a desire of the person for whom it is a reason. But it is hard to see how such an explanation could possibly work for all moral reasons. Consider this case: Katie needs help. So there is a reason to help Katie. It is a reason for you to help Katie, a reason for me to help Katie, and in general, it is a reason for *anyone* to help Katie. Some of the most important moral reasons seem to be like the reason to help Katie—they are reasons for *anyone*, no matter what she is like. But does *everyone* really have some desire that would explain a reason for her to help Katie in the same way that Ronnie's desire to dance explains his reason to go to the party? It seems fairly implausible.

So those who accept versions of the Narrow Humean Theory often take revisionist views about the kind of objectivity that moral claims have. Gilbert Harman, for example, argues for these reasons that moral claims aren't really universally binding, but are only binding on people who have implicitly contracted in certain ways. This is his brand of moral relativism in 'Moral Relativism Defended' and subsequently.[6] Philippa Foot argues for almost identical reasons that moral claims don't provide reasons to everyone, but only to those who care about morality. That is her thesis in 'Morality as a System of Hypothetical Imperatives'.[7] The difference between Harman and Foot is that Foot thinks that there is another, non-reason-giving, sense in which moral claims nevertheless 'apply' to everyone, even to those to whom they don't give reasons. John Mackie argues that it is essential to moral claims that moral requirements give reasons to everyone. Since this is incompatible with the Humean Theory of Reasons, he concludes that moral claims are uniformly false.[8] These are all drastic forms of skepticism about the objectivity or universality of morality that are motivated by the Humean Theory of Reasons. And it is these kinds of arguments which give the Humean Theory so much of its interest for moral theorists. It is in order

[6] Harman (1975). See also Harman (1978) and (1985).

[7] Foot (1975). Foot, however, subsequently rejected this view. See, for example, Foot (2001).

[8] Mackie (1977). The interpretation of Mackie's argument from 'queerness' is controversial, however, since there are at least two other good candidates for the kind of argument that Mackie intended to offer. Richard Joyce, however, does unambiguously endorse this argument as the best argument for a moral error-theory, in the process of motivating his moral fictionalism. See Joyce (2001).

to avoid these kinds of implications that moral philosophers have been so concerned, over so many years, to finally conclusively refute the Humean Theory.

But notice that none of these arguments actually turns on making any particular assumptions about what *kind* of psychological state is necessary in order to explain a reason. No matter what kind of psychological state is necessary in order to explain a reason, it is fairly implausible that we are going to be able to expect that everyone, no matter what she is like, will have some psychological state of the requisite kind in order to explain a reason that is supposed to be a reason for everyone. So the Broad Humean Theory of Reasons best captures what lies at the heart of this kind of worry about the universality or objectivity of morality—the kind of worry that the revisionary Humean takes to be conclusive.

Now if the Narrow Humean Theory of Reasons is the most popular version of the Broad Humean Theory, it is easy to understand for purely sociological reasons why it would receive so much attention. But what we can expect for sociological reasons is quite different from what we should demand of good philosophy. There are any number of supposed refutations of the Narrow Humean Theory of Reasons in the literature, all for the purpose of setting aside the kinds of skeptical arguments run by Harman, Foot, and Mackie. But it's simply faulty reasoning to think that if an argument you want to rebut needs the premiss that *p*, you can rebut it by refuting *p*+, a stronger premiss. If we're really concerned about the kinds of skeptical arguments raised by Harman, Foot, and Mackie, we have to be concerned about the more general Broad Humean Theory of Reasons.

1.3 THE CLASSICAL ARGUMENT FOR THE HUMEAN THEORY

So why haven't philosophers critical of the skeptical arguments of Harman, Foot, and Mackie been more concerned about this more general view? Are they philosophically lazy? No; a much better explanation is easy to find. The better explanation is that it is widely believed to be common knowledge what the *only motivation* for believing the Broad Humean Theory of Reasons is.[9] And it is an argument which, if it works, also establishes the truth of

[9] Hubin (1999: 31): 'I think what is special about the Humean position on reasons for acting is approximately what most defenders and detractors alike are prone to point to as its attraction … What attracts many of us, to the different degrees that we are attracted, to Humeanism is, as many have suggested, a motivational argument.'

the Narrow Humean Theory of Reasons. I call it the *Classical Argument* for the Humean Theory.

Elijah Millgram, a critic of the Humean Theory, puts the Classical Argument most succinctly: 'How could anything be a reason for action if it could not motivate you to actually *do* something? And what could motivate you to do something, except one of your desires?'[10] Millgram's first rhetorical question states the thesis of *reason internalism* and his second that of the *Humean Theory of Motivation*. If having a reason requires being motivatable, and being motivatable requires having a desire, then having a reason must require having a desire. And that is enough of the Humean Theory of Reasons to motivate the kinds of skepticism just discussed.

A great deal of the abundant literature critical of the Humean Theory of Reasons has focused on rebutting the Classical Argument, and many of the points made there are fairly conclusive. The Classical Argument leaves much to be desired, as a motivation for the Humean Theory of Reasons. But if this is the only motivation for the Broad Humean Theory, then we can straightaway draw two conclusions about the kind of view that the Humean Theory takes about desires. First, they have to be motivating states. And second, they have to be *ubiquitous* motivating states: any action whatsoever has to have one of them in its causal etiology.

These two conclusions set enormous constraints on the kind of shape that the Broad Humean Theory of Reasons might take. If they are sound, then refutations of the Broad Humean Theory of Reasons can take for granted some fairly strong conclusions about what kind of psychological state explains reasons, according to the Humean: not only that they are *desires*, but what desires, in fact, *are*. But I think that if we are genuinely interested in the kind of view that can motivate Harman's, Foot's, and Mackie's kinds of skepticism about the objectivity of morality, then we should cast our nets wider. In particular, I don't think that the Classical Argument gives the best or most interesting argument for the Broad Humean Theory of Reasons. It is the purpose of this paper to offer a better and more general motivation for the Humean Theory, one which doesn't commit that theory to any particular story about what explains the difference between Ronnie and Bradley. It is my purpose to show how *few* assumptions about the Humean Theory of Reasons are necessary in order to motivate it.

[10] Millgram (1997: 3). The classical argument is given in Williams (1981), cited in Bond (1983) and Darwall (1983), and discussed extensively in Korsgaard (1986), Hooker (1987), Millgram (1996), and in many other places. Of these authors, Darwall is the only one who allows that there are other motivations for the Humean Theory of Reasons.

2.1 THE POSITIVE MOTIVATION

It is fairly uncontroversial, as I suggested in section 1.1, that the difference between Ronnie's and Bradley's reasons is due to a difference in their psychologies. It is not uncontroversial, of course, *which* difference in their psychologies it is due to. But the central idea behind my motivation for the Humean Theory is to take what we *do* know about Ronnie and Bradley's case, and to put it to work. If there is *any* uniform explanation of all reasons, then maybe what we know about how *some* explanations of reasons work will help to shed light on how *all* explanations of reasons must work. And that is the idea that I will be pushing. There are broad-based theoretical motivations to hope that there might be some common explanation of why there are the reasons that there are—broad motivations to be in search of a uniform explanation of all reasons. If we are after a uniform explanation of all reasons, I will be suggesting, Ronnie and Bradley's case is where we should look.

This may not move you. You may be thinking, 'but maybe there are *two kinds* of reason—one kind that gets explained by psychological states, and one kind that doesn't!' I agree. There *may* be two kinds of reason. But on the face of it, the reason for Ronnie to go to the party and the reason for Ronnie not to murder are both *reasons*—they are both cases of the same general kind of thing. It would be very surprising if these two uses of the word 'reason' turned out to be merely homonyms. So, given that they are both cases of the same kind of thing, it is reasonable to wonder whether there is anything to be said about why they are. And it is this reasonable thing to wonder, I will be suggesting, which will lead to the hypothesis that all reasons are explained in the way that Ronnie's is.

Of course, it doesn't follow from the fact that Ronnie's reason is explained, in part, by his psychology, and the hypothesis that there is a common explanation of all reasons, that psychological features figure in all of these explanations. It could be that the feature of Ronnie's psychology plays a *role* in the explanation of his reason that can be filled by other kinds of thing—for example, by promises or special relationships. And in any case, if we really care about finding a common explanation of all reasons, something must motivate us to pay attention to Ronnie and Bradley's case, in particular. After all, there are many cases of reasons, and we might know something about how many of them work. Where does the pressure come from to try to generalize Ronnie and Bradley's case to cover others, rather than trying to generalize other cases to cover Ronnie and Bradley's?

This last question is really what this paper is about. My aim is to give a principled motivation for looking to cases like Ronnie and Bradley's. And it will come in two steps. First I'll give a principled motivation from a broad methodological principle for looking to cases of reasons that are *merely agent-relational*, rather than to reasons that are *agent-neutral*, in a sense that may be unfamiliar, but which I will explain. The second, more controversial, step will be to isolate psychology-explained reasons as a better candidate to generalize from than other categories of merely agent-relational reason, such as those deriving from promises or from special relationships. The first step will occupy the remainder of part 2; I'll offer two arguments for the second in part 3, and another in part 4.

2.2 A METHODOLOGICAL PRINCIPLE

The argument that if we are looking for a uniform explanation of all reasons, merely agent-relational reasons are the most methodologically promising place for us to look, trades on what I think should be an uncontroversial methodological principle. I'll uncover this principle in two stages. First, suppose that you start noticing a lot of shapes like the ones depicted in Figure 1. These shapes seem to have something interesting in common, and if you investigate, you will be able to find all kinds of interesting things about them. They are, for example, the shape that objects which are actually circular occupy in our visual fields, and so if you are, for example, a painter, it would behoove you to learn more about what they really have distinctively in common that explains why they are *that* shape, rather than some other. It might, after all (indeed, it will), help you to recreate them accurately.

But you'll be going about things all wrong if you start trying to figure out what these shapes distinctively have in common that distinguishes them simply by looking at *them*. It will put you off on all sorts of wild-goose-chases. For example, one of the first things you're likely to notice

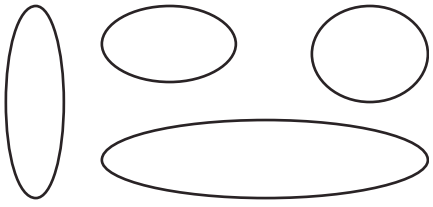

Figure 1

about your shapes is that they are all round. But what ellipses all have distinctively in common—for the shapes that you are trying to investigate are ellipses—is not simply that they are all round *plus something else*. You won't ever find something that you can add to their being round, to give you the right account of what sets them aside as a distinctive class of shapes. To discover the answer to that, you have to look not only at ellipses, but at *foils*—shapes that are like ellipses, but not. In particular, you will want to look at egg-shapes and other non-elliptical ovals. Features that are shared by both ellipses and egg-shapes can be quickly set aside as irrelevant. The Methodological Principle, then, is this:

MP If you want to know what makes *P*s *P*s, compare *P*s to things that are not *P*s.

I want to take this carefully in order to be perfectly clear how uncontroversial the Methodological Principle should be, because I want to emphasize exactly how natural and forceful my motivation for the Broad Humean Theory of Reasons is. But lest I be accused of belaboring the obvious, the Methodological Principle quickly generalizes once we start paying attention to the case of relations. And here my example will be slightly contrived. Suppose that having discovered what ellipses have in common[11] you notice that some people are the *ancestors* of other people, and decide that you want to discover the same thing about this relation, that you have discovered about the property of being an ellipse. It follows from a generalization of the Methodological Principle that some people are not going to be particularly worth investigating, if you are trying to discover what the common explanation is, of what makes one person the ancestor of another.

Eve, who is the ancestor of everyone (I warned you this would be *slightly* contrived) will not be a particularly good place to start, in investigating the *ancestor of* relation. Since she is the ancestor of everyone, she has no non-descendants to compare to her descendants as foils. And so you will suffer from an embarrassment of riches, if you try to sort through all of the things that all of Eve's descendants have in common, in search of the one that makes them her descendants. Since every human being is one of Eve's descendants (as I stipulated), any feature that every human being shares will become a candidate, and you will have no way of ruling any of these out. So Eve's case gives you no privileged *insight* into the *ancestor-of* relation. Being descended from Eve is not being human *plus* anything else, any more than being an ellipse is being round plus something else.

[11] They consist in the set of points whose summed distance from each of two fixed points is the same. (This knowledge *will* help you to depict them more accurately, if you really are a painter, because by tying a thread around two pins, you can use this knowledge to trace any ellipse you like with indefinite accuracy.)

So if you really want to investigate the *ancestor of* relation, the generalization of our Methodological Principle tells us that you need to pay more attention to cases like that of Japheth. Japheth is the ancestor of many people, but he is also not the ancestor of many others. And so we have lots of non-descendants of Japheth to compare to lots of descendants of Japheth. With so many foils, we'll be able to rule out many more potential candidates for what it is that makes Japheth the ancestor of the people who are his descendants. In fact, it is quite likely that there will be *only one* natural candidate for what all of Japheth's descendants have in common but his non-descendants lack: that they are people to whom he stands in the ancestral of the *parent of* relation. So it is quite likely that Japheth's case is going to help you to zero in very quickly on the common explanation of what makes someone the ancestor of someone else. The Generalized Methodological Principle says, then, to pay attention to cases like that of Japheth:

GMP If you want to understand what makes $x_1 \ldots x_n$ stand in relation R, compare cases in which $A_1 \ldots A_n$ stand in relation R but $B_1, A_2 \ldots A_n$ do not, in which $A_1 \ldots A_n$ stand in relation R but $A_1, B_2, A_3 \ldots A_n$ do not, and so on.

Since everyone is a descendant of Eve, Eve's case sets an important *constraint* on a good account of the *ancestor of* relation. The account will be wrong, if it yields the wrong predictions about her case. That is why it is a relief to check and see that Eve does, in fact, stand in the ancestral of the *parent of* relation to everyone. But by the Generalized Methodological Principle, her case is not the right kind of case to give us any particular *insight* into what makes someone the ancestor of someone else. And that is because it leaves us with no useful foils. It allows us to see things that ancestor–descendant pairs have in common, but since it leaves no foils, focusing on this case is like trying to understand ellipses without comparing them to other shapes. It doesn't rule enough out.

2.3 ...APPLIED TO THE CASE OF REASONS

My *ancestor of* case is, as I noted, slightly contrived. It is highly unlikely, to say the least, that Eve is really the ancestor of *everyone*. To be so, she would have to be her own ancestor, which seems rather unlikely to be the case, stipulations aside. So to that extent, the *ancestor of* relation really only approximates the troubles that beset us when we turn our attention to the *reason* relation. For one of the most philosophically salient features of the *reason* relation—and one that we should have fully in view, if we understand the puzzles about the objectivity of morality raised by the

Humean Theory—is that there are some reasons that really *are* reasons for everyone, no matter who she is or what she is like. These *universal*, or *agent-neutral*, reasons of morality, about which the Humean Theory of Reasons is supposed to raise so many puzzles, are supposed to be such reasons. Agent-neutral reasons, in the uncontroversial sense, are like the case of Eve, in that they are reasons for everyone.[12] They may place *constraints* on a good theory about the common explanation of reasons, but they can't give us any important *insight* into what makes some consideration a reason for someone to do something. For in their case we suffer from an embarrassment of riches. There are too many things that everyone has in common for the case to give us any insight into what distinguishes people for whom *R* is a reason to do *A* from those for whom it is not.

So by the Generalized Methodological Principle, it follows that if you want to know what the common explanation of all reasons is, agent-neutral reasons like the reason to help Katie are not going to be a promising place to start. The *promising* place to start is with the case of reasons that are *merely agent-relational*: reasons for some people but not for others. Ronnie and Bradley's is such a case. And so Ronnie and Bradley's case is a much more promising place to look, in order to discover what makes reasons reasons than the case of the agent-neutral reason to help Katie, or any of the other moral reasons.

And that is an interesting result. We might have thought that Humeans are obsessed with cases like that of Ronnie and Bradley because they begin

[12] Unfortunately, both the words 'universal' and 'agent-neutral' turn out to have misleading associations. See Schroeder (forthcoming-a) and (forthcoming-d), for discussion of the difference between the controversial and uncontroversial senses of 'agent-neutral'. In essence, in *The Possibility of Altruism* Nagel (although using the terms 'objective' and 'subjective' at the time) made an uncontroversial distinction between reasons that are reasons for everyone, and reasons that are reasons for only some (1970). But Nagel also adopted the controversial assumption that the only kind of action that a reason can be in favor of, is an action of the form, 'promote state of affairs *p*'. Only given this highly controversial background assumption does Nagel's uncontroversial distinction, which I am putting to use, succeed at tracking the issues of 'agent-relativity' and 'agent-neutrality' that have anything to do with the distinction between consequentialism and deontology. The distinction I am making here therefore has nothing directly to do with the existence of agent-centered constraints, of special obligations, or of agent-centered options.

It is also important to distinguish *universal* reasons from *universalizable* reasons. A reason is *universal* if it is a reason for everyone. A reason is *universalizable*, if its existence follows from a general (universal) principle, of the form, 'for all *x*, if *x* is in conditions *C*, then there is a reason for *x* to do *A*'. So reasons can be universalizable without being universal. See also my Schroeder (2005) for further discussion of this important distinction. For my purposes, getting confused about this is worse than getting confused about whether the distinction has something to do with agent-centered constraints or options, and so I've elected to retain the term 'agent-neutral' as the less confusing of these two options.

with a pre-theoretic prejudice against reasons like the one to help Katie. After all, Christine Korsgaard has claimed repeatedly that the very idea of a Humean Theory of Reasons *starts* with a special focus on reasons like Ronnie's and a chauvinistic attitude about other intuitive examples of reasons, such as the one to help Katie.[13] But the Generalized Methodological Principle explains why it is natural to be interested in cases like Ronnie and Bradley's. For according to the GMP, we *need* to focus on cases of reasons that are merely agent-relational, in order to see what role the agent-place plays in the three-place *reason* relation: *R* is a reason for *X* to do *A*.

But this observation is still insufficient to justify or even motivate the Broad Humean Theory on the basis of our premisses. The observation tells us that *merely agent-relational* reasons are the place that we need to look, in order to see what makes reasons reasons, but Ronnie and Bradley's case is only one *kind* of case of merely agent-relational reasons. The observation explains why the efforts of many philosophers to give explanatory accounts of reasons on the basis of paying special or exclusive attention to moral reasons are straightforwardly methodologically unpromising. But it does not justify paying any more attention to psychology-explained agent-relational reasons than to promise-explained agent-relational reasons, special-relation-explained agent-relational reasons, or any number of others, and that is why the methodological principle only gives us the *first* step in our motivation for the Humean Theory.

Compare: Al promises to meet Rose for lunch at the diner. Andy has made no such promise—he's promised his sick mother to visit her at the hospital. The fact that it's time for lunch is a reason for Al to head to the diner. But it's not a reason for Andy to head to the diner—it's a reason for him to head to the hospital. This difference between Al's and Andy's reasons is explained by their respective promises, rather than as a matter of what they like or dislike, want or don't want, care about or not. In another case, Anne is Larry's infant daughter. That is a reason for him to take care of her. But unless you are in Larry's family or a particularly close friend, it isn't a reason for you to take care of Anne. Now, you might have all manner of reasons to take care of Anne—she might, for example, have been abandoned by her father. But the fact that she is Larry's daughter is not among *your* reasons to take care of her. Here it is Larry's relationship to his daughter that seems to make for a difference between his reasons and yours.

[13] One such argument is the central line of argument in her (1986); a distinct and more general argument to this effect is implicit in the opening pages of her (1997).

So examples of merely agent-relational reasons are ubiquitous.[14] Our Methodological Principle tells us to look at what is distinctive of merely agent-relational reasons, in order to understand reasons in general. But that isn't yet enough to close in on the Humean idea of focusing on Ronnie and Bradley's case, in which the difference in reasons is due to some *psychological* feature. To do that, we need an argument that Ronnie and Bradley's case gives us a *better* insight into what is distinctive of the agent-place in the reason relation than do Al's case or Larry's case. That is, we need to establish an *asymmetry* thesis. My argument for the Broad Humean Theory of Reasons does not rest on ignoring Al's case and Larry's case, or on taking Ronnie's case more seriously. It rests on establishing this Asymmetry Thesis, to which I turn in part 3.

3.1 WEAK ASYMMETRY

I'd like to offer three motivations for the Asymmetry Thesis: a weak, a middling, and a strong. The weak motivation motivates a weak version of the Asymmetry Thesis, but rests on less controversial grounds, the middling motivates a middling version of the Asymmetry Thesis and rests on middlingly controversial grounds, and the strong motivation motivates a very strong version of the Asymmetry Thesis, but rests on very controversial grounds. So they vary from weak to strong in three different dimensions. I'll summarize the weak motivation in this section, rehearse the arguments for the middling motivation in the remainder of part 3, and end up with the strong motivation in part 4; the middling motivation is the one on which I wish to place the most weight for the purposes of this paper, but the broad strategy that I am developing for motivating the Humean Theory can be developed in different ways.

One relevant asymmetry between the case of psychology-explained reasons and other cases of merely agent-relational reasons would be if one of these kinds of reason were a better candidate to generalize in order to explain universal or agent-neutral reasons such as the fact that Katie needs help, which is a reason for anyone to help Katie. According to a common view, it is hopeless to generalize what we know about cases like Ronnie's to cases like that of the reason to help Katie, and that is part of why the

[14] Again, to be clear, since what I am after is agent-relational reasons in the uncontroversial sense, what is crucial here is that the reason for Al to go to the diner is not also a reason for Andy *to go to the diner*—not that it is not also a reason for Andy to make sure that Al ends up at the diner. This further feature of Al's reason is highly relevant—but it is not what the uncontroversial sense of 'agent-relational' tracks.

Humean Theory of Reasons is hopeless. But I have argued elsewhere that it *is* promising to think that the Humean Theory of Reasons may be able to explain agent-neutral reasons such as the reason to help Katie.[15] There is unfortunately no space to rehearse these arguments here.

There is space, however, to consider why it might be thought unpromising to use cases like those of Al and Larry in order to explain reasons like the reason to help Katie. Al has a reason to meet Rose for lunch because of something that he has *done*—some *promise* that he has made. So one might think about contractualist theories of morality as trying to subsume moral reasons under the case of promises, as in Al's case, in this way. But whatever the promise of contractualism in general, we can only use it to subsume reasons like the one to help Katie under cases like Al's if it is based on *actual* contracts, not merely on *hypothetical* contracts. Al has a reason to meet Rose for lunch because he has *actually* made a promise, not because he *might* have made such a promise, if things were different. So only a contractualism based on actual promises could succeed at subsuming moral reasons to cases like Al's. Since that seems unpromising, this seems like an unpromising way to go.

What about cases like Larry's? Could it be that merely agent-relational reasons like Larry's, based on the fact that he is Anne's father, are used to explain reasons like the reason to help Katie? Well, not unless it turns out that everyone is Katie's father. So that doesn't look like a promising view, either. Some authors, however, seem recently to have suggested that being a *fellow human being with* someone is relevantly similar to being the *father of* someone, and that this general relationship, which everyone bears to Katie, can be used to explain reasons in the same kind of way that the fact that Larry is Anne's father can explain agent-relational reasons that Larry has to help Anne.[16] But even supposing this to be true, it would not really be a case of generalizing what we know about Larry's case to all other reasons, because Larry's merely agent-relational reason to help Anne does not derive from the fact that he is a fellow human being with Anne (we all have that reason to help her) but from the fact that *he* is her *father*.

So it is not at all obvious how to generalize other cases of merely agent-relational reasons in a way that would account for the reason to help Katie. It therefore follows that if I am right that Ronnie and Bradley's case *can* plausibly be generalized to account for such reasons, then there is a relevant asymmetry among the obvious cases of merely agent-relational reasons. If we are to look to *any* kind of merely agent-relational reason for insight into

[15] Schroeder (forthcoming-b), (forthcoming-c).

[16] See, for example, Darwall (2006), although I'm not certain that this is the right way to understand Darwall's claims about second-personal authority.

the common explanation of all reasons, as the methodological principle suggests that it should be promising to do, then this asymmetry directs us to look to cases like Ronnie and Bradley's. I haven't discharged the antecedent of this argument, here—that requires another paper.[17] But this illustrates one, weak, way in which we *might* motivate the asymmetry thesis. In the remainder of part 3, I turn to a middling way of motivating the asymmetry thesis that we need, on which I wish to place the most weight for the purposes of this paper. And then in part 4, I will use the results of part 3 in order to state a strong version of the asymmetry thesis.

3.2 THE STANDARD MODEL

Recall that the Methodological Principle does not tell us that cases of agent-neutral reasons *don't matter* for an adequate account of reasons. What it tells us is that like Eve's case, they should operate as a *constraint* on a good account, but they are not likely to give us any particular *insight* into the common explanation of all reasons. My first, weak, strategy for motivating the asymmetry thesis had us look at the prospects for each kind of merely agent-relational reason of being used to account for agent-neutral reasons. My second, middling, strategy for establishing the Asymmetry Thesis goes the other way around. It is to show that most merely agent-relational reasons can be *subsumed* under the case of agent-neutral reasons, but psychology-explained reasons like Ronnie's and Bradley's plausibly cannot. If that is right, then we can treat Al's case and Larry's case as setting constraints on an adequate account of reasons, but like Katie's case, not being particularly good sources of insight into that relation. But if it is right, then we *can't* treat Ronnie's case in this way. And that will be my argument that if we want to look for a common explanation of all reasons, psychology-explained reasons like Ronnie's and Bradley's are the first place that we should look. And this is my central presumptive argument for the Broad Humean Theory.

So consider the case of Al and Andy. Al promises Rose to meet her for lunch at the diner, and Andy promises his mother to visit her at the hospital. As a result, the fact that it is almost noon is a reason for Al to head to the diner and a reason for Andy to head to the hospital. But plausibly, this difference in Al and Andy's reasons can be traced back to a reason that they have in common—to keep their promises. One such reason is that breaking promises tends to destroy their usefulness. Another is that

[17] Schroeder (forthcoming-b).

breaking promises is a breach of trust. Since this is a reason for Al to keep his promises, the fact that he has promised Rose to meet her at the diner for lunch makes heading for the diner at noon necessary for keeping his promises. And since Andy has promised to visit his mother at the hospital, that makes heading to the hospital at noon necessary for *him* to keep *his* promises. So the facts about what promises they have made explain why going *different* places at noon are *ways* for Al and Andy to do the thing that they both have a reason to do—to keep their promises.[18]

It is non-trivial to hold that the difference in Al and Andy's reasons is explained by a further reason that they both share, in this way. Logically speaking, all that we need in order to explain the difference between Al and Andy, is to appeal to the following *conditional*:

Conditional Promise For all x and a, if x promises to do a, then there is a reason for x to do a.

Logically speaking, no one need have any reasons whatsoever in order for Conditional Promise to be true. But I appealed to something *further* in order to explain Al and Andy's reasons:

Categorical Promise There is a reason r such that for all x, r is a reason for x to keep her promises.

In this case, it does seem like Categorical Promise is true. I named two such reasons, and likely there are more. And in this case, that seems to be *why* Conditional Promise is true. So though Al and Andy's reasons differ, that difference can be traced back to an agent-neutral reason. Some philosophers seem to believe, in fact, that *no* conditional like Conditional Promise could ever be true without being backed up with a categorical reason like that in Categorical Promise.[19] But this would be a bold substantive thesis. Logically speaking, Categorical Promise does not follow from Conditional Promise.

Yet the difference between your reason and Larry's can be explained in this same kind of way. Anne is Larry's infant daughter, and that is a reason

[18] Let me immediately head off one source of misunderstanding. When I say that one reason to keep promises is that breaking promises is a breach of trust, I do *not* mean to be suggesting that there is a *further* agent-neutral reason not to breach trust (but not saying what that reason is), and that since breaking promises is a breach of trust, this reason transfers its force to a derivative reason to keep promises. All I am saying is that the fact that breaking promises is a breach of trust is an agent-neutral reason to keep promises. So the explanation that I gave *discharged* the obligation to say *what* the agent-neutral reason from which Al and Andy's reasons derive *is*. But the explanation that I did *not* give *failed* to discharge this obligation—it merely passed it on to the further claim that there is an agent-neutral reason not to breach trust.

[19] I have written about this theory in detail in Schroeder (2005).

for Larry to take care of her, but not a reason for you to take care of her. This, it seems, is because the following conditional is true:

Conditional Child For all x and y, if y is x's infant child, that is a reason for x to take care of y.

Conditional Child backs up a reason for Larry to take care of Anne, but it doesn't back up a reason for you to take care of her. But in this case, also, it doesn't seem like Conditional Child is true all by itself. Like Conditional Promise, it seems to be backed up by a reason that you and Larry *share*—one to take care of whatever children you *do* have:

Categorical Child There is a reason r such that for all x, r is a reason for x to take care of whatever children she brings into the world.

Again, it is easy to come up with such reasons. One is that a person's children are moral subjects who cannot provide for themselves, for whom she is causally responsible. This reason seems to back up Larry's reason to take care of Anne, but to avoid backing up the same reason for you to take care of Anne—Anne, after all, is not *your* child.[20]

Cases like these, in which differences in agent-relational reasons are backed up by an agent-neutral reason, follow what I call the *Standard Model* for reason-explanations.[21] The Standard Model is important and interesting, but all that we need to understand about it here is that in a Standard Model explanation, some class of merely agent-relational reasons is collectively subsumed under an agent-neutral reason from which they derive. What I've illustrated here is that merely agent-relative reasons like Al's and like Larry's can be explained in this kind of way, and hence subsumed under the case of agent-neutral reasons. As such, they place *constraints* on a good account of the common explanation of all reasons, but they don't promise to give us any special *insight* into it.

It is natural to think that all cases of merely agent-relational reasons will be like Al's and Larry's cases in this way—that every time some contingent feature of an agent's circumstances plays a role in explaining why something is a reason for *her* to do something, even though it is not a reason for others to do it, it does so by subsuming her case under a more general agent-neutral reason. The theory that all explanations of agent-relational reasons work in this way is the *Standard Model Theory*. According to the

[20] Again, I do not mean to be saying that there is some more basic agent-neutral reason to take care of moral subjects for whom one is causally responsible. That would not answer the challenge to say what this reason is; it would only put it off. I only mean to be saying that the fact that your children are moral subjects for whom you are causally responsible is a reason for you to take care of them.

[21] See Schroeder (2005), (forthcoming-a), and (forthcoming-c).

Standard Model Theory, though Ronnie's psychological state does play some role in explaining his reason, the role that it plays is a *contingent* one, that can also be played by other kinds of thing. So the possibility of Standard Model explanations is why it doesn't follow from the conjecture that all reasons are explained in fundamentally the same way, and that Ronnie's reason is explained in part by his psychology, that all reasons are in part explained by psychological features. It gives a natural story about how it could be that all reasons really are explained in the same way, and Ronnie's psychological state plays a role in the explanation of his reason, but there are not psychological states in the explanation of every reason. According to the theory, this is because the *role* played by Ronnie's psychology can also be played by other kinds of thing.

But what I'll argue in the next section is that the class of psychology-explained reasons like Ronnie's *can't* be subsumed under agent-neutral reasons in this kind of way. The Standard Model Theory, that is, is false. And that will be the asymmetry that I will argue gives us middling warrant to hold that Ronnie's case is a more promising place to look in order to see what role the *agent*-place plays in the *reason* relation.

3.3 IS THERE AN AGENT-NEUTRAL REASON TO PROMOTE YOUR DESIRES?

To have a Standard Model explanation of reasons like Ronnie's, we need two things. First, we need an action-type A such that in every case like Ronnie's, the action the reason is for is a *way* for the agent to do A. And second, we need a reason, R, that is a reason for anyone to do A. It is easy to see how to construct the appropriate A and R in the paradigmatic cases in which the Standard Model is motivated. What Rachel has a reason to do on both Monday and Thursday is to write about whatever she is thinking about at the time. And the reason for her to do this is that it has been assigned by her poetry professor. Because this is a reason for Rachel to write about whatever she is thinking about, it follows that no matter what Rachel is thinking about, she has a reason to write about that.[22]

But unfortunately, it is quite difficult to construct the appropriate A and R for the full range of cases like Ronnie's. Here I will assume for the sake of argument that there *is* some action A such that all actions for which there are psychology-explained reasons are *ways* of doing A. For the sake of argument, I will assume that this is the action of *doing what you want*. It is

[22] See Schroeder (2005) for an extended discussion of Rachel's case.

unclear, I think, whether any such action-type will do the required work for the Standard Model, but the issues are complicated. I will confine myself to arguing that even if there is some such action *A*, there is no good candidate, *R*, for what the agent-neutral reason is to do this thing. If there is not, then the Standard Model Theory is, I think, wrong, and wrong in an interesting way. The way in which it is wrong leaves a relevant asymmetry between psychology-explained and other merely agent-relational reasons. And from the preceding considerations, that means that reasons like Ronnie's are the most promising place to look for a unified explanation of all reasons.

This may seem like a silly view. It may seem obvious that there is a reason to do what you want. But we have to be careful how we understand that claim, and consequently we should be suspicious about whether the thought supports the Standard Model in any way. Compare the following:[23]

Easy For all *x* and *a*, if doing *a* is what *x* wants, then there is a reason *r* for *x* to do *a*.

Mid For all *x*, there is a reason *r* for *x* to: do what *x* wants.

Hard There is a reason *r* that is a reason for all *x* to: do what *x* wants.

The problem is that in order to get a Standard Model explanation of the full range of cases like Ronnie's, **Hard** must be true. But it is not at all obvious that **Hard** is true (that is why I called it '**Hard**'). At best, it is **Easy** that is obvious.

Consider the case of Brett. Brett wants to finish his Ph.D. in philosophy. Working on his dissertation on the pragmatics of context-dependence promotes finishing his Ph.D. in philosophy, and so there is a reason for Brett to work on his dissertation on the pragmatics of context-dependence. Moreover, it is easy to see what this reason is. It is that working on his dissertation will enable him to finish his Ph.D. But Brett also wants to become a rock star. Recording a new album with his band will promote this aim. And so it seems that there is a reason for Brett to record a new album with his band. Moreover, it is easy to see what this reason is. It is that recording a new album with his band is necessary in order to get picked up by a label, and hence in order to become a rock star.

Obviously, the reasons for Brett to do these two things are different. Examples like this (at least, enough of them—one for every want) are enough to make **Easy** true. But for **Mid** to be true, there must be a *further*

[23] Here I bracket the question of whether these claims are *sufficient* as stated. We're interested in the view that psychological states like desire play a *necessary* (but not necessarily sufficient) role in the explanation of reasons. If you think some further condition is also required in order to *complete* this explanation, by all means build it in. This question is orthogonal to the one that I am pursuing here.

reason for Brett to *do what he wants*, some fact about the world that is both a reason for Brett to work on his dissertation and a reason for him to record a new album with his band. And for **Hard** to be true, this reason, whatever it is, must also be a reason for Ronnie to go to the party, for Vera to practice playing chess, for Christina to buy a new cookbook, for Bill to hike the Appalachian Trail, and so on. What single state of the world could possibly tell in favor of such a rich and diverse class of actions? I don't see what it could be, and no one who believes that there is such a reason has ever given me a good answer as to what they think that it is, either.

The idea I hear most often is also the most unpromising, so let me set it aside, here. The conjecture that I hear most often is that the reason *r* which makes **Hard** true is just the truth of **Hard** itself! How convenient! Unfortunately, also how circular. Even if the truth of **Hard** does satisfy the condition that **Hard**'s existential quantifier governs, it simply *can't* be the only thing that does. For in order to be such a reason, it must first be *true*. But in order for it to be true, there must first be such a reason. So it can't be the only one. The fact that I so often hear this hopeless answer seems to me to be evidence that no one does have any good idea of what consideration it could be that makes **Hard** true.

So despite appearances, it should not be at all obvious that there *must* be *some* agent-neutral reason to do what one likes. What should be obvious is that a Standard Model explanation of psychology-explained reasons like Ronnie's owes us something significant. It is committed to holding that there *is* some such reason. And so it should be able to tell us what this reason is. I myself don't know what this reason is. I have no *proof* that there is no good answer as to what it is, but no one, no matter how confident that there *must* be some such reason, has ever given me a satisfactory answer as to what it is. And so I remain suspicious that their convictions that there is such a reason arise not from knowing what it is, but because they are in the grip of a theory—the Standard Model Theory. This constitutes my second, middling, motivation for the asymmetry thesis.

4.1 THE ARGUMENT IN BRIEF

So in sum, this is my argument for the Broad Humean Theory of Reasons, given the middling motivation for the asymmetry thesis:

1 Ronnie's reason is explained by some feature of his psychology.

2 All reasons are, at least at bottom, explained in the same kind of way.

3 From the *Generalized Methodological Principle*, agent-neutral reasons should function as a *constraint* on a good unified explanation of reasons,

but they don't give us a promising place to look for how that explanation *works*.

4 From the *Asymmetry Thesis*, all merely agent-relational reasons *other* than the psychology-explained ones can be successfully subsumed under the case of agent-neutral reasons.

C So psychology-explained reasons like Ronnie's are the *most methodologically promising* place to look for features of how the uniform explanation of all reasons must work.

I don't claim that this argument gives more than a presumptive motivation for the Broad Humean Theory of Reasons. All it tells us is that Ronnie and Bradley's case is a *methodologically promising place to look* for an explanation of reasons, *so long as* we aspire for a uniform explanation. But I *do* claim that this argument gives us a *very good* presumptive motivation for the Humean Theory, which is all that I am after.

Premiss **1** is weak enough to be uncontroversial—or at least, to create a quite significant cost to rejecting it. Premiss **2** is *not* uncontroversial, but it represents an appropriate and reasonable ambition for philosophical theory. Premiss **3** is backed by a genuinely uncontroversial methodological principle. And I've argued carefully for premiss **4** in part 3 of this paper—if you think it is false, you're welcome to propose what the action and reason could possibly be that would make a Standard Model explanation of all of the reasons like Ronnie's turn out to work, without raising problems of its own. And if that fails, there is still the weak motivation for the asymmetry thesis from section 3.1. Once we recognize the Methodological Principle and apply it to reasons, we only need *some* relevant asymmetry in order to generate *some* kind of motivation for the Broad Humean Theory of Reasons.

4.2 REVISIONIST AND CONSERVATIVE HUMEANISM

Notice that I have *not* claimed that Katie's case, Al's case, Larry's case, and others like them, do not place important *constraints* on an account of reasons. On the contrary, I compared these cases to that of Eve in the *ancestor of* case. Though Eve's case did not in and of itself give us any special insight into the *ancestor of* relation, I claimed that it did place an important constraint on a successful account of that relation. Similarly, I claim that Katie's case, Al's case, and Larry's case place important constraints on a successful account of reasons. I hold that it is a serious mark against any theory of reasons that it fails to account for such reasons.

Distinguish two kinds of Humeanism—*revisionist* and *conservative*. The revisionist Humean is happy to embrace the kinds of skeptical results

about the objectivity of morality that I discussed in section 1.2. When the revisionist Humean says that all reasons must be explained by a psychological state just like Ronnie's is, she means that there is no special reason for everyone to help Katie, nor for Al to meet Rose for lunch, and so on. But when the *conservative*, or *sophisticated*, Humean says that all reasons must be explained by a psychological state just like Ronnie's is, he doesn't mean to be denying that there is a reason for anyone to help Katie no matter what he is like; he is merely making a theoretical claim about that reason's genesis.[24]

The sophisticated Humean's theory may ultimately fail to successfully explain all of the reasons for which he wants to account. If it does so, then he is forced to take a revisionist view. And that can lead, ultimately, to skeptical results about the objectivity of morality. But the motivation that I am offering for the Broad Humean Theory of Reasons is, at least initially, *sophisticated* in outlook. What I am offering is simply a methodological consideration in favor of expecting that Ronnie and Bradley's case should give us a special *insight* into what explains all reasons. And *that*, I would have thought, is all that we need in order to have excellent presumptive motivation for finding the Broad Humean Theory of Reasons *attractive*. It is certainly enough to dispel the illusion that the only reason anyone would believe the Humean Theory is because they were committed to the Classical Argument. And that should be enough to dispel the idea that motivation by the Classical Argument can be taken for granted when evaluating the prospects of the Broad Humean Theory of Reasons.

4.3 CODA: *HOW IS* RONNIE'S REASON EXPLAINED?

One of the principal advantages that I've claimed for my motivation for the Humean Theory of Reasons is that it makes no discriminations among *forms* that the Humean Theory of Reasons might take. It leaves for investigation just *how* the explanation of Ronnie's reason actually works—for example, what kind of psychological state explains it, but also many other questions about how the explanation works. Since we've seen that the Humean Theory cannot accept the Standard Model explanation of Ronnie's reason, and since I've argued in part 3 that this explanation is suspicious anyway, I want to close by offering an alternative way of understanding how Ronnie's reason *does* get explained by his psychology, which leads to an interesting conjecture, which leads to a third, strong, version of the asymmetry thesis, and hence a further, related, argument for the Broad Humean Theory of Reasons.

[24] See Schroeder (forthcoming-b).

The fact that there will be dancing at the party tonight is a reason for Ronnie to go there, but not for Bradley to go there. And this is because Ronnie, but not Bradley, desires to dance. For this explanation to be true, something like the following has to be the case:[25]

Expl For all agents *x*, if *R* helps to explain why *x*'s doing *A* promotes *p*, and *p* is the object of one of *x*'s desires, then *R* is a reason for *x* to do *A*.

Expl is a generalization under which we can subsume Ronnie's case. In Ronnie's case, the fact that there will be dancing at the party tonight helps to explain why going to the party will promote one of Ronnie's desires. For it helps to explain why going to the party will be a way for Ronnie to go dancing, and dancing is something that Ronnie desires to do. But since Bradley doesn't desire to go dancing, it doesn't follow from **Expl** that this is a reason for Bradley to go to the party.

The Standard Model Theory would have it that positing generalizations like **Expl** is not enough to explain Ronnie's reason. For on the Standard Model Theory, as we have seen, **Expl** itself needs to be explained. *Why* is it that **Expl** is true? On the Standard Model Theory, this question must be answered by appealing to a *further* action that there is a reason for everyone to do. But as I've argued, we *can't* successfully do that in this case.

But that doesn't mean that **Expl** must be unexplained. Compare **Expl** to another explanatory generalization. We can say that the Bermuda Triangle is a triangle, in part, because it has three sides. This is because the following generalization is true:

Tri For all *x*, if *x* is a closed plane figure consisting of three straight sides, then *x* is a triangle.

But no one thinks that for **Tri** to be true, there has to be a further shape, over and above triangularity, that is had by everything, and explains why everything has the conditional property postulated by **Tri**. On the contrary, people are likely to think that **Tri** is true simply because it states *what it is* for something to be a triangle. It is because triangularity *consists* in being a closed plane figure consisting of three straight sides, that **Tri** is true.

So I offer **Tri** to the Humean as a model for how the explanation of how Ronnie's reason works, if it does not follow the Standard Model. On this view, a desire helps to explain Ronnie's reason, because there being such a desire is part of *what it is* for Ronnie to have a reason. That is just what reasons are, just as triangles are simply three-sided plane figures. Like the Standard Model, this is a substantive view about *how* Ronnie's

[25] The account given here is the one that I defend in Schroeder (forthcoming-c), but the details are irrelevant for this point.

desire helps to explain his reason. But it is an intelligible alternative to the Standard Model. And as such, it suggests the following alternative simple argument for the Humean Theory of Reasons, based on what we might call the *Standard-Constitutive Conjecture*:

1 Ronnie's psychology helps to explain his reason.

2 The Standard Model does not successfully account for how it does so.

3 Conjecture: the constitutive model of **Tri** is the only alternative to the Standard Model.

HTR If so, then being in the kind of psychological state that Ronnie is in must be part of *what it is* to have a reason. So in every case of a reason, there must be some such psychological state.[26]

REFERENCES

Bond, E. J. (1983) *Reason and Value* (Cambridge: Cambridge University Press).

Darwall, Stephen (1983) *Impartial Reason* (Ithaca, NY: Cornell University Press).

____ (2006) *The Second-Person Standpoint: Morality and Accountability* (Cambridge, Mass.: Harvard University Press).

Foot, Philippa (1975) 'Morality as a System of Hypothetical Imperatives,' reprinted in *Virtues and Vices* (Oxford: Oxford University Press, 2002).

____ (2001) *Natural Goodness* (Oxford: Oxford University Press).

Harman, Gilbert (1975) 'Moral Relativism Defended,' reprinted in *Explaining Value and Other Essays in Moral Philosophy* (Oxford: Oxford University Press, 2000).

____ (1978) 'Relativistic Ethics: Morality as Politics,' reprinted in *Explaining Value and Other Essays in Moral Philosophy* (Oxford: Oxford University Press, 2000).

____ (1985) 'Is There A Single True Morality?,' reprinted in *Explaining Value and Other Essays in Moral Philosophy* (Oxford: Oxford University Press, 2000).

Hampton, Jean (1998) *The Authority of Reason* (Cambridge: Cambridge University Press).

Hooker, Brad (1987) 'Williams' Argument against External Reasons,' *Analysis* 47: 42–4.

Hubin, Donald (1999) 'What's Special about Humeanism,' *Noûs* 33: 30–45.

Joyce, Richard (2001) *The Myth of Morality* (Cambridge: Cambridge University Press).

Korsgaard, Christine (1986) 'Skepticism about Practical Reason,' *Journal of Philosophy* 83: 5–25.

[26] Special thanks to Stephen Darwall, David Copp, Sari Kisilevsky, Russ Shafer-Landau, Ralph Wedgwood, Rob Shaver, Gideon Rosen, Gilbert Harman, Michael Morreau, Scott James, Aaron James, two readers for Oxford University Press, and audiences at the College Park Conference on Practical Rationality and the second annual Wisconsin Metaethics Workshop. Mark Murphy, in particular, provided very helpful comments and discussion at the Maryland conference and following.

____ (1997) 'The Normativity of Instrumental Reason', in G. Cullity and B. Gaut (eds.), *Ethics and Practical Reason* (Oxford: Oxford University Press).
Mackie, J. L. (1977) *Ethics: Inventing Right and Wrong* (New York: Penguin).
Millgram, Elijah (1996) 'Williams' Argument against External Reasons,' *Noûs* 30: 197–220.
____ (1997) *Practical Induction* (Princeton, NJ: Princeton University Press).
Nagel, Thomas (1970) *The Possibility of Altruism* (Princeton, NJ: Princeton University Press).
Schroeder, Mark (2004) 'The Scope of Instrumental Reason,' *Philosophical Perspectives* 18: 337–64.
____ (2005) 'Cudworth and Normative Explanations,' *Journal of Ethics and Social Philosophy*, 1, **www.jesp.org**.
____ (forthcoming-a) 'Reasons and Agent-Neutrality,' forthcoming in *Philosophical Studies*.
____ (forthcoming-b) 'Weighting for a Plausible Humean Theory of Reasons,' forthcoming in *Noûs*.
____ (forthcoming-c) *Slaves of the Passions*, forthcoming from Oxford University Press.
____ (forthcoming-d) 'Teleology, Agent-Relative Value, and "Good",' forthcoming in *Ethics*.
Setiya, Kieran (2004) 'Hume on Practical Reason,' *Philosophical Perspectives* 18: 365–89.
Williams, B. (1981) 'Internal and External Reasons,' in *Moral Luck* (Cambridge: Cambridge University Press).

10

Responding to Normativity

Stephen Finlay

To many it seems obvious that normativity or justification depends upon desire. Few answers to the question, 'Why should I?' seem more natural than 'Because I want to,' and if we are told, 'You should do this,' there is something natural about the objection, 'But I don't want to, so why?' I believe that the very nature of normativity can be comprehensively explained in terms of desire: the mysterious 'force' of value, reasons, and obligation are explicable by appeal to the 'force' of our motivating psychological states. This *desire-based normativity* (DBN) thesis faces serious difficulties, however, that seem insuperable to most sophisticated minds who contemplate them. I remain convinced that DBN is correct, although as yet unvindicated. This paper seeks to lay the cornerstone of what could prove a successful strategy, sketching an Argument from Voluntary Response that is based on the autonomous character of our experience of normative authority and the voluntary character of our responses to it.

In the first section, I consider the fortunes of its ancestor, the rickety standard Argument(s) from Motivation. The second section sketches an account of what it is to desire, the third explores the character of experience and response to normativity, and the fourth examines the necessary conditions for voluntary behaviour. The fifth section explores what implications the argument contained in sections 2–4 have for the plausibility of the DBN and anti-DBN models of response to normativity, and the final section provides some reflections on the question of how to bridge the gap

I would like to thank Sarah Buss, Pamela Hieronymi, Sam Schpall, Mark Schroeder, Gideon Yaffe, audiences at the 2nd Annual Metaethics Workshop on Madison, Wisconson and the 10th Annual Southern California Philosophy Conference, and the Oxford University Press referees.

between DBN and the argument's more modest conclusion, that *response* to normativity is based on desire.

1. ARGUING FROM MOTIVATION

Why even suppose that normativity depends on desire? Desires, it is objected, are merely motivating psychological states. How could it follow from the fact that somebody is motivated to make it the case that *p* that he has a reason or *ought* to make it the case that *p*? Hume's own 'Law' of no-ought-from-is can be utilized here against the 'Humean' desire-based view of normativity. The case for DBN is standardly presented by various forms of argument from motivation.[1] (In this paper I focus on practical reasons rather than value or ought-facts.) These arguments have two main premisses: the first is some form of *motivational internalism* (MI): having, judging that one has, or judging that something is a normative reason to act has some especially close connection to being motivated to act. The second premiss, sometimes known as 'motivational Humeanism' (MH), holds that being motivated requires desire. The arguments conclude that normative reasons are based on desires.

These arguments differ considerably from one another, but it is now widely recognized that they all seem to fail somehow. First, strong forms of MI appear simply implausible. Some agents' normative judgments (particularly moral judgments), for example, seem not to provide them with any motivation whatsoever, even overridden motivation. Normative judgments seem only *sometimes* to motivate us. But plausibly weak forms of MI are insufficiently strong to support the inference to the conclusion, DBN. A weaker motivational connection might be explained by a contingent combination of normative judgment and independent desire (e.g. to act on one's best reasons), in which case the reason itself need not be based on a desire in order to have the requisite connection with motivation.

Opponents of DBN however mostly accept some form of MI, and are more interested in pointing out the flaws in MH. Motivation of action is a causal process,[2] and it is a contingent and a posteriori matter as to what

[1] Such arguments are presented in Hume (1978: 457) and Williams (1981), and discussed in Cohon (1988), Wallace (1990), Wedgwood (2002), Heuer (2004). This is not the only argument around, however: see for example Mark Schroeder's contribution to this volume.

[2] Some philosophers disagree, but I shall not explore this controversy here, and instead direct the reader to the discussion in Mele (2003: ch. 2).

causes what. Then what are the grounds for the claim that beliefs are never sufficient and desires always necessary for motivation of action? We do not have satisfactory empirical evidence to justify the claim, and indeed the premiss is generally regarded by supporters of DBN as an a priori truth in no need of empirical support. But surely, it is objected, a claim about causal conditions cannot rightly be thought a priori.[3] MH is therefore accused of being a mere dogma.

The only solution, it seems, is to define 'desire' such that MH has to be true. *Motivational* and *dispositional* analyses of the concept of desire do precisely this. According to motivational accounts, to 'desire' that *p* just is to be motivated towards making it the case that *p*.[4] Some version of MH does then acquire the status of an a priori principle, but at the cost of triviality, and the Argument from Motivation is rendered obviously invalid. Indeed this is the aim of many proponents of motivational accounts, which support a version of MH that requires desire only as a *logical consequence* of motivation and therefore not as a possible cause or metaphysical condition of it. One is motivated to action if and only if one has some desire, but not *because* one has that desire; rather it is one's being motivated to action that makes it the case that one has the desire. This version of MH undermines rather than supports DBN; the claim that normativity is *based* on desire clearly invokes some kind of metaphysical dependence of normativity on desire.

According to dispositional accounts, to 'desire' that *p* is just to be disposed under certain circumstances to act in certain ways: in particular to try to make it the case that *p*.[5] While dispositions can be causes rather than mere 'logical shadows', these accounts also succeed only at the cost of triviality: they rule out no coherent account of the causation of action.[6] Like motivational accounts, therefore, they open the way for cognitivist accounts of motivation: certain (normative) beliefs or their contents themselves directly motivate action, thereby entailing the existence of any requisite desires. This cognitivist strategy is classically illustrated by Stephen Darwall's story about Roberta, who learns of the suffering of textile workers in the southern United States and is motivated to act by her recognition that their

[3] It might be suggested that although not a priori, MH is confirmed by the empirical observation that even paradigms of normative beliefs (e.g. that ϕ-ing is what I ought to do all things considered) sometimes fail to motivate. We could therefore infer that something besides normative belief is needed. But (a) it is unclear why this must be a desire, and (b) such cases may be due to the presence of an inhibitor rather than the absence of an enabler; see Cohon (1988); Dancy (2000).

[4] Nagel (1970: 29); Darwall (1983); Schueler (1995); Dancy (2000).

[5] Smith (1994); Stalnaker (1984); Heuer (2004).

[6] Darwall (1983: 42); Platts (1979: 256); Heuer (2004: 57); see also Ross (2002: 205).

plight is a reason for her to assist in the efforts to force labour reform, without having any preceding desire that explains her motivation (1983: 39–41). Arguments from Motivation for any interesting form of DBN need MH to claim something less anemic.

Can the Argument from Motivation be made to work? I remain convinced of the truth of DBN not because normative beliefs are causes of behaviour, but because of the character of our experience of and response to normativity. We experience normativity as autonomous authority, and we respond to it with voluntary activity. I shall now attempt to sketch such an Argument from Voluntary Response, which can be seen as a revision rather than a replacement for the Argument from Motivation provided that the concept of motivation is understood as I shall suggest.[7] I believe DBN is true not because normativity or normative belief merely *causes behaviour*, but rather because we *respond voluntarily* to it (i.e. it *motivates action*). Motivation is a form of causation and action a form of behaviour, but they are special kinds of causation and behaviour. While I have no objections to the possibility of a belief causing behaviour unassisted by any desire, I shall argue that no behaviour can be voluntary (or 'motivated') if its causes do not include in the appropriate way some desire. I conclude that desire is necessarily a cause of any response to normativity. The argument has the following general form:

VB-RN: Necessarily, all responses to normativity are voluntary behaviours;
DB-VB: Necessarily, all voluntary behaviours are caused by desire;
Therefore,
DB-RN: Necessarily, all responses to normativity are caused by desire.

Two concessions are needed: (i) it will not yet be clear how **DB-RN** presents a difficulty for anti-DBN models of normative motivation. To this end I shall address Roberta's case in detail in section 4. (ii) There is a significant gap between **DB-RN** and DBN, the desire-dependence of our *response* to normativity and the desire-dependence of normativity itself; this is addressed programmatically in section 6. But first in the order of business is an investigation of the concept of desire.

2. WHAT IS DESIRING?

Against dispositional accounts we must observe the difference between 'dispositional' and 'occurrent' desires (or 'wants'). There are all sorts of

[7] Hume (1978: 457) and those following his treatment (e.g. Cohon 1988) do not make it clear that this is their conception, writing rather of normative beliefs merely 'moving' or 'influencing' us.

things I can be said to want of which there are no traces in my current psychological activity. But actively desiring something is different, and involves some form of mental activity or process. 'Dispositional desires' are just a kind of disposition to desire occurrently.[8] I shall therefore switch from the noun to the verb: what is it *to desire* something?[9] The correct analysis of occurrent desiring, I believe, is *teleological* and *intrinsic*: to desire that *p* is for one's mental activity to aim at the goal that *p* 'for its own sake'.[10]

Unlike motivational and dispositional accounts, this account does not falsely claim that whatever we aim at we desire, since all genuine desiring is intrinsic: I am only (genuinely) desiring that *p* if I am aiming at its being the case that *p* 'for its own sake,' i.e. not in virtue of my aiming at any further end.[11] I mean here to deny the existence of 'motivated' or 'derivative' desiring altogether; i.e. to claim that all desiring is 'basic' or 'brute'. It is commonly thought that desire can be 'motivated' in two different ways: (a) by other desires—hence 'derivative' desires—and (b) by reasons (or value, norms, etc.)—hence 'rational' desires. In rejecting motivated desires I do not deny that desires are caused,[12] or even that they can be caused by other desires or normative beliefs, but merely that such causation is ever an instance of motivation. The case against (b) rational desires requires the entire argument of this paper, and so must here be set aside. I defend the rejection of (a) derivative desires on the grounds of (i) their incompatibility with our considered desire-ascriptions, and (ii) their redundancy.

(i) Ordinary wisdom tells us that we can perform actions that we don't desire to perform and pursue states of affairs that we don't desire to obtain. While this is sometimes thought to deliver a decisive blow to MH, against this account it has no force at all. It is, I submit, precisely the things that we do or pursue merely as means (e.g. visiting the dentist, rising at the crack of dawn, inserting coins in a vending machine) that we are disposed to

[8] There is a common intuition that we don't attribute agents' desires on the basis of dispositions that have never been activated.

[9] There are numerous theories of desire which I cannot discuss here, including the phenomenological theory, the judgment theory, the directed attention theory (Scanlon 1998), and the reward theory (Schroeder 2004). To borrow a joke, for every five moral psychologists there are seven theories of desire.

[10] Smith (1994); Lenman (1996); Ross (2002) also offer teleological accounts. Smith presents his dispositional account as an elucidation of his teleological account, but I doubt their compatibility: aiming at something is not the same as being disposed to act in certain ways.

[11] See also Chan (2004).

[12] Some are concerned that this Humean view of desire is committed to the implausible denial of the possibility of acquiring new desires beyond those we have innately at birth (e.g. Cohon 1988). I see no grounds for this concern: we are psychologically disposed to develop new desires through association, transference and other contingent mechanisms.

concede, at least on close questioning, that we don't *really* desire or want.[13] (ii) Since desires are individuated by their ends, rather than by the actions that they motivate, we can explain pursuit of means by appeal to the desire for the end without having to invoke any desire for the means.[14] Derivative desires are therefore redundant:[15] given desire for an end, we have reason to pursue the means, not any reason to acquire in addition a new desire.

The two most pressing objections can be met by a single response. First, it may seem implausible that desiring that *p* entails aiming at making it the case that *p*, for we have many desires that we do not act on and desire many ends that do not become objects of our pursuits.[16] Second, this account may appear to share the failing of dispositional and motivational analyses with respect to its support for DBN: if desiring an end is identical with the activity of aiming at that end, then we cannot coherently maintain that desiring the end *causes* or motivates us to aim at the end: rather the desiring and the aiming both must have some other cause and explanation.

The solution to both problems is to resist an excessively simple-minded view of action or activity. Desiring is a mental and not a physical, bodily, or *overt* activity, and by 'aiming' I mean to refer to mental rather than overt behaviour or action.[17] The thought activity that precedes overt pursuit of ends is also a form of 'aiming'. Suppose I desire that I drink a soda. The overt action of making it the case that I drink a soda may be constituted by leaving my office, going to the vending machine, inserting coins, etc. None of this activity is desiring. But before I can perform these actions I must direct my practical thought,[18] plotting a path to the end. This involves

[13] Our frequent ascriptions of such desires may be thought to show that this analysis fails to capture the ordinary concept. I think they can be accounted for by appeal to the difficulty of identifying the true objects of our desires and the low precision required for ordinary communication, and that they can be discounted by appeal to our disposition to withdraw them under cross-examination.

[14] I defend this view against Korsgaard and Nagel in my (forthcoming). Mele (2003: 93) rejects such views on the ground that force doesn't 'flow out of' (i.e. diminish the strength of) motivation toward the end when we derive motivation towards the means. But this seems to assume that motivational force resembles the flow of water rather than the flow of electricity.

[15] Don't we need derivative desires to enable pursuit of long-range goals without having to keep them constantly in mind? The evolutionary fitness of such desires suffices to explain our disposition to be *caused* to form desires for the means to our desired ends (de re); there is no need for a *motivational* link.

[16] My use of 'ends' may cause confusion: it has a connotation of *intention* which I do not intend. As I mean it, an 'end' is simply a conceivable state of affairs.

[17] This is perhaps a non-standard use of 'action', and it therefore ought to be flagged that this paper employs the term in this broader way.

[18] These processes may be fleeting. The period of cogitation involved in an instance of desiring may be relevant to its degree of phenomenological presence. We are especially conscious of desirings that occupy considerable amounts of our time.

thinking my way around not merely physical obstacles (identifying means), but also mental ones: desiring must contend with other desires, and takes the form of seeking a path to the end around the constraints posed both by the world and by conflicting desires. This includes, minimally, looking for ways of preserving the prospects for the end while pursuing other more pressing ends.

It may be objected that since desiring can occur without intending it need not involve any aiming at ends. But to have an intention, in my view, is approximately to have found a path to one's desired end that is not blocked by any of one's other desirings; (occurrently) intending that *p* therefore entails desiring something (*p* or some projected consequence), but differs from mere desiring simply in involving being *settled* on making it the case that *p*. To reach intention, desiring must first survive some hazards. Many desirings are stunted by recognition that there are no available means to their ends and, if not immediately abandoned, are diminished to activities of surveying the scene for emergence of a route.[19] Other desirings awaken slumbering beasts that devour them: stronger desires. For example it is arguable that in becoming mature deliberators we acquire through negative reinforcement a prudential disposition to desire that we not bring suffering upon ourselves, activated upon the contemplation of action.

These mental activities are often causal antecedents of physical actions and of further desiring activity. We can therefore say that desiring causes action. Indeed (I shall argue) behaviour not preceded by such mental activities[20] is not 'action,' as it does not stem from any agency, and therefore desiring is a conceptually *necessary* cause of acting, just as intending to kill is a necessary cause of murder, provided that intentional killing is in its definition.[21] However, explanation of a particular action by appeal to a

[19] A problem arises from desires for states of affairs obviously out of our control: Dancy (2000: 87–8); Mele (2003: 22–7); Schroeder (2004: 16). If I desire that the Chicago Cubs win the World Series, whatever I am thereby doing surely it is not seeking to make it the case. Three responses: (i) this may be a case of misdescribing the content of desire, which may be rather that I savour such victory—the means to which are partially within my control. (ii) If the impossibility of advancing the end is to inhibit our behaviour, there must be a primitive stage of mental activity at which we encounter it. Perhaps such desires are stopped short by such recognitions. Typical fan behaviour supports this: shouting at the TV, muttering prayers, egging the team on, and rehearsing advice to coach or players. (iii) I am skeptical that theories immune to this problem can individuate desires by their content. If (per Mele) such a desire might manifest itself in seeking to learn whether the team wins, what differentiates it from the desire to learn whether the team wins?

[20] We must include aversion, the negative form of desire, but for simplicity's sake I will not differentiate.

[21] Mele's example: the US Treasury is a necessary cause of a US dollar bill (2003: 53).

particular desire is non-trivial, because it is contingent *which* desire causes any action.[22]

I maintain that the concept of desiring an end is the concept of engaging in practical thought or mental activity aiming at promotion of that end for its own sake. This account plays an important role in the argument that follows. I acknowledge, however, that closer scrutiny is needed. In particular, I am relying on an unanalyzed notion of *aiming at an end* (which I cannot, and would not, attempt to explain by appeal to desire). Notwithstanding the difficult philosophical problems in explaining teleology, I trust that it has sufficient intuitive clarity to legitimize my doing so. Those unpersuaded by my analysis of the concept of desire may therefore read me as arguing for a kind of *teleology*-based normativity. I argue for the further link to desire nonetheless because of its central role in this debate.

We can reach another significant result when we combine this analysis with the following reasonable claim: necessarily, all mental activity of aiming at some end performed by finite creatures is either intrinsically directed at that end, or an instance of mental activity of aiming at some further end intrinsically. This rules out the possibility of infinite (linear or circular) regresses of ends, and yields the conclusion that all end-directed mental activity must constitute desiring some end or other.

To establish the promised link to the Argument from Motivation, consider the question of what kind of causation of behaviour we mean by 'motivation'. The word itself gives us the crucial clue: to motivate behaviour (in the 'success' sense)[23] is, I suggest, to cause it by way of providing a *motive* for it—i.e. an end or goal at which the agent aims. Motivation is therefore an essentially teleological form of mental causation,[24] and as such necessarily involves causation by desire, given the account of desire defended above. This is so despite the fact that mental states can 'provide' motives for actions in different ways. While desires constitute or contain motives, beliefs can 'motivate' an action either by stimulating (i.e. causing us to commence) desiring that produces the action, or by instructing us that the action promotes some already desired end. This is not to say that there is more than one kind of process by which action is produced, but rather

[22] It is even non-trivial to explain a particular thought process that constitutes some desire by appeal to that very desire, just as it is non-trivial to explain a murder as a knifing, shooting, or poisoning.

[23] There is also a non-success sense on which one can be motivated towards some action without attempting it. This is easily accommodated: motives can be provided for actions that nonetheless fail to eventuate.

[24] See also Smith (1987–8: 251). Here 'teleological causation' means merely non-deviant causation by end-directed psychological states, and should not be confused with causation effected by the future.

that beliefs can be said to motivate on the basis of two different relations to desire: as cause/stimulus, and as channel/navigator.[25] Motivation of behaviour is therefore teleological or non-deviant causation of behaviour by desire.

My case for this account of motivation has been hasty, I concede, but as nothing significant will ride on it in this paper, the unpersuaded may take it as stipulative. This may seem to beg the question in favour of DBN, since the easiest path onwards to that conclusion would be to claim next that normative judgments do indeed *motivate* us, and hence that they must depend on desires. But I will not argue this way; instead I shall take seriously as an idea needing refutation that there may be agential or voluntary forms of causation other than motivation. (R. Jay Wallace (1990) and Michael Smith (1987–8: 252), for example, propose that inferences between mental states may vindicate a non-DBN model of normativity.) No significant questions will therefore be begged. Indeed this account of motivation may seem to concede DBN's opponents everything they want: beliefs like Roberta's *can* motivate us by causing us to desire something. However the question is whether this kind of motivation by belief can constitute motivation or causation by normativity, and I will argue that only the other kind—in which beliefs motivate an action by revealing its relation to an occurrent desire—can constitute a response to normativity, because of the voluntariness of such responses.

3. RESPONDING TO NORMATIVITY

I have conceded to DBN's opponents that beliefs can stimulate desiring without the contribution of any occurrent desiring. This is not sufficient to establish that response to normativity is possible without contribution from occurrent desiring, however, because mere causation of desire by belief is not sufficient for a response to normativity. Two further conditions must be met: (i) the 'response' must have the right sort of causal antecedent, and (ii) which must causally operate in the right sort of way (non-deviantly).

What do I mean by a 'response to normativity'? First, for behaviour to count as a response to normativity in the sense I intend, it must be caused by *cognition* of the normativity of some consideration. Roberta's desire to aid the workers is a response to normativity only if it is caused by her awareness of their plight as being a *reason* for her to act. It is not sufficient

[25] Hume (1978: 459). Mele (2003: 19) distinguishes similarly between 'motivation-encompassing' and (merely) 'motivation-providing' attitudes.

that the behaviour merely be caused by a belief or perception. Suppose that I come to believe that there are no custard squares in the kitchen (you've just announced it to me), and that this thought causes me to desire to eat a custard square. We here have a desire caused by a belief, but presumably in coming to my desire I do not see the fact that there are no custard squares in the kitchen as a *reason* for me to eat or to desire to eat one, and my reaction is not a response to the content of my belief as a normative reason.[26] Neither is it sufficient that the behaviour is caused by a belief whose content is in fact a reason, nor even that it be caused by a belief whose content one judges to be a reason, since that judgment itself might not be causally responsible for the behaviour. (I am not claiming that responding to normativity is a condition of behaving as we ought: it usually suffices that we act merely in accordance with our reasons, as we commonly do.)

A response to normativity must be caused by something more like a perception than a judgment or belief. I may *believe* that p is a normative reason for me to ϕ simply because you told me so and I accept your authority. In doing so, I lack something important: an appreciation of the normative character of p. If I then ϕ, this cannot be a response to the normativity of p, but rather to the normativity of the general proposition that one acts on normative reasons, or something of that kind. In order to be able to respond to the normativity of a proposition, one needs to perceive or grasp that proposition as normative. To understand what it is to respond to normativity, therefore, we need to understand the experience of normativity.

We can agree with the contemporary consensus that experiencing the consideration that p as normative, or as a reason to ϕ, is to experience it as 'counting in favour' of ϕ-ing. But the vital feature here is that the experience of normativity is essentially an experience of autonomous authority. 'Autonomy' can mean a lot of different things; here I mean only that normative authority is not alien to the thinking self or self-determining agent. Experiencing something as normative for you forestalls sincere declarations of indifference or skepticism about its practical relevance, of challenging 'So what? What does that matter?'[27] In order to forestall practical challenges like this, the experience of normativity must be an experience of understanding that the counted-in-favour-of action matters, an experience of having the importance of this action explained or made transparent to oneself. (This is not to deny that any particular normative

[26] Arguably the desire is a response to a value I perceive custard squares to have. Ultimately I'll argue against non-DBN versions of this suggestion, but all I need from the example here is that beliefs can cause desires without being seen as reasons for them.

[27] Korsgaard (1996: 9); Joyce (2001: 81).

consideration may be overridden by considerations that seem to matter *more*.) These observations on the character of normative experience are familiar from Kantians' arguments, and seem obvious enough that I shall not argue for them further.

Since the experience of normativity is that of autonomous authority, of behaviour being required of us as self-determining agents, the proper character of response to normativity must be that of voluntary behaviour (or 'action'), as premiss **VB-RN** claims. When we respond to normativity we voluntarily initiate our behaviour because we recognize that our behaving thus matters. The relevant response to normativity is therefore voluntarily to initiate or choose some course of action for that reason.[28] This is implicit in the platitude that normativity is a *guide*: that is, it provides counsel that we are psychologically free either to heed or flout.

To qualify as a response to normativity, behaviour therefore must be voluntary, and it must be appropriately caused by the experience of normativity. However there is potential tension between these requirements;[29] in order for behaviour to be voluntary, it must be caused in the right way (non-deviantly). Much behaviour is not voluntary: salivating at the prospect of food, wincing at pain, etc. We must address the conditions for voluntary action; the difference between what we do by willing it, and what we do without willing it (or what merely happens to us).[30]

4. THE CONDITIONS FOR VOLUNTARINESS

Voluntary behaviour is 'self-initiated' behaviour. So what exactly is this 'self' whose activity is suitably voluntary? The experience of normativity provides the answer: in order for authority to be autonomous it must come from the same entity or faculty that poses the 'So what?' challenge to demands—hence the thinking self or our thought processes themselves. Kantians have typically rejected DBN on this basis: the operation of the will consists in the free exercise of practical *reason*, while desiring is not a free action of thought. But if my claims about the concept of desire are correct,

[28] See also Audi (2002).

[29] This is Kant's problem of how there can be a law of freedom. Kant and other libertarians maintain that free actions must be without causes, but I shall assume without argument here that this anticausalist view of freedom is a non-starter.

[30] This way of drawing the contrast may be infelicitous. We are usually content to describe many of our non-voluntary behaviours as things that we 'do'. (See Hieronymi 2006; Hieronymi unpublished for an argument that we have non-voluntary control over our beliefs and intentions.) We don't generally find our behaviour alien unless it is *in*voluntary, or *contrary* to our will, but this is not what is at issue here.

this objection to DBN is mistaken. Desiring is an activity constituted by thought and so not thereby disqualified. 'Passion' without 'reason' is not merely blind—it is oxymoronic.

Causation by mental processes or events is a necessary condition for voluntary behaviour, but it is not a sufficient condition—the nature of the causal link from thought to behaviour is also crucial. It is not enough for voluntariness that some behaviour is caused by the experience of normativity. Suppose that whenever you saw that you had a reason to scratch, this directly or without identifiable intermediary volitions caused you to blink. This reaction would not be voluntary or self-activated any more than is the reflex to kick when your knee is tapped. We should not suppose matters to be any different if the causal effect was rather to make you scratch.[31] To be voluntary, your behaviour must constitute an activity you effect as an agent, and we need to identify the other conditions necessary for this. Indeed the point even extends to our thought behaviour. Not all of it is thinking that we actively do; some of it consists in thoughts that just 'strike' us.

Premiss **DB-VB** claims that all voluntary behaviour is caused non-deviantly by desire (i.e. it is motivated). Given my account of desire, this is just to say that all voluntary behaviour is teleologically caused, the intentional result of mentally aiming at some end. But why should anyone accept this premiss? I submit that intuitively this just is the essential difference between the behaviour, both physical and mental, that comes upon us (or that we do without volition), and that which we actively and voluntarily perform. If some behaviour *B* is not an intentional result of my aiming at it (or of my *trying* to produce it)[32]—directing my motions and thoughts in the ways I think will or may lead to *B*—then I cannot recognize it as something that I do voluntarily.

Some philosophers are skeptical that intellectual activity and epistemic responses to normativity, at least, require desire, and point to inferential processes as instances of agential responses to normativity that do not depend on desire.[33] Presented with a valid deductive argument, for example, with premisses that I accept, I am presented with a (subjective) reason to believe the conclusion. If I am rational I will respond to this reason by forming this belief (or by reconsidering my acceptance of the premisses). But no *desire*

[31] Nagel (1970: 34); Davidson (1980b: 78–9).

[32] McCann (1974) argues in a similar vein that the essential mental component of action is *trying*. Trying to ϕ entails aiming to ϕ. I disagree, however, with McCann's claim that trying is a *spontaneous* mental action.

[33] Wallace (1990) and Smith (1987–8) advance this possibility as the chink in the Humean's armor. I thank Michael Smith for pushing these concerns against me in discussion.

is required to explain my recognition of a reason to believe the conclusion or my forming that belief. Therefore we might justifiably suspect that there could similarly be *practical* inferences, with actions as conclusions, that do not depend on desires.

In response, it is first important to distinguish evidence or 'reasons for belief' from reasons to *form* beliefs. A justified belief always constitutes evidence for its logical consequences but does not always provide a reason (i.e. normative pressure) to form such beliefs, simply because many of the logical consequences of our beliefs are utterly trivial.[34] If (*A*) Los Angeles is in California, then (B) either Los Angeles is in California or the moon is made of blue cheese. But my believing proposition *A* gives me no reason by itself to *form* the belief that B. So the existence of normative reasons to form certain beliefs is conditional on more than simply logical or evidentiary relations. My perceiving myself to have such a reason, and my being motivated to form the belief, I argue, depends upon my intellection being motivated by some desire such as the desire to know about subject *X*, or (more accurately) to settle whether something is the case.[35]

It will be objected that my drawing the inference does not depend on any such desire. I (qualifiedly) concede this. But (i) we typically 'draw' inferences automatically. We are disposed to form beliefs non-voluntarily on the basis of evidence. There is no need to think that this belief formation is a response to normativity in the relevant sense. We might note, further, that epistemically we are unresponsive to at least some kinds of normativity: the perception that we have good practical reasons to form some beliefs is notoriously impotent in producing them. Furthermore, (ii) there is a difference between voluntarily *drawing* an inference, and an inference just striking you.[36] Only the former case is a voluntary action of thought and a response to normativity. I maintain that the difference between voluntarily and non-voluntarily forming an inferentially derived belief is precisely that in the former but not in the latter case one is *aiming* at drawing the correct inference—hence that one draws the inference as a result of desiring some (typically epistemic) end. If this is correct, then inference does not constitute a form of voluntary action or response to normativity that rivals motivation or causation by desire.

[34] I owe this point to Aaron James.

[35] As Pamela Hieronymi points out to me, we typically form beliefs without attending to our own mental states.

[36] Again, I have to acknowledge that some inferences (especially those that come with expertise) are non-passive, despite being non-voluntary. I also maintain that they are not responses to normativity in the privileged sense.

5. SQUARING OFF OVER ROBERTA

The significance of the argument so far for anti-DBN, cognitivist theories of motivation will not be immediately obvious, and therefore it is time to square off over Roberta. We can grant (i) that there is nothing that she (henceforth R) desires prior to or concurrent with forming her belief about the plight of the textile workers that is relevant to explaining her motivation (her pre-existing dispositions being causally irrelevant or trivial), (ii) that she perceives the fact that *p* (the workers are exploited) to be a reason for her to *A* (assist in the reform efforts), and (iii) that she is motivated to *A* by that perception of a reason. There remain at least two rival models of the causal process. On the anti-DBN model, R's perception of the consideration that *p* as a reason for her to *A* causes all her relevant desires and motivations. On the DBN model, R's belief that *p* causes a desire (an episode of desiring), under the influence or from the perspective of which R experiences the consideration that *p* as a reason to *A*, and by which she is then motivated accordingly.

The anti-DBN model fails because it portrays the response to normativity as passive and non-voluntary.[37] This will not be obvious. If R responds to her perception that the consideration that *p* is a reason by attempting to *A*, her attempting to *A* is teleologically caused by her desiring (let us suppose) to *A*, which I have conceded to be sufficient for volition and agency. However my concern lies elsewhere in the causal story. R's *immediate* response to her perception of a reason, on this model, is not to *A* but to form a desire (commence end-directed practical thought). The desire indeed produces voluntary action, but this action is only a voluntary *response to the reason* if its volitional and agential causation reaches back beyond the desire to the perception of the reason. That is to say, R voluntarily *As for the reason* that *p* only if R's motivation to *A* arises voluntarily from that perception of normativity. The formation of R's desire, therefore, must be a voluntarily chosen activity: R must voluntarily set herself to aim at *A*-ing.

This spells trouble for the anti-DBN model. It follows from my argument that the formation of desire (adoption of a new end) can only be voluntary if it is motivated by some further desire. In order to maintain an agential or volitional link to the perception of the reason, it would have to be the case that R was desiring something concurrently with her recognition of the consideration that *p* as a reason that motivated her to form her desire

[37] I also think that it has no plausible account to give of the perception that something is a reason, although this calls for a different argument.

that motivates her to act (overtly). But this is to concede to **DB-RN**, my conclusion. In order for a new end to be generated voluntarily from the perception of a reason, it must be motivated by a desire that is not itself generated from that perception.

I adduce two further problems that arise here for the anti-DBN model. First, it is independently implausible that we voluntarily initiate our desires (i.e. we are unable to 'desire at will'), as has been widely observed,[38] and hence it is implausible that R voluntarily comes to desire to *A*. This result is accommodated by my argument: voluntary actions all proceed from desiring some end, but the causal processes by which desires themselves are initiated are merely causal and not teleological. Even when desires are stimulated by thought, the process of their generation from those thoughts occurs below the level of even unconscious thought.

The other problem emerges from consideration of which desires play a role in the process. There are two versions of the anti-DBN model, corresponding to the two desires we might reasonably expect Roberta to acquire. R's perception of her reason to *A* might cause her (a) to desire to *A* (an action desire), or (b) to desire that not-*p* (that it's not the case that the workers are exploited; a state desire).[39] Note that given my account of desire, to say that the effect is simply motivation to *A* for its own sake is just option (a) again. Which of these desires does she possess, and which constitutes her immediate response to normativity? Suppose first that her response to normativity is constituted by her coming (a) to desire to *A*. It seems implausible (i) that she would come to desire to *A* (assist the reform efforts) without also coming to desire that not-*p* (the workers not be exploited), and (ii) that her desiring to assist the reform efforts would not then be derived somehow from her desiring that the workers not be exploited.[40] Suppose instead, therefore, that her direct response to normativity is to come to desire that not-*p*, which motivates her to *A*. The

[38] Hutcheson (1969: 139); Stampe (1987: 370); Millgram (1997: 11). Might the intrinsic-teleological account of desire favoured here provide an explanation of this inability? We might reason that if desiring an end is aiming at it for no further end, and if initiating some behaviour voluntarily is to initiate it by aiming at some end, then to initiate desiring at will would be to initiate aiming at an end for no further end for some end, which we might think to be self-contradictory and hence impossible. In this form, however, the argument doesn't work: what is done for some further end is the action of *initiating* behaviour (of aiming at an end for no further end), not the action of aiming at an end for no further end.

[39] I am assuming that Roberta's motivation is altruistic rather than dutiful: an alternative scenario has her belief stimulating rather the desire that she not shirk her moral duty. A similar objection would still apply.

[40] Here I'm granting the possibility of derived desires, for the sake of argument. My preferred interpretation of the scenario is rather that R's desiring that not-*p* itself motivates her to *A*, obviating any need for a desire to *A*.

problem here is that this does not seem to be the appropriate response to the reason she perceives herself to have. Intuitively and as the case has so far been described, R perceives *p* to be a reason in the first instance to *A*; she sees the fact of the textile workers' exploitation as a reason for her to assist in reform efforts, and not as a reason to form a desire that the textile workers are not exploited. Forming the desire that not-*p* would thus appear an inappropriate response to her perception of her reason. Neither horn of the dilemma facing the anti-DBN model looks comfortable, because the model fails to find a satisfactory fit between the reason and the desires it allegedly causes.[41]

The DBN model fares much better under close scrutiny. According to this, R's belief that *p* causes her (being a sympathetic soul) to desire that not-*p*; awareness of exploitation—but not any perception of a reason—prompts a desire that it be eliminated. From this motivated point of view she now experiences the fact that *p* as normative for her; it requires action of her in virtue of her desired end. She thus recognizes the exploitation of the workers to matter from her own point of view, and her immediate response is voluntarily to choose[42] the course of action (*A*-ing) that she judges the reason to count in favour of, motivated by her desire that not-*p*. A *reason* (with normative authority for R) for R to *A*, on this model, is roughly a fact that indicates that *A*-ing might promote some end that R cares about, and its counting in favour of *A* is just its so indicating.[43] As opposed to the rival model, (i) the response to normativity is voluntary, (ii) the relevant desires and motivations are all in place playing appropriate roles in the story, and (iii) the reason counts in favour of what intuitively is the relevant behaviour.

The basic reason the anti-DBN model fails is this: we engage voluntarily and actively in the exercise of our desires but not in their formation, because the activity of our desires though not their formation occurs through teleological thought. The anti-DBN model fails because it identifies response to normativity with the non-voluntary and passive formation of desires, whereas the DBN model succeeds because it identifies response to normativity with the voluntary and active exercise of desires. Ironically, it is the respect in which desiring is non-voluntary that shows us that the voluntary character of response to normativity entails that it is desire-based.

[41] A further option for the anti-DBN model is that the recognition of a reason stimulates something like a desire to act for reasons. This seems unattractively indirect, however.

[42] I assume here that choice is fully compatible with being non-deviantly caused by desiring. Choice, as I see it, is constituted by the interplay of our desires, rather than an act of external arbitration upon them.

[43] Finlay (2006).

6. EXTENDING THE ARGUMENT

The Argument from Voluntary Response provides only qualified support for DBN. Its conclusion, **DB-RN**, maintains only that *response* to normativity is desire-based. This is compatible with the possibility that normativity itself is not desire-based. This would be true, for example, on the popular view of desires as cognitive states that essentially involve representation of their objects as having some 'desirability characteristic' or normative quality;[44] on this view desires rather are normativity-based. However, this reversed dependency is not compatible with the *argument* I have given for **DB-RN**, on which to desire is by definition to engage in intrinsic teleological activity. Regardless of whether my hypothesis about the concept of desire is correct, the point of the argument is that the motivation constitutive of a response to normativity cannot be caused by a normative belief or perception alone, but must derive from antecedent motivation that does not depend upon that belief or perception.

However, there may also be normative beliefs by which we are not motivated, and which therefore need owe nothing to our desires. I concede that we can and do recognize practical reasons that lack even the power to motivate us and that are not connected to our desires. But I maintain and have argued elsewhere (2006) that these are reasons that we do not experience as normative (i.e. as having autonomous authority) for us. I concede, therefore, that there can be reasons for agents to act in certain ways (as well as value and ought-facts) that are not based on those agents' desires, but I maintain that the normative force or authority, or *importance*, of these reasons for any agent is based on that agent's desires. As I construe it, therefore, DBN is a doctrine concerned with importance rather than with practical reasons (or value or 'oughts') per se.

The opponent of DBN could concede that experiencing and responding to considerations as normative depends upon desires, but maintain that this merely has to do with the *appearance* of importance, which is oftentimes an illusory appearance. Importance can outstrip our awareness of it, after all, so arguably *it* is an objective desirability characteristic tracked by our desires. The reason why this objection fails, and why **DB-RN** does support DBN, I would argue, is that the experience of normative authority ('finding something important') does not even purport to represent some

[44] Anscombe (1957); Davidson (1980a); Stampe (1987); Millgram (1997); Scanlon (1998); Raz (1999); Hurley (2001); Darwall (2001).

independent facts about (*intrinsic*[45]) importance; rather it involves having something matter to you. (The perception involved in the experience of normativity is the perception that something is a *reason*: experiencing that reason as important rather involves its mattering to you.) What is important for you—as opposed to what is important *to* you—outstrips both your awareness and your occurrent desires, it is true. But I would argue that this is a consequence of the fact that the concept of a person is the concept of a temporally (and even counterfactually) extended being; 'You' are more than your present mental activities, and it is because of this that what is important for 'you' outstrips what you desire and what is important *to* you at any moment. This complexity in the ontology of persons would yield objectivity in the concept of importance for persons that remains grounded in (actual, future, and counterfactual) desires.

There remains much work yet to do before I can claim that my Argument from Voluntary Response proves that DBN is true and normativity depends upon desire. The presentation of the argument itself here is unavoidably sketchy in many places, requiring in particular a much more scrupulous investigation of the nature of the voluntary than I have provided. And my closing suggestions on how the gap between **DB-RN** and DBN might be closed provide only a promissory note in need of redemption. But I hope to have introduced a new argument supporting DBN that deserves further development and consideration.

REFERENCES

Anscombe, G. E. M. (1957) *Intention* (Oxford: Blackwell).

Audi, Robert (2002) 'Prospects for a Naturalization of Practical Reason: Humean Instrumentalism and the Normative Authority of Desire,' *International Journal of Philosophical Studies* 10: 235–63.

Chan, David (2004) 'Are there Extrinsic Desires?' *Noûs* 38: 326–50.

Cohon, Rachel (1988) 'Hume and Humeanism in Ethics,' *Pacific Philosophical Quarterly* 69: 99–116.

Dancy, Jonathan (2000) *Practical Reality* (New York: Oxford University Press).

Darwall, Stephen (1983) *Impartial Reason* (Ithaca, NY: Cornell University Press).

____ (1992) 'Internalism and Agency,' *Philosophical Perspectives* 6: 155–74.

____ (2001) ' "Because I Want It",' *Social Philosophy and Policy* 18: 129–53.

Davidson, Donald (1980a) [1963] 'Actions, Reasons, and Causes,' reprinted in *Essays on Actions and Events* (New York: Oxford University Press).

[45] Finding something derivatively or instrumentally important does, however, have a cognitive element. It involves the representation that something stands in a promotive relation with something else that matters intrinsically to you.

Davidson, Donald (1980b) [1973] 'Freedom to Act,' reprinted in *Essays on Actions and Events* (New York: Oxford University Press).
Finlay, Stephen (2006) 'The Reasons that Matter,' *Australasian Journal of Philosophy* 84: 1–20.
____(forthcoming) 'Motivation to the Means,' in David Chan (ed.), *Values, Rational Choice, and the Will.*
Heuer, Ulrike (2004) 'Reasons for Actions and Desires,' *Philosophical Studies* 121: 43–63.
Hieronymi, Pamela (2006) 'Controlling Attitudes,' *Pacific Philosophical Quarterly* 87: 45–74.
____(unpublished) 'Responsibility for Belief.'
Hume, David (1978) [1739] *A Treatise of Human Nature*. ed. L. A. Selby-Bigge, rev. P. H. Nidditch (Oxford: Clarendon Press).
Hurley, Paul (2001) 'A Kantian Rationale for Desire-Based Justification,' *Philosophers' Imprint* 1 (2): **www.philosophersimprint.org/001002/**
Hutcheson, Francis (1969) [1738] *An Inquiry into the Original of our Ideas of Beauty and Virtue* (Farnborough: Gregg International Publishers).
Joyce, Richard (2001) *The Myth of Morality* (Cambridge: Cambridge University Press).
Korsgaard, Christine (1996) *The Sources of Normativity* (Cambridge: Cambridge University Press).
Lenman, James (1996) 'Belief, Desire and Motivation: An Essay in Quasi-Hydraulics,' *American Philosophical Quarterly* 33: 291–301.
McCann, Hugh (1974) 'Volition and Basic Action,' *Philosophical Review* 83 (4): 451–73.
Mele, Alfred (2003) *Motivation and Agency* (New York: Oxford University Press).
Millgram, Elijah (1997) *Practical Induction* (Cambridge, Mass.: Harvard University Press).
Nagel, Thomas (1970) *The Possibility of Altruism* (Princeton, NJ: Princeton University Press).
____(1986) *The View from Nowhere* (New York: Oxford University Press).
Platts, Mark (1979) *Ways of Meaning* (London: Routledge & Kegan Paul).
Raz, Joseph (1999) 'Explaining Normativity: On Rationality and the Justification of Reason,' *Ratio* 12: 354–79.
Ross, Peter W. (2002) 'Explaining Motivated Desires,' *Topoi* 21: 199–207.
Scanlon, T. M. (1998) *What We Owe to Each Other* (Cambridge, Mass.: Harvard University Press).
Schroeder, Timothy (2004) *Three Faces of Desire* (New York: Oxford University Press).
Schueler, G. F. (1995) *Desire: Its Role in Practical Reason and the Explanation of Action* (Cambridge, Mass.: MIT Press).
Smith, Michael (1994) *The Moral Problem* (Oxford: Blackwell).
____ (1987–8) 'Reason and Desire,' *Proceedings of the Aristotelian Society* 88: 243–58.
Stalnaker, Robert (1984) *Inquiry* (Cambridge, Mass.: Harvard University Press).
Stampe, Dennis W. (1987) 'The Authority of Desire,' *Philosophical Review* 96: 335–81.

Velleman, J. David (2000) *The Possibility of Practical Reason* (Oxford: Clarendon Press).
Wallace, R. Jay (1990) 'How to Argue about Practical Reason', *Mind* 99: 355–85.
Wedgwood, Ralph (2002) 'Practical Reason and Desire,' *Australasian Journal of Philosophy* 80: 345–58.
Williams, Bernard (1981) [1980] 'Internal and external reasons,' reprinted in *Moral Luck* (Cambridge: Cambridge University Press).

11

Normativity

Judith Jarvis Thomson

1. What are commonly called normative judgments fall, intuitively, into two classes. First, there are what I will call evaluatives: these include judgments to the effect that a certain state of affairs would be good or bad for Smith, or for England, judgments to the effect that a certain experience was delightful or dreadful, judgments to the effect that a certain carving knife is a well-made or defective carving knife, and so on. The evaluatives also include comparatives, such as that this is or would be a better so and so than that. Second, there are what I will call directives: these include judgments to the effect that something or someone ought or ought not do this or that—as, for example, that Jones ought to be kind to his little brother.

Intuitively, the two classes of normative judgment interconnect. What is good or bad on the one hand must surely link with what ought or ought not be done on the other hand. What I will focus on is the question how they link.

Consequentialism is the most familiar theory about how they link. We can take the Consequentialist to say two things. (i) It is a necessary truth that a person ought to do a thing at a time just in case the world will be better if he does it than if he does any of the other things it is open to him to do at the time, and (ii) when a person ought to do a thing at a time that is *because* the world will be better if he does it than if he does any of the other things it is open to him to do at the time.

I stress that the Consequentialist says (ii) as well as (i). He does not merely offer us a necessarily true biconditional: he says it is the truth of the appropriate evaluatives that make directives true. That is worth stress

The following is a draft of parts of Chapters VIII and IX of a work in progress entitled *Normativity*. Parts of it were presented as the Howison Lecture for 2005 at the University of California at Berkeley; other parts were presented at the Metaethics Conference at the University of Wisconsin at Madison. I thank those who attended for their comments. I also thank Paul Bloomfield and a referee for Oxford University Press for telling me that a theory in some ways like my own appeared in Bloomfield (2001).

since there is room for a theory according to which it is the truth of the appropriate directives that make evaluatives true. For example, there is a theory according to which it is the truth of certain directives that make ascriptions of intrinsic goodness true, one possible world being better than another just in case it contains more intrinsic goodness than the other.[1]

That second kind of theory has not been found attractive by many people, however, and it is easy to see why. For such a theory to be true it is required that there be truths of the form "*A* ought to *V*" that are not made true by any facts about what is or would be good or bad. That is metaphysically implausible. For such a theory to be true it is required that there be truths of the form "*A* ought to *V*" that we can find out the truth of without finding out facts about what is or would be good or bad. That is epistemologically implausible.

In sum, directives do seem to call for justification by appeal to the truth of evaluatives.

On the other hand, very few people regard Consequentialism as acceptable. What I will do is to point to the possibility of an alternative, a theory that is like Consequentialism in that it says that what makes directives true, when they are, is the truth of certain evaluatives; but the evaluatives it focuses on are different from those that the Consequentialist focuses on. I will suggest that although what the world will be like if a person does or does not do a thing has a bearing on whether the person ought to do the thing, it is not how good the world will or will not be if he does the thing that fixes whether he ought to do it.

2. I begin with three preliminary remarks about the directives.

(i) Some people have claimed that the word "ought" is in many ways ambiguous. Thus that it has a chess meaning such that people who say "Smith ought to move his rook" are likely to mean by it "Smith ought$_{\text{chess}}$ to move his rook," and a moral meaning such that people who say "Jones ought to be kind to his little brother" are likely to mean by it "Jones ought$_{\text{morality}}$ to be kind to his little brother"—and indeed, a medical meaning such that when your doctor says to you "You ought to get more exercise," what he is likely to mean is "You ought$_{\text{health}}$ to get more exercise," and a Wall Street meaning such that when your investment adviser says to you "You ought to diversify your portfolio," what he is likely to mean is "You ought$_{\text{Wall Street}}$ to diversify your portfolio." Why might a person think that?

[1] Michael J. Zimmerman says that for *x*, *y*, *z*, and so on, to possess intrinsic goodness is for it to be the case that "*there is a moral requirement to favor them* (welcome them, admire them, take satisfaction in them, and so on) *for their own sakes*" (his italics); see Zimmerman (2001: 24).

Suppose that Alfred can see that Smith will be checkmated in three unless he moves his rook. Then Alfred may say "Smith ought to move his rook." Suppose also that Bert can see that if Smith moves his rook, then that will cause hundreds of people to die. Then Bert may say "Smith ought to refrain from moving his rook." Considerations of chess strategy seem to yield that Smith ought to move his rook. Moral considerations seem to yield that Smith ought to refrain from moving his rook. Which ought Smith do? The people I have in mind think that there is nothing that fixes which he ought to do. And they conclude that that is because there is no such question as the question which he ought to do. There is a question whether Smith $\text{ought}_{\text{chess}}$ to move his rook (the answer to which is Yes), and a question whether Smith $\text{ought}_{\text{moral}}$ to refrain from moving his rook (the answer to which is Yes), and no further question whether Smith ought to move his rook or to refrain from moving his rook.

It is a bad argument. If those 'moral considerations' are true, and yield that Smith ought to refrain from moving his rook, then Smith ought to refrain from moving his rook. What follows from those true moral considerations is not: Smith $\text{ought}_{\text{moral}}$ to refrain from moving his rook. A morality doesn't tell you only that morality tells you this or that; it tells you what you ought to do. Period. And if it is a true morality, then that *is* what you ought to do.

By contrast, books on chess strategy don't tell you what you ought to do. They tell you what will make it more likely that you will win, but whether you ought to do what will make it more likely that you will win is quite another matter. No author of a book on chess strategy believes, and therefore tells his readers, that when you will be checkmated in three unless you move your rook, then you ought to move your rook come what may, thus no matter how many deaths you will thereby cause. It may be that nothing will go wrong if you move your rook; it may indeed be the case that you ought to move it. But what settles that you ought to is not *just* the fact that you will win only if you do—and no author of a book on chess strategy, or sensible reader of such a book, believes that it does.

I take it to be clear that there is such a question as the question which Smith ought to do, move his rook or refrain from moving his rook. And moreover that, given that Smith will cause the deaths of hundreds of people if he moves his rook, then it is clear that the answer is, simply, that he ought to refrain from doing so.[2]

[2] A further consideration is worth noting. Whenever a philosopher says about a philosophically important term that it is ambiguous, it is *always* called for that we ask whether he means that it is what we might call strongly (or happenstance) ambiguous, as "bank" and "bat" are, or what we might call weakly ambiguous, as "healthy" is. (As

(ii) In (i), I rejected an argument for the view that "ought" is many ways ambiguous, but we really do need to grant that it is two ways ambiguous. Consider the sentence "Jones ought to pass by us soon." Typically, I suppose, a person who says those words means that Jones is *called on* or *supposed to* pass by us soon, as Jones would be if he had promised to pass by us soon. But suppose that we are looking out of a window on the fifth floor of the Empire State Building. We believe that Jones has decided to throw himself off the roof at 4 p.m., and it is 4 p.m. now, so I say to you "Jones ought to pass by us soon." When I do, I mean that Jones is likely to pass by us soon—a very different affair.

Let us be clear that there really is an ambiguity here. Suppose that Alfred thinks that Jones has promised to pass by us soon, and therefore says "Jones ought to pass by us soon," meaning that Jones is called on or supposed to pass by us soon; and suppose that Bert thinks that it is 4p.m., and that Jones intends to wait till 6p.m. before jumping off the roof, and therefore says "It's not the case that Jones ought to pass by us soon," meaning that it is not the case that Jones is likely to pass by us soon. There is no further question who is right, Alfred or Bert. Alfred and Bert do not contradict each other: both could be speaking truly, both could be speaking falsely.

Let us call the two meanings of "ought" its normative meaning and its probability meaning. A sentence of the form "*X* ought to *V*" understood normatively says roughly that *X* is called on or supposed to *V*. A sentence of that form understood probabilistically says roughly that *X* is likely to *V*.[3] And let us say that anyone who says a sentence of the form "*X* ought to *V*," meaning "ought" normatively, makes a normative judgment, and anyone who says a sentence of that form, meaning "ought" probabilistically, makes a probability judgment. What I referred to as the class of directives

Aristotle pointed out, to say of a foodstuff that it is healthy is not to ascribe to it what one ascribes to a person when one says that the person is healthy. But for a foodstuff to be healthy is for it to be conducive to being healthy in those who eat it.) When people say that "ought" is many ways ambiguous, which do they mean? It is the height of implausibility to say that "ought" is many ways strongly ambiguous. What, then, would they have us take to be the relations among "$\text{ought}_{\text{chess}}$", "$\text{ought}_{\text{morality}}$", "$\text{ought}_{\text{health}}$", and "$\text{ought}_{\text{Wall Street}}$"?

On some views, we do better to say, not that "ought" is ambiguous, but rather that it is 'incomplete': very roughly, "A ought to V" is true only relative to a body of rules—as, for example, the rules of chess, the rules of morality, and so on. I do not take space to discuss this kind of view here. I draw attention only to the fact that on these views, the following outcome is the same as on the 'ambiguity view': once the information is in about the rules of chess and morality, there is no question remaining as to whether Smith ought to move his rook, all things considered. As I said in the text above, I take that to be clearly wrong.

[3] And is this strong or weak ambiguity? For a suggestion, see footnote 10 below.

includes only normative judgments—thus it excludes those judgments in the making of which the judger says "*X* ought to *V*," meaning "ought" probabilistically.

From here on I will throughout mean "ought" normatively unless I explicitly say otherwise.

(iii) We should notice, finally, that there are true directives that are not about people. A toaster ought to toast toastables—bread, bagels, frozen waffles, and the like; I'll just say bread, for short. A valve of a certain kind ought to blow when the pressure in the pipe it is installed in reaches so and so many degrees. A seeing eye dog ought to stop its master at street corners. The pancreas ought to secrete digestive enzymes. I don't mean that a toaster, a valve of that kind, a seeing eye dog, a pancreas is likely to do these things, though that may well be true. I mean rather that they are called on or supposed to. My judgments about them are normative, not probabilistic.

This third remark about the directives is of great importance. I suggest that it is precisely by virtue of what we learn when we attend to directives that are about non-human things that we can best understand all of the directives, and thus those that are about people as well.

3. Let us begin with artifacts. Toasters, for example. I said: a toaster ought to toast bread. So let *A* be a toaster. Then the following is true:

(1) *A* ought to toast bread.

What makes (1) true? Two facts. First, the fact that A is by hypothesis a toaster. Second, the fact that toasters are manufactured *to* toast bread. Since toasters are manufactured *to* toast bread, a toaster that comes off the assembly line but does not toast bread is a toaster that does not do what it is manufactured to do, and therefore is a defective toaster. It is *that* that makes (1) true.

To forestall an objection, I should perhaps stress that the words I wrote in writing (1) have to be understood as an abbreviation. A toaster is a defective toaster if it doesn't toast bread, but not in just any circumstances. A toaster is marked as defective when it fails to toast bread only if it has been plugged in, the bread was inserted in the slots, the bar was depressed, and you aren't sitting in the bathtub while doing all of that. A toaster is marked as defective when it fails to toast bread only if it fails to toast bread in suitable circumstances. So what a toaster ought to do is only to toast bread in those suitable circumstances. I won't try to spell out what all those suitable circumstances are; I merely abbreviate when I say that a toaster ought to toast bread. I will be helping myself to similar abbreviations throughout what follows.

Consideration of (1) suggests an idea, namely that the directives are kind-dependent: that what marks a directive as true of a thing turns on what kinds the thing is a member of—indeed, that it turns on what kinds the thing would be a *defective* member of if the thing does not do what the directive says it ought to do. Let us say that a kind *K* is a directive-generating kind—a directive kind, for short—just in case there is such a property as being a defective *K*. Then consideration of (1) suggests the following idea:

(First Candidate Thesis) For it to be the case that *X* ought to *V* is for it to be the case that there is a directive kind *K* such that *X* is a *K*, and if *X* does not *V*, then *X* is a defective *K*.[4]

There are no such properties as being a defective pebble and being a defective piece of wood; therefore the kinds pebble and piece of wood are not directive kinds. That leaves room for the possibility that a given pebble or piece of wood ought to do such and such, for it leaves open the possibility that the thing is a member of some directive kind *K* such that it is a defective member of *K* if it does not do the such and such. But as we might put it: there is nothing that it ought to do *qua*, or just in virtue of, being a pebble or piece of wood.

Seeing eye dogs are not artifacts: they are not manufactured to do things, they are instead trained to do things—in particular, to serve as eyes for the blind. Suppose *A* is a seeing eye dog. Then

(2) A ought to stop its master at street corners

is true. What makes (2) true? There is such a property as being a defective seeing eye dog, so the kind seeing eye dog is a directive kind; and *A* is a member of it, and if *A* does not stop its master at street corners, then it is a defective member of it.

The kinds toaster and seeing eye dog are function-kinds. That is, there is a function associated with each of those two kinds which is such that it is a member's failing to carry out that function, or carrying it out badly, that *marks* it as a defective member of the kind. The function in the case of the kind toaster is to toast bread; the function in the case of the kind seeing eye dog is to serve as eyes for the blind. Among the functions of the pancreas

[4] Nicholas Wolterstorff says: "Many, though not all, kinds are such that it is possible for them to have properly formed and also possible for them to have improperly formed examples. Let us call such kinds, *norm*-kinds." (See Wolterstorff 1980: 56.) The kind lion, then, is, as he says, a *norm*-kind. The notion 'directive kind' that I defined in the text above is more general, since while the kind lion is a directive kind (since it is possible for a thing to be a defective lion), so also is the kind tennis player (since it is possible for a thing to be a defective tennis player), but I take it that the kind tennis player is not a '*norm*-kind'.

is to secrete digestive enzymes, and an instance that fails to do that, or does it badly, is thereby marked as a defective instance. It is a disputed issue in the philosophy of biology just what it is that gives the pancreas and other human organs the functions they do have, that is, whether it is evolution, or the role they currently play in the bodily economy, or both; I leave aside the question what should be said about that issue. Whatever explains why the pancreas has the function it does, it nevertheless does have the function of secreting digestive enzymes. So if *A* is your pancreas, then *A* is a member of a function-kind, and therefore of a directive kind, such that if *A* does not secrete digestive enzymes, then *A* is a defective member of it, and

(3) *A* ought to secrete digestive enzymes

is therefore true.

But we should notice that the directive kinds are not limited to the function-kinds. Beefsteak tomatoes are bred to be big and fat at maturity, and if a particular beefsteak tomato turns out to be little at maturity—perhaps because of some freak in the weather—then it is a defective beefsteak tomato. But being big and fat at maturity isn't a function of a beefsteak tomato—it is just a feature such that if a beefsteak tomato lacks the feature, then it is a defective beefsteak tomato. So though the kind beefsteak tomato isn't a function-kind, it is all the same a directive kind; and we can say that if "*A*" is the name of a beefsteak tomato, then

(4) *A* ought to be a big, fat tomato at maturity

is also true.

So far, so plausible, I hope.

4. Alas, it won't do. Consider the kind jewel thief. It is a function-kind, since the function of a jewel thief is to steal jewelry, and a jewel thief who can't tell good jewelry from junk, and who therefore steals the junk instead of the good jewelry, is a defective jewel thief. It follows that the kind jewel thief is a directive kind. Then let A be a jewel thief. If

(First Candidate Thesis) For it to be the case that *A* ought to *V* is for it to be the case that there is a directive kind *K* such that *A* is a *K*, and if *A* does not *V*, then *A* is a defective *K*

were true, then it would follow that

(5) *A* ought to steal good jewelry

was true. But we had really better not opt for a theory that yields *that* outcome.

I hope it won't strike anyone to say "Well, there's an ambiguity here. In the jewel thief meaning of 'ought', *A* ought to steal good jewelry. Though

of course in the moral meaning of 'ought', *A* ought not steal good jewelry." There is no such ambiguity.

We needn't have climbed all the way up from toasters to people to find ourselves in trouble. Let us imagine that a breeder develops a dog-kind that he advertises as Studio Apartment Dogs: these are dogs bred to be especially suited to masters who live in small apartments—they are bred to be small and obedient, and they are operated on early to cut their vocal cords.[5] If the operation fails in the case of a particular Studio Apartment Dog, so that it remains capable of barking, then the result is a defective Studio Apartment Dog. Then let *A* be a Studio Apartment Dog. If the First Candidate Thesis were true, then

(6) *A* ought to be unable to bark

would be true. But that is surely implausible. Surely it is not true of any dog that it ought to be unable to bark. A dog that is unable to bark is a defective dog![6]

A revision all but suggests itself. Let us say:

(Second Candidate Thesis) For it to be the case that *A* ought to *V* is for it to be the case that there is a directive kind *K* such that *A* is a *K*, and

- (α) if *A* does not *V*, then *A* is a defective *K*, and
- (β) there is no directive kind *K*+ such that *K* is a sub- kind of *K*+, and such that *A* is a defective *K*+ if it does *V*.

(I take *K* to be a sub-kind of *K*+ just in case necessarily, every *K* is a *K*+.) That provides the needed ground for saying that if *A* is a jewel thief, then (5) is false, namely: the kind jewel thief is a sub-kind of the kind person, and although *A* is a defective jewel thief if he does not steal good jewelry, he is a defective person if he does. Similarly for (6) if *A* is a Studio Apartment Dog: the kind Studio Apartment Dog is a sub-kind of the kind dog, and although *A* is a defective Studio Apartment Dog if it is able to bark, it is a defective dog if it is unable to.

At the same time, the Second Candidate Thesis makes no trouble for the supposition that if *A* is a toaster, then

(1) *A* ought to toast bread

[5] It was with considerable surprise that I learned recently that some people do in fact have this done to their dogs. By contrast, it is a familiar enough fact that many people have their cats declawed in order to keep them from shredding the furniture.

[6] A participant at the Wisconsin conference told me that there is a species of the genus dog for which it is normal and natural—thus non-defective and non-damaged—to be incapable of barking. If that is true, readers are invited to substitute for "dog", throughout, the name of some species of dog for which it is normal and natural to be capable of barking. (Terrier? Dachshund?)

is true. Suppose that *A is* a toaster. The kind toaster is a directive kind. So there is a directive kind *K* such that *A* is a member of *K*, and such that (α) if *A* does not toast bread, then *A* is a defective member of the kind. So clause (α) is true. So far so good. What about clause (β)? It is very plausible to think that there is no directive kind *K*+ such that the kind toaster is a sub-kind of *K*+, and such that if *A does* toast bread, then it is a defective *K*+. If that is right, then that would explain why (1) is true.

Similarly, I should think, for judgments (2), (3), and (4). Note in particular the contrast between (6) and (2): while a dog that is unable to bark is a defective dog, it is not true that a dog that stops its master at street corners is a defective dog.

5. We should stop for a moment to take note of a question that might well arise here. Consider

(5) *A* ought to steal good jewelry

again, and suppose that *A* is a jewel thief. We know that *A* is a defective jewel thief if he doesn't steal good jewelry. According to the Second Candidate Thesis, that fact does not fix that *A* ought to steal good jewelry since there is a directive kind person such that jewel thief is a sub-kind of it, and such that *A* is a defective member of it if he does steal good jewelry. In short, the fact that *A* is a defective jewel thief if he does not steal good jewelry is trumped by the fact that *A* is a defective person if he does. In shorter still, the directive super-kind person trumps its conflicting directive sub-kind jewel thief.

Indeed, we might rewrite our thesis more briefly as follows:

(Second Candidate Thesis, abbreviation) For it to be the case that *A* ought to *V* is for it to be the case that there is a directive kind *K* such that *A* is a *K*, and

(α) if *A* does not *V*, then *A* is a defective *K*, and

(β) *K* is not trumped by any directive super-kind.

It might well be asked, however, why we should think that true. Why not instead suppose that what fixes what *A* ought to do is not what issues from the fact that *A* is a member of the super-kind person, but rather what issues from the fact that *A* is a member of the sub-kind jewel thief?

That the super-kind person trumps the sub-kind jewel thief in the way the Second Candidate Thesis says it does is intuitively plausible. That a directive super-kind *K*+ trumps a directive sub-kind *K* in that way is intuitively plausible quite generally. When we reason about what a thing ought to do, we look for generalizations, and we take what issues from the more general to have more weight than what issues from the less general

if what issues from the more general conflicts with what issues from the less general.

Why so? Appeal to the concept 'defect' supplies a justification. Consider *A*, the jewel thief. He is a defective jewel thief if he doesn't steal good jewelry, and a defective person if he does. The fact that he is a jewel thief guarantees that he is in some way defective. Not so the fact that he is a person. *That* is why what fixes what he ought to do is not the fact that he is a jewel thief but instead the fact that he is a person.

Again, suppose *A* is a Studio Apartment Dog. If the operation on *A* failed, so that *A* is able to bark, then *A* is a defective Studio Apartment Dog. If the operation on *A* did not fail, so that *A* is unable to bark, then *A* is a defective dog. The fact that *A* is a Studio Apartment Dog guarantees that *A* is in some way defective. Not so the fact that *A* is a dog. That is why what fixes what *A* ought to be is not the fact that it is a Studio Apartment Dog but instead the fact that it is a dog.

6. Let us stop for another moment to take note of something more general that is suggested by what we have so far.

We were looking at an idea, namely that what marks a directive as true of a thing turns on what kinds the thing would be a *defective* member of if the thing does not do what the directive says it ought to do. It is not easy to see how exactly that idea should be made more precise, but I suggest that it is right to think that the concept 'defect' lies at the heart of the concept 'ought'.

Similarly, the concept 'defect' lies at the heart of the normative concept 'normal'—that is, the normative rather than the statistical concept 'normal'. Someone who says "Adult human beings normally have 32 teeth" might mean that adult human beings mostly have 32 teeth; but he might instead mean that it is a 'norm' for the species for adults to have 32 teeth, the truth of which is compatible with its being the case (as perhaps it is) that adult human beings mostly have fewer than 32 teeth, having lost some due to baseball or gum disease. What is it for it to be a norm for the species for adults to have 32 teeth? It is a physical defect in an adult to have fewer.

Again, consider the sentence

Dogs are normally capable of barking.

Someone who says that sentence may mean that dogs are mostly capable of barking; but he may instead mean that it is a norm for dogs to be capable of barking. What is it for it to be a norm for dogs to be capable of barking? It is a physical defect in a dog to not be.

We can think of normative judgments such as that adult human beings normally have so and so many teeth, and that dogs are normally capable

of barking, as themselves directives—they tell us what adult human beings and dogs ought to be like, that being true in that they are physically defective if they are not.

7. Let us go back to where we were. *A* is a jewel thief, and

(5) *A* ought to steal good jewelry

should turn out to be false. Why does it? I said that the Second Candidate Thesis supplies us with the needed ground for saying it is false—for although *A* is a defective jewel thief if he doesn't steal good jewelry, the directive kind jewel thief is a sub-kind of the directive kind person, and *A* is a defective person if he does steal good jewelry.

So I was assuming that the kind person is a directive kind. Is it?—*is* there such a property as being a defective person?

I am sure that when I said that *A* is a defective person if he steals good jewelry, you assumed I meant that *A* is a morally defective person if he steals good jewelry. Indeed, I take it that there is such a property as being a defective person, and that what it consists in *is* being a morally defective person.

But if we so take it, then there are other directives we will have trouble with. For example, let *A* be any human being. Then we should be able to say, truly,

(7) *A* ought to be capable of seeing,

and of hearing, speaking, walking, and so on. (It is a norm for the species that its members are capable of such things.) Yet it is no moral defect in a man born blind that he is incapable of seeing. So if being a defective person is being a morally defective person, then we can't explain what makes (7) true by appeal to the Second Candidate Thesis.

Again, let *A* be any human being. Then we should be able to say, truly,

(8) *A* ought to be capable of reasoning.

(It is another norm for the species that its members are capable of reasoning.) Yet a man who has been caused by an ailment to be incapable of reasoning is not thereby marked as a morally defective person.

Should we say that being a defective person is, after all, a disjunctive property—the disjunction of being a morally defective person, being a physically defective person, and being a mentally defective person? To do so is to commit oneself to the idea that a man born blind, or caused by an ailment to be incapable of reasoning, is a defective person. That strikes me as unacceptable. Physically defective in the one case, mentally defective in the other. In neither case, all simply, *a defective human being*.

If you disagree, then (7) and (8) make no trouble for you: that is, you can explain why they are true by appeal to the Second Candidate Thesis. For those who agree, I offer a different explanation.

More precisely, a different explanation of what makes a directive about a human being true or false. Directives about toasters and beefsteak tomatoes make no trouble for the Second Candidate Thesis. For those kinds, being a defective member is being a physically defective member—they lead such narrow lives that for them there is no such thing as being a morally or mentally defective member. Dogs and cats are presumably halfway houses. They lead richer lives than toasters and beefsteak tomatoes, and while there is no such thing as being a morally defective dog or cat, there is such a thing as being a mentally defective dog or cat. I will bypass them. From here on, I will focus on people—"A" will from here on always refer to a human being, indeed, to an adult human being.

Moreover, I will focus only on some among the many different kinds of directives about people. And while I take the word "morally" to be redundant in "morally defective person", I will retain it—in order to keep clearly before us that I am distinguishing among the properties being a morally, physically, or mentally defective person.

8. Let us say that V_{body}-ings are V-ings that consist in a person's being in a certain bodily state. I take it that we can say:

(T_1) For it to be the case that A ought to V_{body} is for it to be the case that if A does not V_{body}, then A is a physically defective person.

("T_1" is short for "first thesis".) Then

(7) A ought to be capable of seeing

is true since if A is not capable of seeing, then A is a physically defective person.

Let us say that V_{mind}-ings are V-ings that consist in a person's being in a certain mental state. We might well think that we can say:

(T_2) For it to be the case that A ought to V_{mind} is for it to be the case that if A does not V_{mind}, then A is a mentally defective person.

If we can, then

(8) A ought to be capable of reasoning

is true, since if A is not capable of reasoning, then A is a mentally defective person. (T_2), alas, will call for revision; we will return to it later.

Let us say that V_{act}-ings are V-ings that consist in 'doings'—as it might be, eating a banana or giving Smith a banana. I think it is at first sight plausible to think we can say:

(T_3) For it to be the case that *A* ought to V_{act} is for it to be the case that if *A* does not V_{act}, then *A* is a morally defective person.

Then

(5) *A* ought to steal good jewelry

is false since the following is false: if *A* does not steal good jewelry, then *A* is a morally defective person.

(T_3) won't do, however. Suppose that *A* is aware that his child has a fever, and believes that giving it aspirin would cure it, and that he has some aspirin. Then the following is the case: if *A* does not give his child aspirin, then he is a morally defective person. (T_3) therefore yields that

(9) *A* ought to give his child aspirin

is true. But suppose that aspirin would in fact be bad for it. I take it to be implausible to think that *A* ought all the same to give it aspirin.

Again, let us also suppose that we know that aspirin would in fact be bad for it. If *A* knew what we know, then the following would be the case: if *A* gives his child aspirin, then he is a morally defective person. So if *A* knew what we know, then (T_3) would yield that

(10) *A* ought to refrain from giving his child aspirin

was true. It cannot plausibly be thought that (9) is true, though we know something such that if *A* knew it, then it would instead be (10) that was true.

Finally—and I take this to be conclusive—suppose that, just to be sure, *A* asks us "Ought I give it aspirin?" It would be utterly wild in us to reply: "Well, tell us what you believe about aspirin. If you believe that aspirin would be good for your child, then you ought to give it aspirin, but if you believe that aspirin would be bad for it, then you ought not give it aspirin."

Quite generally, when a person asks us "Ought I V_{act}?" it cannot be thought that we must find out what *he* thinks will happen if he V_{act}s and if he does not V_{act}, it being *that* that our answer will have to turn on. Thus I suggest that we must accept, quite generally:

(Objectivity-Thesis-Acts) Whether "*A* ought to V_{act}" is true turns, not on what *A* believes, but on what is in fact the case.[7]

[7] There is a phenomenon that tends to annoy teachers of moral philosophy. We describe a hypothetical case and invite our students to express an intuition as to what the agent in the case ought to do; our students then ask how we can be sure that the facts are as we said they are, and that there aren't any other relevant facts that we have overlooked. There are of course ways of dealing with that question; but I take their asking it to be a sign that the Objectivity-Thesis-Acts is (rightly) at work in them.

It should be clear that opting for that thesis does not commit us to regarding a person's beliefs as irrelevant to any moral judgment at all. Suppose that *A*'s giving his child a certain antibiotic Alpha would be good for it, and that *A* has some Alpha. Then, as the Objectivity-Thesis-Act says, *A* ought to give his child Alpha. What if *A* thinks that Alpha would cause his child's death, and gives it some in order to kill it? We can say that although it was true that *A* ought to give his child Alpha, many other things are also true. "Ought" is not the only normative term in our vocabulary: there are plenty of others available to us. We can say (i) *A* was at fault for giving his child Alpha. We can also say (ii) *A*'s giving his child Alpha marks *A* as a morally bad person. (This is what John Stuart Mill suggested that we should say.) We can also say (iii) *A*'s giving his child Alpha was a morally bad act. (This is what W. D. Ross suggested that we should say.) A fourth possibility will turn up in the following section.

But then do we have to give up the idea that the concept 'defect' lies at the heart of the concept 'ought'? No. We must just be a little more careful about the connection between those concepts. In particular, we must go counterfactual. We should not say:

(T_3) For it to be the case that *A* ought to V_{act} is for it to be the case that if *A* does not V_{act}, then *A* is a morally defective person.

We should prefer the likes of

(T_3*) For it to be the case that *A* ought to V_{act} is for it to be the case that if *A* knew everything that would be the case if he V_{act}-ed and if he did not, then *A* would be a morally defective person if he did not.

Thus given that if *A* gives his child aspirin, his doing so will be bad for it, *A* ought not give his child aspirin—and that is because if *A* knew that his giving his child aspirin would be bad for it, then he would be a morally defective person if he gave it aspirin.

I am sure that some objections to (T_3*) will have struck you straightway. I will take space to discuss only three of the most important ones.

9. But let us first stop to look at some *V*-ings that I do not mean to include among the V_{act}-ings, but that involve V_{act}-ings in a certain way. I said in the preceding section: let us say that V_{act}-ings are *V*-ings that consist in 'doings'—as it might be, eating a banana or giving Smith a banana. Other examples are refraining from giving Smith a banana, and causing Jones's death.

Now people very often do one thing in order to do another. Let us say that Φ-ings are *V*-ings that consist in doing one thing in order to do

another—thus V_{act-1}-ing-in-order-to-V_{act-2}. Here is an example: giving Smith a banana in order to cause Jones's death. I stress: *A* V_{act-1}s-in-order-to-V_{act-2} only if he V_{act-1}s, but it is possible that *A* V_{act-1}s-in-order-to-V_{act-2} even if he does not succeed in V_{act-2}-ing. Thus you stab a man in order to kill him only if you stab him; but you might stab a man in order to kill him without succeeding in killing him.

It is often true to say "*A* ought to V_{act}." Thus, for example, we were supposing that Alpha would be good for *A*'s child, and that *A* therefore ought to give his child Alpha. Is there some intention such that *A* ought to give his child Alpha with that intention? I suggest that there isn't.

But it is intuitively plausible to think that there are intentions such that *A* ought *not* give his child Alpha with that intention. Thus if *A* gives his child Alpha in order to cause its death, then it is intuitively plausible that he does something he ought not do. It is not the case that he ought not give his child Alpha. But it is intuitively plausible that he ought not give his child Alpha in order to cause its death.

I said in the preceding section that if *A* thought that Alpha would cause his child's death, and gave it some in order to cause its death, then although it was true that *A* ought to give his child Alpha, many other things are also true. I said that we can say (i) *A* was at fault for giving his child Alpha, and (ii) *A*'s giving his child Alpha marks *A* as a morally bad person, and (iii) *A*'s giving his child Alpha was a morally bad act. It is intuitively plausible that we can also say (iv) *A* ought not have given his child Alpha in order to cause its death.

What would make it true that *A* ought not V_{act-1}-in-order-to-V_{act-2}? I suggest that we should say the following:

(T_4) For it to be the case that *A* ought not V_{act-1}-in-order-to-V_{act-2} is for it to be the case *either* that *A* ought not V_{act-1}, *or* that *A* ought not V_{act-2}.

The role of the first disjunct is obvious enough. If *A* ought not give his child aspirin, then he obviously ought not give his child aspirin in order to V_{act-2}, whatever V_{act-2}-ing may be. Consider Alpha, however. *A* ought to give his child Alpha, so the first disjunct is not true. But if, as I should think we can presume, *A* ought not cause his child's death, then the second disjunct is true. (T_4) therefore yields—as it should—that *A* ought not give his child Alpha in order to cause its death.

My inclusion of that second disjunct relies on the quite general idea that if one ought not do a thing, then one also ought not try to do it. That seems to me very plausible.

10. To return from Φ-ings to V_{act}-ings. I said that I would discuss three of the most important objections to

(T_3*) For it to be the case that *A* ought to V_{act} is for it to be the case that if *A* knew everything that would be the case if he V_{act}-ed and if he did not, then *A* would be a morally defective person if he did not.

The first is that it is intuitively implausible to think that it is moral considerations alone that fix whether a person ought to V_{act}. Consider

(11) Smith ought to get his teeth fixed this week.

Are we really to say that what makes (11) true, if it is, is the fact that if Smith knew everything that would happen if he got his teeth fixed this week and if he did not, then he would be a *morally* defective person if he did not? Suppose that what Smith knows in knowing everything that will happen if he gets his teeth fixed this week and if he does not includes the following:

(i) If he gets his teeth fixed this week, he will suffer some pain in the process,

and

(ii) If he does not get his teeth fixed this week, he will suffer considerable pain in the weeks to come, and will in any case have to get them fixed later.

Other things being equal, he ought to get his teeth fixed this week. But is it a *moral* defect in him if he does not?

Well, if he doesn't, why doesn't he? One possibility is fear of the pain that he will suffer in the process of getting his teeth fixed this week. That would be cowardice, a moral defect on any view.

Alternatively, he might want to avoid pain this week, not caring now about future pain. That would be imprudence. My own view is that imprudence is itself a moral defect, and thus that if he fails to get his teeth fixed this week for this reason, then it is a moral defect in him to not do so.

Perhaps you don't regard imprudence as a *moral* defect? I will not argue the matter here. If you don't regard imprudence as a moral defect, you are invited to take it that when I say "morally defective" I mean "morally defective or imprudent". In short, (T_3*) is to be so understood that since Smith would be imprudent if—knowing (i) and (ii)—he did not get his teeth fixed this week, (T_3*) yields that he ought to get his teeth fixed this week.

It should be stressed, though, that accepting (T_3*), so understood, does not commit us to accepting that if *A*'s V_{act}-ing would be imprudent, then *A* ought not V_{act}. It might be that *A*'s giving his child a certain medicine would be imprudent, given the limited information *A* has in hand, compatibly with its being the case that the medicine would be good for the child, and

therefore that he ought to give it to the child. What (T_3*) says matters to the truth of a directive is not the imprudence in a person who acts imprudently on limited knowledge, but rather the imprudence in a person who knows what will happen but discounts the far future in favor of the near.

11. A second objection to

(T_3*) For it to be the case that *A* ought to V_{act} is for it to be the case that if *A* knew everything that would be the case if he V_{act}-ed and if he did not, then *A* would be a morally defective person if he did not

emerges as follows. Suppose that *A* knows that his child has a fever, and does not know what to do. Suppose that *A*'s medical advisor has given *A* a letter describing what to do in case *A*'s child has this or that ailment. Then

(12) *A* ought to open the envelope

is surely true. But now let us ask: *can* it be the case that *A* knows everything that would be the case if he V_{act}-ed and if he did not? Well, what are the things that would be the case if he did and if he did not?

Suppose that the letter inside the envelope truly says that Alpha would cure the child; so Alpha would cure the child—indeed, Alpha would cure the child whether or not *A* opens the envelope. So if *A* opens the envelope, and also if *A* doesn't open the envelope,

(13) Alpha would cure the child.

Suppose also that, as a matter of fact, *A* will not know that Alpha would cure the child unless and until he opens the envelope. Alternatively put:

(14) If *A* knows at *t* that Alpha would cure the child, then *t* post-dates *A*'s opening the envelope.

Suppose, finally, that the time is NOW, and that *A* has not yet opened the envelope. And let us ask: can it be supposed that *A* knows at NOW that both (13) and (14) are true? If he knows at NOW that (13) is true, then he knows at NOW that Alpha would cure the child. If he also knows at NOW that (14) is true, then he knows at NOW that NOW post-dates *A*'s opening the envelope. Thus he knows that he has already opened the envelope. But by hypothesis, he hasn't yet opened it.

So it can't be the case that *A* knows *everything* that would be the case if he opened the envelope and if he did not.[8] Let us therefore revise (T_3*). Let us say, instead:

[8] The difficulty here does not arise merely for cases in which what is in question is doing something one needs to do in order to find out what to do. Suppose that if Smith kills Jones he will later come to know that Alfred hates Bert, and that if Smith does not

(T_3**) For it to be the case that *A* ought to V_{act} is for it to be the case that if *A* knew everything it is consistent to suppose he knows about what would be the case if he V_{act}-ed and if he did not, then *A* would be a morally defective person if he did not.

12. A third objection to (T_3*) is one which opting for (T_3**) is no defense against. It emerges from the same example. We supposed that *A* knows that his child has a fever, and does not know what to do. We supposed that *A*'s medical advisor has given *A* a letter describing what to do in case *A*'s child has this or that ailment. Then, I said,

(12) *A* ought to open the envelope

is surely true.

But now suppose, as we also supposed, that the letter inside the envelope truly says that Alpha would cure the child; so Alpha would cure the child whether or not *A* opens the envelope. Suppose, as we also did, that *A* knows this. Suppose, finally, that *A* also knows everything else it is consistent to suppose he also knows about what would be the case if he opened the envelope and if he did not. Would he be a morally defective person if he did not open it? No. For given he knows that Alpha would cure the child whether or not he opens the envelope, there is no reason for him to open it. Given this piece of knowledge about what would happen if he opened it, he *already* knows what would cure the child and hence has no need to open the envelope to find out.

The kind that this case falls into is important. Cases of this kind are cases in which the ground—indeed, the only ground—for thinking that a person ought to V_{act} is that there is something he ought to do, such that he will find out that he ought to do it if and only if he V_{act}s. It should therefore be no surprise that imagining *A* to know what will happen if he V_{act}s and if he does not makes trouble for the possibility of explaining why he ought to V_{act} in such a case.

So if we think, as I do, that (12) is true, then we must emend (T_3**). One way of doing so is for us to accept, instead,

(T_3***) For it to be the case that *A* ought to V_{act} is for it to be the case that *either*

- (α) if *A* knew everything it is consistent to suppose he knows about what would be the case if he V_{act}-ed and if he did not, then *A* would be a morally defective person if he did not, *or*
- (β) for some act-kind V_{act}*-ing, *A* ought to V_{act}*, and *A* can find out that he ought to V_{act}* only by V_{act}-ing.

kill Jones he will not later come to know that Alfred hates Bert. Same problem; same solution called for.

Thus let us ask: is

(12) *A* ought to open the envelope

true? Well, consider

(I) For it to be the case that *A* ought to open the envelope is for it to be the case that *either*

- (α) if *A* knew everything it is consistent to suppose he knows about what would be the case if he opened the envelope and if he did not, then *A* would be a morally defective person if he did not, *or*
- (β) for some act-kind V_{act}*-ing, *A* ought to V_{act}*, and *A* can find out that he ought to V_{act}* only by opening the envelope.

Clause (α) of (I) is false. What about clause (β) of (I)? Since giving his child Alpha would cure his child, I take it that *A* ought to give his child Alpha. Is that true? Well, consider

(II) For it to be the case that *A* ought to give his child Alpha is for it to be the case that *either*

- (α) if *A* knew everything it is consistent to suppose he knows about what would be the case if he gave his child Alpha and if he did not, then *A* would be a morally defective person if he did not, *or*
- (β) for some act-kind V_{act}*-ing, *A* ought to V_{act}*, and *A* can find out that he ought to V_{act}* only by opening the envelope.

I take it that clause (α) of (II) is true, and therefore that *A* ought to give his child Alpha. Since *A* ought to give his child Alpha, and *A* can find out that he ought to only by opening the envelope, clause (β) of (I) is true. Therefore (12) is true—as it should be.

My inclusion of that second disjunct (β) in (T_3***) relied on the quite general idea that if one ought to do a thing, then one also ought to do what is necessary to find out that one ought to do it. That seems to me plausible. But perhaps it is over-strong? Suppose that *A* can find out that he ought to give his child Alpha only by torturing Smith, who villainously withholds the information from *A*. On some views, we ought never torture anyone, however villainous, and however dreadful the outcome will otherwise be. I bypass the question in moral theory whether those views are correct. I merely indicate a way in which we can accommodate them if we accept them—we can replace (β)of (T_3***) with:

for some act-kind V_{act}*-ing, *A* ought to V_{act}*, and *A* can find out that he ought to V_{act}* only by V_{act}-ing, *and* it is not the case that *A* ought to refrain from V_{act}-ing.

On those views, it is always and everywhere the case that *A* ought to refrain from torturing Smith, and we can explain what makes *that* true by appeal to

clause (α) of (T_3***). I leave it open whether we should accept this further revision.

13. There is room for other objections to (T_3***), but none, I think, that calls for still further revision in it. In any case, I will bypass them. Let us instead turn from V_{act}-ings back to V_{mind}-ings.

I said: let us say that V_{mind}-ings are V-ings that consist in being in a mental state. I said that we might well think that we can say:

(T_2) For it to be the case that A ought to V_{mind} is for it to be the case that if A does not V_{mind}, then A is a mentally defective person.

If we can, then

(8) A ought to be capable of reasoning

is true, since if A is not capable of reasoning, then A is a mentally defective person. I said, however, that (T_2) would call for revision.

The need for revision that I have in mind issues from consideration of certain kinds of V_{mind}-ings. Consider, first, the class of propositions that we get by inserting some sentence in for "p" in "A ought to believe that p" - as, for example,

(15) A ought to believe that Smith is taller than Jones.

Believing something is being in a mental state, so believing that Smith is taller than Jones is an instance of V_{mind}-ing. Should we take it that (T_2) is right about what it is for (15) to be true?

We should stop first, however, to take note of an argument that many people have taken seriously—an argument to the effect that the likes of (15) are never true. The argument proceeds as follows. First premiss: it cannot be true to say that a person ought to V unless the person can V at will. Second premiss: it is not possible for a person to believe a thing at will. Conclusion: it cannot be true of any person that he ought to believe a thing. So (15) and its ilk are all false.

I say that many people have taken that argument seriously, though many of them have argued that it should be rejected, thus that one or the other of the premisses is false. But the argument is not really worth taking seriously, since its first premiss is so obviously false.

For (i) believing a thing is being in a certain state, and one can't be in a state—*any* state—at will. Trusting a person is being in a mental state. So is preferring X to Y. Being in Chicago is being in a physical state. So also is weighing so and so many pounds. One can't be in any of these states at will. That is not a deep point, it is right up at the surface. Being in a state isn't something that is done, and a fortiori one can't do it at will.

Yet (ii) there are states such that it is very often true to say of a person that he ought to be in this one or that. It may be true that *A* ought to trust *B*. It may be true that *A* ought to prefer *X* to *Y*. Suppose that *A* has promised to be in Chicago today, and that people have been counting on his being there today. I run across *A* in Harvard Square today, and I say, in some surprise, "You ought to be in Chicago today!" What I say may be true. Again, a doctor may say of a child "Given its age, that child ought to weigh more than 37 pounds," and be speaking truly when he does.

And (iii) there is no reason at all to think that believings are unique among states in that it is not possible for a person to say the likes of (15) and be speaking truly when he does.

It may be objected that "A ought to be in Chicago today" is true only if *A* could at will have done something that would have caused him to be in Chicago today. (In that he could at will have caught a plane or train.) And it might therefore be suggested that a weaker pair of premisses will suffice for the conclusion. Weaker first premiss: it cannot be true to say that a person ought to *V* unless the person could, at will, have caused himself to *V*. Weaker second premiss: it is not possible for a person to, at will, cause himself to believe a thing. Conclusion: it cannot be true of any person that he ought to believe a thing.

I leave open whether it is right to think that "*A* ought to be in Chicago today" is true only if *A* could, at will, have caused himself to be in Chicago today. (We must in any case allow that even if *A* could, at will, have done something yesterday that would have caused him to be in Chicago today, he may by now have left it too late: there may be nothing at all that he could, at will, do now that would cause him to be in Chicago today—compatibly with its being the case that he ought to be there today.) But it is in any case wrong to think that "*A* ought to trust *B*" is true only if *A* could, at will, have caused himself to trust *B*. And wrong to think that "*A* ought to prefer *X* to *Y*" is true only if *A* could, at will, have caused himself to prefer *X* to *Y*. And wrong to think that "That child ought to weigh more than 37 pounds" is true only if the child could, at will, have caused itself to weigh more.

So let us ignore this argument, and turn to the question what might make it true that a person ought to believe a thing.

14. I said that we might well think we can say:

(T_2) For it to be the case that *A* ought to V_{mind} is for it to be the case that if *A* does not V_{mind}, then *A* is a mentally defective person.

If we can, then for

(15) *A* ought to believe that Smith is taller than Jones

to be true is for it to be the case that if *A* does not believe that Smith is taller than Jones, then *A* is a mentally defective person.

This kind of view is very popular. It is often said that *A* ought to believe that Smith is taller than Jones just in case the total body of evidence *A* has in hand for that hypothesis supports it.[9] Why so? Because if the total body of evidence *A* has in hand for the hypothesis supports it, then it would be irrational in him to not believe it.

I suggest, however, that there is a conclusive objection to (T_2)—a first cousin of the objection to

(T_3) For it to be the case that *A* ought to V_{act} is for it to be the case that if *A* does not V_{act}, then *A* is a morally defective person

that I drew attention to earlier. I said: suppose that, just to be sure, *A* asks us "Ought I give my child aspirin?" It would be utterly wild in us to reply: "Well, tell us what you believe about aspirin. If you believe that aspirin would be good for your child, then you ought to give it aspirin, but if you believe that aspirin would be bad for it, then you ought not give it aspirin." Quite generally, when a person asks us "Ought I V_{act}?" it cannot be thought that we must find out what *he* thinks will happen if he V_{act}s and if he does not V_{act}, it being *that* that our answer will have to turn on. Thus I suggested that we should accept, quite generally:

(Objectivity-Thesis-Acts) Whether "*A* ought to V_{act}" is true turns, not on what *A* believes, but on what is in fact the case.

Suppose that, just to be sure, *A* asks us "Ought I believe that Smith is taller than Jones?" It would be utterly wild in us to reply: "Well, tell us what evidence you have in hand about Smith and Jones. If the evidence you have in hand is evidence for the hypothesis that Smith is taller than Jones, then you ought to believe that he is, but if the evidence you have in hand is evidence for the hypothesis that Jones is taller than Smith, then you ought

[9] A view of this kind appears in Feldman (2000). Feldman inserts a condition: one ought to believe a hypothesis one's evidence supports *if* one is going to have any doxastic attitude toward it at all.

It is perhaps worth stressing that what concerns us here is only normative judgments of the form "*A* ought to believe *P*." I thank Johanna Goth for drawing my attention to the fact that, like "*A* ought to V_{act}," "*A* ought to believe that *p*" has a probabilistic as well as a normative meaning. Thus someone who says "*A* ought to believe that giving his child Alpha would be good for it" is likely to be making a normative judgment. But he might be making a probabilistic judgment instead. Suppose that there is a kind of pill that much experience suggests works as follows: if you give one to a person at 10 a.m., then at 11 a.m. he believes that giving his child Alpha would be good for it. Then if we gave *A* one of those pills at 10 a.m., we may say at 11 a.m., with confidence, "By now, *A* ought to believe that giving his child Alpha would be good for it," meaning that it is by now likely that *A* believes that. See Goth (unpublished ms.).

not believe that Smith is taller than Jones." Quite generally, when a person asks us "Ought I believe that *p*?"—and it hardly needs saying that people do very often ask such questions—it cannot be thought that we must find out what evidence *he* has in hand, it being *that* that our answer will have to turn on. Thus I suggest that we should accept, quite generally:

(Objectivity-Thesis-Beliefs) Whether "*A* ought to believe that *p*" is true turns, not on the evidence that *A* has in hand, but on what is in fact the case.

Now it is plain enough that we cannot say: "*A* ought to believe that *p*" is true just in case *p*. Suppose that the Governor of Massachusetts in fact prefers chocolate ice cream to vanilla. It isn't the case that anyone and everyone ought to believe that he prefers chocolate ice cream to vanilla.

There are matters about which a person need have no opinion, and matters about which he ought to. Children are subject to a variety of ailments, and in the case of the common childhood ailments, a parent ought to have an opinion about what to do in case his child suffers from this one or that. By contrast, it is not the case for most of us anyway that we ought to have an opinion about whether the Governor of Massachusetts prefers chocolate to vanilla.

Indeed, a parent ought not merely have an opinion about what to do in case his child suffers from one of the common childhood ailments: he ought to have a true opinion about what to do in such a case. (The true opinions surely include that when in doubt, it is safest to phone the doctor.) It may be the case that a certain parent who lacks a true opinion about what to do is not morally defective, and is not irrational. No matter. Whether he ought to have a true opinion is an objective matter, turning not on his beliefs but on what is in fact the case.

I suggest that what is needed here is that we appeal to considerations of the kind I drew attention to in section 12 above, namely considerations having to do with finding out that one ought to do a thing—something along the following lines:

For it to be the case that *A* ought to have a true opinion about whether *p* is for it to be the case that there is, or may well come to be, some act-kind V_{act}-ing such that *A* ought to V_{act}, and *A* does, or would then, know that he ought to V_{act} only if he has a true opinion about whether *p*.

As I said, children are subject to a variety of ailments, and a parent ought to have a true opinion about what to do in case his child suffers from this one or that. By contrast, while it just might be the case that Smith or Jones ought to have an opinion about whether the Governor of Massachusetts prefers chocolate to vanilla, that is not the case for most of us, and that is because there is nothing we need to do, or may well need to do, that we can know of only if we have a true opinion about his preference.

And then I suggest that we should say:

(T_2-belief) For it to be the case that *A* ought to believe that *p* is for it to be the case that

(α) *A* ought to have a true opinion about whether *p*, and

(β) *p*.[10]

Analogously for trustings and preferrings, which I mentioned in section 13—and also for admirings, wantings, and so on. If *A* asks us whether he ought to trust *B*, we don't take our answer to be fixed by what *A* thinks about *B*; if we know that *B* is untrustworthy, we say "No, you ought not," whatever we believe *A*'s beliefs to be. On the other hand, it isn't required of people that they trust everybody who is trustworthy, just as it isn't required of people that they believe everything that is true. So I suggest that we should say:

(T_2-trust) For it to be the case that *A* ought to trust *B* is for it to be the case that

(α) *A* ought to have a true opinion about whether *B* is trustworthy, and

(β) *B* is trustworthy.

(I hope it is clear that this thesis does not tell us that trusting *B* itself is, or includes, believing that *B* is trustworthy. It tells us only what it is for it to be the case that *A ought to* trust *B*.) Analogously for preferrings, admirings, and so on. Let us call these the T-Theses.

Then let us revise the terminology I introduced earlier. Let us restrict the V_{mind}-ings to believings, trustings, preferrings, admirings, and so on. Then I suggest that we should say: for each V_{mind}-ing, it is the appropriate T-Thesis that tells us what it is for it to be the case that *A* ought to V_{mind}.

Being capable of reasoning is a mental norm for our species, and presumably there are others. Let us take them to be instances of $V_{\text{mental capacity}}$, and say

(T_2*) For it to be the case that *A* ought to $V_{\text{mental capacity}}$ is for it to be the case that if *A* does not $V_{\text{mental capacity}}$, then *A* is a mentally defective person.

Then we can still have that

(8) *A* ought to be capable of reasoning

is true, though it is (T_2*) rather than (T_2) that tells us why it is.

[10] If this idea is right, then perhaps we can say that the ambiguity in "ought" as between a normative and a probabilistic meaning is weak ambiguity, explainable as follows: "*A* $\text{ought}_{\text{probabilistic}}$ to *V*" is equivalent to "For any person *X*, if it is or were the case that *X* $\text{ought}_{\text{normative}}$ to have a true opinion about whether *A V*s, then it is or would be probable that *X* $\text{ought}_{\text{normative}}$ to believe that *A V*s."

This section supplies only a sketch of an account of the directives about people of the form "*A* ought to V_{mind}." However, I thought it should be included because I take it to be important to keep in mind that there are directives *everywhere* in our thinking: focusing exclusively on this or that narrow sub-class of the directives is bound to mislead us about what and how much the concept 'ought' does for us.

15. In sum, we have reached theses governing directives about people of a great many kinds: "*A* ought to V_{body}," "*A* ought to V_{act}," "*A* ought to Φ," "*A* ought to V_{mind}," and "*A* ought to $V_{\text{mental capacity}}$." There are other V-ings for which it may be true that a person ought to V. For example, there is being at a certain place at a certain time: as I said earlier, it may be true that *A* ought to be in Chicago today. It may be true that a person ought to engage in a practice, as for example exercising, cutting down on empty calories, and making regular appointments with the dentist. It would be very welcome if we could arrive at a generalization covering all normative judgments of the form "*A* ought to *V*" that are about people. I do not try to produce such a generalization here. What matters now is just that it be plausible that the concept 'defect' lies at the heart of the concept 'ought'. That it does is easily seen in the case of directives about things that are not people. Similarly for directives about people of the form "*A* ought to V_{body}" and "*A* ought to $V_{\text{mental capacity}}$." If, as I have suggested, the other directives we looked at are analyzable in terms of instances of "*A* ought to V_{act}," and those, in turn, in terms of counterfactuals that ascribe defects, then the concept 'defect' is at work in all of them.

If that is right, then we have in hand the possibility of an alternative to Consequentialism. I said at the outset that directives do seem to call for justification by appeal to the truth of evaluatives. The theory in the offing here is like Consequentialism in that it says that what makes directives true, when they are, is the truth of certain evaluatives; but the evaluatives it focuses on are more or less complex, and it is the concept 'defect' rather than the concept 'goodness' that gives them their normative power.

Moreover, in focusing on defects rather than goodness, the theory in the offing here is safe against the familiar objections to Consequentialism that lie in the fact that Consequentialism may require doing what is in fact unjust, and that it is too demanding. For the theory in the offing here requires of a person only that he do what it would be a defect in him to not do—what it would be unjust or ungenerous or irresponsible or cruel or imprudent … in him to fail to do—if he knew what the consequences of his acting or refraining would be. That strikes me as a major improvement.

I add that I think it also preferable to those theories about what a person ought to do that are nowadays called virtue theories: according to those

theories, what a person ought to do is what a virtuous person would do. The theory in the offing here is instead a vice theory. That too strikes me as an improvement, since "ought" is weak: what it requires of a person is only meeting minimal standards—it requires only such conduct as one would be marked as morally *defective* for *failing* to engage in if one knew what would happen if one did and what would happen if one did not engage in it. Of course it isn't nothing to do what one ought to do, and on occasion, it may be much. But it typically isn't much. "He did what he ought" is very rarely high praise.

REFERENCES

Bloomfield, Paul (2001) *Moral Reality* (Oxford: Oxford University Press).

Feldman, Richard (2000) "The Ethics of Belief," *Philosophy and Phenomenological Research*, 60: 667–95.

Goth, Johanna, "Understanding 'Ought to Believe Sentences'" (unpublished MS.).

Wolterstorff, Nicholas (1980) *Works and Worlds of Art* (Oxford: Clarendon Press).

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

This page intentionally left blank

Index

Note: Footnote numbers in brackets are used to indicate the whereabouts on the page of authors who are quoted, but not named, in the text.