OXFORD

Oxford Studies in Metaethics Volume 1

OXFORD STUDIES IN METAETHICS

This page intentionally left blank

Oxford Studies in Metaethics

VOLUME 1

Edited by
RUSS SHAFER-LANDAU

CLARENDON 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 trademark 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 2006

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 0–19–929188–8 978–0–19–929188–5
ISBN 0–19–929189–6 (Pbk.) 978–0–19–929189–2 (Pbk.)

1 3 5 7 9 10 8 6 4 2

Contents

Notes on Contributors

Terence Cuneo is Assistant Professor of Philosophy, Calvin College

Justin D'Arms is Associate Professor of Philosophy, The Ohio State University

James Dreier is Professor of Philosophy, Brown University

Allan Gibbard is Richard B. Brandt Distinguished University Professor of Philosophy, University of Michigan

Terry Horgan is Professor of Philosophy, University of Arizona

Nadeem J. Z. Hussain is Assistant Professor of Philosophy, Stanford University

Daniel Jacobson is Associate Professor of Philosophy, Bowling Green State University

Derek Parfit is Senior Research Fellow, All Souls College, University of Oxford

Peter Railton is John Stephenson Perrin Professor of Philosophy, University of Michigan

Mark van Roojen is Professor of Philosophy, University of Nebraska, Lincoln

Nishi Shah is Assistant Professor of Philosophy, Amherst College

Sergio Tenenbaum is Associate Professor of Philosophy, University of Toronto

Mark Timmons is Professor of Philosophy, University of Arizona

Pekka Väyrynen is Assistant Professor of Philosophy, University of California, Davis

Ralph Wedgwood is CUF Lecturer in Philosophy, University of Oxford

Introduction

Russ Shafer-Landau

This is the inaugural volume of *Oxford Studies in Metaethics*. This series is devoted exclusively to original philosophical work in the foundations of ethics. It provides an annual selection of much of the best new scholarship being done in the field. Its broad purview includes work being done at the intersection of ethical theory and metaphysics, epistemology, philosophy of language, and philosophy of mind. The essays included in the series provide an excellent basis for understanding recent developments in the field; those who would like to acquaint themselves with the current state of play in metaethics would do well to start here.

The contents of this volume of *Oxford Studies in Metaethics* nicely mirror the variety of issues that make this area of philosophy so interesting. The volume opens with Peter Railton's exploration of some central features of normative guidance, the mental states that underwrite it, and its relationship to our reasons for feeling and acting. In the next offering, Terence Cuneo takes up the case against expressivism, arguing that its central account of the nature of moral judgements is badly mistaken. Terence Horgan and Mark Timmons, two of the most prominent contemporary expressivists, then present their thoughts on how expressivism manages to avoid a different objection—that of collapsing into an objectionable form of relativism. Daniel Jacobson and Justin D'Arms next offer an article that continues their research program devoted to exploring the extent to which values might depend upon, or be constrained by, human psychology. Ralph Wedgwood engages in some classical metaethical conceptual analysis, seeking to explicate the meaning of *ought*. Mark van Roojen then contributes a new take on the Moral Twin Earth Argument, a prominent anti-realist puzzle advanced in the early 1990s by Horgan and Timmons.

Allan Gibbard next presents his latest thoughts on the nature of moral feelings and moral concepts, crucial elements in the overall project of defending the expressivism he is so well known for. James Dreier then takes up the details of Gibbard's recent efforts to provide a solution to what many view as the most serious difficulty for expressivism, namely, the Frege-Geach problem. Dreier identifies difficulties in Gibbard's expressivist account, and offers a suggestion for their solution. Sergio Tenenbaum explores the concept of a direction of fit, relied on so heavily nowadays in accounts of moral motivation. Nadeem Hussain and Nishiten Shah then consider the merits of Christine Korsgaard's influential critique of moral realism. T. M Scanlon's widely discussed buck-passing account of value attracts the critical eye of Pekka Väyrynen, who attempts to reveal the reasons that we might resist it. Derek Parfit's contribution concludes this volume, with an article on normativity that presents his most recent thinking on this fundamental notion.

Most of the articles included here took initial shape as papers delivered at the first annual Metaethics Workshop, held at the University of Wisconsin in October 2004. I'd like to thank those who served as members of the Program Committee for that event, and so as de facto referees for this volume: David Brink, David Copp, Nicholas Sturgeon, and Robert Audi. Robert also did double duty as one of the reviewers commissioned by Oxford to assist me in evaluating the contents of this first volume. He was joined in this work by Michael Brady; their criticisms and suggestions were always informed, judicious and delivered in a manner designed to be most helpful to the authors. Their efforts have led to substantial improvements in many of the papers in this inaugural volume. Finally, I'd like to express my gratitude to Peter Momtchiloff, philosophy editor at the Press, whose enthusiasm and unfailing good sense have made him the ideal partner in this exciting new enterprise.

1

Normative Guidance

Peter Railton

Introduction

I've been told that there are two principal approaches to drawing figures from life. One begins by tracing an outline of the figure to be drawn, locating its edges and key features on an imagined grid, and then using perspective to fill in depth. The other approach proceeds from the 'center of mass' of the subject, seeking to build up the image by supplying contour lines, the intersections of which convey depth—as if the representation were being created in relief. The second approach need not adopt a unified perspective, and is more concerned with evoking the volume and 'presence' of the subject than with accurate placement of edges and features. Call this second approach drawing 'from the inside out', meant to capture the living force of the subject rather than freeze it in a coordinate frame.

I sometimes feel that those of us who hanker after system in ethics tend to opt unconsciously for the first approach, tracing the outlines of moral practice from the outside and setting it into a coordinate scheme and unified perspective external to the agents themselves. We should probably try more often to work from the inside of agents, from their centers of mass as agents and moral beings. For such an approach, questions of normative guidance become questions about how normative guidance occurs within the agent, what gives norms their life, and how they enter into the shape and meaning of the agent's experience, thought, feeling, and action.

Working 'from the inside out' might suggest starting with an agent exercising reflective choice, facing the question of whether to accept a given norm and thereby endow it with life. But such higher-order reflection

Thanks to Justin D'Arms, David Copp, Allan Gibbard, Daniel Hausman, Daniel Jacobson, Michael Smith, Mark Timmons, and others who made useful comments at the original presentation of this paper. Christian Miller generously provided written comments that were exceptionally helpful in writing the current version.

occupies a small fraction of our normative lives—concentrating on it tends to locate the center of mass of our norm-guided selves *too high*, that is, too much in the domain of self-conscious, deliberative judgment. Moreover, such a focus tends to encourage us to view the agent taken in reflective isolation, giving special prominence to the individual's self-construal and making less evident the social sources of norms and meaning that make self-construal possible and practical.

So I'll suggest that we begin somewhere else, taking as our life-studies everyday human activity involving the most ordinary of norms. Using examples, I will explore some central features of normative guidance, the mental states that underwrite it, and its relationship to our reasons for feeling and acting.

Normative Guidance Caught in the Act

Martha and Rick

Martha and Rick are walking and talking together as they head for classrooms across campus in order to teach their separate classes. They aren't late, but must move fairly briskly to keep it that way. Like most such conversations, this one is pretty humdrum in content—what's doing in the department, why the lecture halls are always overheated, what to make of last week's visiting speaker, and the like. Together they must navigate their way up and down staircases, through doors, and across streets, working their way upstream in a current of hurrying students.

They accomplish this without the need to devote much thought to it. Otherwise, they'd be hard put to maintain any sort of conversation, let alone a moderately engaging one. What seems so spontaneous, however, is the result of complex inner workings, largely below the surface of experience. As they walk and chat, they are guided continuously by what they see and think, as well as what each thinks the other sees and thinks, and by a sense of how the moments before class time are ticking away. They freely begin a sentence without knowing how it will end, coordinate small changes in their shared trajectory through subtle body language, and communicate their intentions to oncoming pedestrians, cars, and bicycles by tiny eye and head movements. Similarly coordinated changes occur in when they speak and what they speak about, each giving the other small cues to direct the pace and course of the conversation. They are comfortable enough with one another that they can talk rather unguardedly, but some things will nonetheless remain unsaid. Martha is quite a bit senior to Rick, who is coming up for promotion. Much as certain issues are on both their minds, it would not occur to them to bring these into the conversation.

In all this mutually adjusted 'whir of organism', deliberate choice and self-conscious effort are largely absent. That is not at all because the underlying structure of norms and intentions is simple. Their shared intention to walk together to class, as well as the communicative intentions that underlie their conversation, are elaborately and reflexively iterated. There is no plan concerning how to walk or what to talk about, other than the constraint of getting to class on time. But for that very reason their path and conversational foci must emerge in real time, through mutual observation and adjustment—all without commentary, and without interfering with other thoughts.

This vignette is important for our purposes simply because it is so unremarkable. A complex constellation of norms is hard at work throughout—norms of sociability, language, assertion, communication, politeness, professional relations, sidewalk etiquette, and privacy—despite the automatic character of much of their action (see Bargh and Chartrand, 1999). The role of these norms in shaping the surface contours of Martha's and Rick's behavior becomes salient only if we compare how they comport themselves with the comportment of two comparably related academic colleagues in a different culture walking to class together. There we would see different norms at work not only in language, but in gestures, conversational distance, turn-taking and interrupting, modulation of voices, deference to seniority and from students, and gender appropriateness. Transplant Martha or Rick as a visitor to such a society, and the elaborate, fluent, unselfconscious mutual choreography each achieves at home would be replaced by behavior more self-conscious, tentative, effortful, and uncoordinated.

Martha and Kim

Let us now imagine that Martha is traveling, flying home after a brief visit to another department. Her connecting flight in Dallas has been canceled, and she finds herself stranded overnight. The gate agent hands her a voucher good for a meal and a night's stay at a budget airport hotel, but the thought fills her with dread. She's tired, and more than ready to be home. A friend from college, Kim, lives in Dallas. Though they haven't been in touch lately, they have kept up somewhat regularly over the intervening years. Without a further thought, Martha looks Kim up in her address book and calls—perhaps they can get together for a meal? Kim can hear the fatigued and somewhat lonesome tone in her friend's voice, and promptly invites Martha to spend the night at her place. She's got an extra room, and planned to take the morning off work tomorrow anyhow. She'll be happy to pick Martha up—there's no traffic at this hour—and deliver her back to the airport tomorrow in plenty of time for her flight. Had Kim's voice shown the slightest hesitancy in making this offer, Martha would feel she was imposing, thank Kim, and say that she's so exhausted that she prefers simply to head straight over to the hotel and bed. But Kim sounds genuinely eager. 'Great!' Martha responds, 'But I insist on taking you out to dinner.' All is agreed. They share a lively meal, and talk late into the night. Martha is up first in the morning. She pads down to the kitchen and quietly fixes herself breakfast.

As before, there are many layers of norms at work in this interaction. Norms of conversation, sociability, and coordination, to be sure, but also norms of friendship, hospitality, reciprocity, privacy, property, and propriety. The shelves in Kim's spare bedroom contain dated volumes labeled 'Journal', but Martha skips over them without a thought when looking for a bit of bedtime reading—though Kim's diaries would be of much more interest to her than the indifferent collection of short stories she ultimately settles on. In the morning, however, Martha shows no similar inhibition about making free with various contents of Kim's refrigerator.

Had Martha been stranded in Tokyo, where Kisho, an exchange student she knew well as an undergraduate now lives, she would have been much more reluctant to initiate such a phone call. She still has his phone number, and would love to see him again, but she'd be stymied by lack of normative knowledge. She would not know what Kisho might make of a call out of the blue. Would it be welcome, or even polite? If Kisho had a partner, would such a call strike her as inappropriate? Would Kisho feel bound by customary obligations of hospitality to go out of his way to arrange a proper get-together, even if he was not at all eager to do so? Might Kisho take it as a slight for Martha to be in Tokyo overnight and *not* call? Would inviting Kisho—and his partner?—out to dinner seem an affront to his hospitality? Notions of gender, friendship, reciprocity, propriety, property, and privacy are culturally articulated, and Martha would be unsure of how to translate her simple desire to see him again after all these years straightforwardly into an appropriate course of action.

Guidance by Norms

To begin our inquiry into normative guidance, we need a clearer grasp of what it is for a bit of behavior to be guided by a norm. Norm-guidance is a sufficiently complex phenomenon that we might do well to build up to it piece by piece, taking as our initial focus individual instances of conduct rather than the agent considered more globally. To start, consider what might seem a necessary condition:

> (1) Conduct *C* is guided norm *N* only if *C* is in accord with *N*.[1]

(1), of course, can't be strictly correct. Conduct can be guided by a norm even when it falls short for various reasons. For example, an agent can try wholeheartedly to conform to *N* by performing *C*, but fail because *C*-ing is insufficient to meet *N*. Still, it is worth pausing with (1) long enough to note that it has some pull. A limo driver who assures passengers that he makes it a rule to observe all traffic laws is not in fact being guided by *that* norm if he cavalierly exceeds posted speed limits while weaving in and out of lanes without signaling. Norm-guidance requires more than lip-service, however earnest. In this respect, it is like belief. Even when a person is mistaken about what she believes, these unacknowledged beliefs can nonetheless guide her expectations and conduct without her awareness. Similarly, an agent can fail to understand or acknowledge which norms are actually guiding her behavior. To be norm-guided is a matter of how one is disposed to think, act, and feel, not simply of how one sees oneself, or would like to.

Still, (1) is too strong, so let's consider replacing it with:

> (2) Conduct *C* is guided by norm *N* only if *C* is the manifestation of a reliable disposition to act in a way conducive to compliance with *N*.

Yet (2) is too stringent as well. Although mere declaration of *N* does not make one *N*-guided, a disposition to *N*-directed effort can count even when not very reliable. Someone can adopt and be guided by a norm of

[1] Here and elsewhere I make some simplifying assumptions. (i) We will assume that the agent does not have false beliefs concerning the relation of *C* to *N*. (ii) We will set aside cases in which behavior has an indirect relation to *N*, while being in some sense guided by it—e.g. you avoid Bilko's company because you suspect he plans to cheat you by violating *N*. (iii) We will also set aside various cases in which conduct is guided by *N* because *N* figures in the agent's practical reasoning, even though the upshot is not performing the action *N* requires—e.g. the agent decides to abandon or alter *N* rather than perform the act it requires, or the agent compares *N* with other applicable norms, weighs it, and determines that it is outweighed. (iv) We will ignore the complication that in virtually all cases (as our examples involving Martha, Rick, and Kim suggest) whether a bit of conduct accords with a norm *N* will depend not only upon *N* itself, but other norms besides—e.g. whether a guest, in using her host's kitchen and food without permission in preparing her breakfast, violates the host's private sphere depends upon socially variable norms. We will mostly limit our discussion to cases of *first-order behavioral* norm-guidance, *other things equal.*

eating a decent breakfast each morning even though he alters this norm after he fails on the very first day to find the time. If his new norm is 'weekend mornings only', and he succeeds in taking time to make a decent breakfast only one weekend morning in two, the revised norm can still be said to guide his behavior. It matters crucially, as we will see below, how he responds to failures to take the time. And even here we must be flexible. For he might respond to his failures in part by developing a lower self-image, which, given his psychology, actually increases his frequency of failure in the future. Still, guidance by the 'weekend mornings' norm is playing a role in his conduct both on days when he succeeds and on days when he fails. So let us relax the condition of reliability and interpret 'conducive' loosely.

In another respect, however, (2) needs strengthening. For even when conduct is attributable to a disposition to 'act in a way conducive to compliance with *N*', it need not involve guidance by *N*. Harry the receptionist cares about looking sharp, and is disposed to dress with just the degree of formality and restraint required by his company for front-office employees. But he is guided by his own sense of style rather than the company dress code, of which he is only vaguely aware. So (2) can be tightened up:

> (3) Conduct *C* is guided by norm *N* only if *C* is the manifestation of a disposition to act in a way conducive to compliance with *N*, such that the fact that *C* conduces to compliance with *N* plays an appropriate role in the explanation of the agent's *C*-ing.[2]

What could it mean to attribute an explanatory role to a seeming *abstractum* like 'the fact that *C* conduces to compliance with *N*'? One answer, the very paradigm of normative guidance in some eyes, runs like this: *A* has a mental representation of *N*, judges that *C*-ing would conduce to compliance with *N*, takes this to be a reason for *C*-ing, and this judgment (partially) causes *A*'s *C*-ing virtue of its content.

The great bulk of cases of normative guidance, however, lack this explicit character. Indeed, in many cases of norm-guided behavior, individuals do not even form the belief that their conduct conduces toward norm-compliance. For example, we typically come to be guided by norms of language, conversation, and social comportment by an age when we could hardly form a clear idea of what these norms might be or how they might, taken together, apply in our circumstances. Even as adults, when our adroitness in being guided by these norms is nearly perfect, our knowledge of

[2] The appearance of 'appropriate' in (3) is needed in part to avoid deviant causal chains, though we won't pause to ask how this might be spelled out. For more substantive questions about appropriateness, see below.

them remains very imperfect. We need, then, an account of norm-guidance that makes room for this tacit sort of explanatory role.

Regulative Role

If 'appropriate explanatory role' need not be a matter of self-conscious judgment or an application of *N* in practical reasoning, can we nonetheless say something informative about what this role might amount to? In asking this, we should not lose touch with our ambition of recovering how the agent herself experiences and understands things. It will help, I think, to look at some more examples.

Fred

Fred is disposed to validate his ticket when riding the bus. His family and friends did so when he was a child, and he has followed their example as a matter of course. Indeed, his disposition is highly reliable, so much so that he confidently *expects* himself to validate, and is mildly surprised when he occasionally notices that he has taken his seat without having done so. Since the bus system in his city issues many special passes that do not require validation on each ride, neither bus drivers nor other passengers pay any attention to who has inserted a ticket in the stamping machine and who has not. What might lead us to say that Fred's ticket-stamping has the right kind of explanation to constitute norm-guided behavior rather than mere habit?

Part of the answer comes when we see what happens on those occasions when he discovers that he has absent-mindedly boarded without validating. If, on such occasions, he thinks only, 'Funny, I don't usually do that', and continues the ride unperturbed, then ticket-validation would appear to be a habit. That is how I am about kitchen cupboards. I am disposed to leave cupboard doors and drawers open, and do so with such regularity that I am mildly surprised if I happen to notice that I have closed everything back up. Surprised, but not discomfited. In such cases, I think only, 'Funny, I seldom do that', and do not feel the least impelled to return to the kitchen to carefully set several doors or drawers ajar.

Fred, however, *does* feel discomfited upon discovering that he is riding without validating. Moreover, even if he can on a given day ignore this mild discomfort or mitigate it by rationalization, still, what matters is that he feels this discomfort or need to rationalize, and that this discomfort, unlike many others, has a sure remedy. All Fred need do is to make his way back to the machine and stamp his ticket. Unlike me, then, Fred tends to treat departures from his usual practice as calling for *correction*. Similarly, Fred will show persistence and effort when crowding on the bus prevents him from reaching the validating machine as he boards. Fred will watch for his chance and squeeze his way between his fellow riders to reach the machine. By contrast, if I find myself in a kitchen with self-closing cabinets, I make no effort to prevail against them. These various ways in which Fred's response to departures from his normal conduct differs from mine suggest a *regulative* explanation of his conduct, as opposed to explanation by the *regularity* of a habit.

What is a regulative explanation? For an engineer, a *regulator* is a device with a distinctive functional character. One component continuously monitors the state of a system—the *regulated* system—relative to an externally set value, e.g. temperature, water pressure, or engine velocity. If the system departs from the set-point value, the monitor sends an 'error signal' to a second component, which modulates the inputs into the system—e.g. electricity, water, or fuel—until the set-point value is restored. The error signal then ceases, and the modulation of inputs stops. A simple example is the home thermostat. Regulative explanations of action deploy what is in effect the structure of a regulator, involving some form of self-monitoring for conformity to a standard or aim, departures from which cause the agent to make corresponding alterations in her course of thought, amount of effort, or direction of action to attain compliance—at least, insofar as possible. We might, then, add to (3):

> (4) Agent *A*'s conduct *C* is guided by norm *N* only if *C* is a manifestation of *A*'s disposition to act in a way conducive to compliance with *N*, such that *N* plays a regulative role in *A*'s *C*-ing, where this involves some disposition on *A*'s part to notice failures to comply with *N*, and to feel discomfort when this occurs, and to exert effort to establish conformity with *N*.

As before, we must not be too strict in how we interpret these conditions. For example, the process of noticing departures and making corresponding adjustments can be imperfect, and need not occur at the level of self-conscious awareness. Social psychologists, for example, have observed the tendency of individuals when being interviewed for a job to make rapid, unnoticed adjustments in posture, position, and voice volume that mirror the comportment of the interviewer (Davis, 1982).

Condition (4) needs further refinement, however, in order to discriminate Fred's ticket-validating from another disposition of Fred's. He is disposed to purchase a snack on his way to work in order to have it on hand for his mid-morning break. He does so reliably enough that he expects this of himself. And if he discovers at break-time that he has failed to purchase a snack that morning, he's annoyed with himself and treats this as something to be remedied—by resort to the office's wretched vending machine if need be. Is Fred's conduct guided by a norm of snack-buying or against snackless mornings? To be sure, there is a norm of prudence at work in the background—he's learned that without a snack he's usually uncomfortably hungry well before noon.

But observe what happens when, on a particular Monday morning, Fred is so engrossed in the project he's completing that he forgets to buy a snack on the way to the office and works right through the morning without let-up. He does not notice his failure to purchase or eat a snack until a co-worker pokes her head into his cubicle at noon to suggest that they go over to the canteen for lunch together. Fred does not regard this failure as something that calls for correction. Instead, he thinks only, 'Funny, I didn't even notice'. Call this a *non-consequential* and *unsanctioned* failure to fit his standing behavioral expectations. Non-consequential because, as it happened, he suffered no ill effects from the omission;[3] unsanctioned because no authority would take any interest in his missed snack, or impose any penalties.

To purchase and consume a mid-morning snack is not mere habit for Fred, nor is it a personal norm. Rather, it is a daily routine acquired for its instrumental value. Let's call such routines *default plans*. Plans, like norms, bring regulative structures into play—we are disposed to monitor our progress toward carrying out our plans, to notice departures from plan, and to adjust action accordingly. But plans and policies are of many kinds, and the agent need not see a departure from plan or policy, if otherwise non-consequential and unsanctioned, as warranting any criticism, correction, or self-reproach.[4]

[3] The existence of *actualist* consequentialisms—as opposed to *expected value* versions of consequentialism—makes formulating this intuitive idea a delicate matter, since the very fact of whether actual-consequentialist norms are violated is a matter of how things turn out. For such normative conceptions, we need to distinguish those phenomena within the purview of the norm (e.g. welfare effects), and those not.

[4] Even though plans, like norms, involve regulative structures within the agent, there are quite general reasons for distinguishing plans as such from norms. This difference is most clearly manifest in the feature adverted to above, namely, that agents typically respond differently when they realize they have violated a norm they hold vs. deviated from a plan they have made. Similar considerations serve to distinguish norms from personal policies or strategies. Two individuals with the same norms and values can differ in their plans or personal policies, and, indeed, it can be a criticism of someone that her plans or personal policies are not consonant with her norms. Despite the difference in attitude between planning and treating as a norm, it is possible to spell out the implications of a norm for action—its 'practical extension', as it were—in terms of a plan specifying indicated actions for all possible contingencies. For a seminal discussion of plans, policies, and self-regulation, see Bratman (2000). For a philosophically illuminating use of plans in providing a systematic treatment of norms, see Gibbard (2003).

Contrast Fred's reaction when, fishing in his pocket for change that Monday in the lunch line, he finds an unstamped bus ticket and realizes that in his distraction he also failed to validate his bus ticket on the way to work. That failure, too, has turned out to be non-consequential and *de facto* unsanctioned. Although he's glad his free riding went unnoticed by anyone—at least, he hopes it did—he still sees himself as having done something that warrants criticism, and finds himself cooking up a quick mental rationalization ('I'd have to take a month of free rides to make up for all the perfectly good tickets I've lost or ruined in the laundry'). Fred thus manifests his sensitivity to pressures of *consistency* in thought and action with respect to *N*. Such pressure is characteristic of norm-guidance in cases where the agent is at least tacitly aware of the norms at work, for example, Fred's 'Pay your way' norm or Martha's 'Respect privacy' and 'Preserve confidentiality' norms.[5] Fred's feelings of discomfort and defensiveness can be thought of as self-imposed *internal sanctions* for the bare fact of norm violation, considered independently of other effects.[6]

Interestingly, such pressures for consistency can be triggered and felt even when the norm of the agent in question is one of which she herself is unaware. One intriguing piece of evidence for this is the phenomenon of 'over-regularization' in children's speech. As their linguistic ability develops, some children who have previously mastered the past tenses of irregular verbs begin 'correcting themselves' by forming irregular past tenses using the <verb stem + -ed> rule for regular verbs, for example, saying 'go-ed' instead of 'went'. This occurs despite the fact that these children have never heard 'go-ed' spoken by adult speakers, and have never been sanctioned for using 'went' as the past tense of 'go'. As adults, we feel similar pressures toward consistency in language use. We can sense that grammatical anomaly is creeping into a sentence we are uttering, and struggle to correct ourselves on the fly. We treat such anomalies as mistakes, even when they have no effect on—or even improve—sentence intelligibility, and even when we would be at a loss to identify the particular incompatibility with grammatical rules involved.

[5] A further manifestation of this pressure is the tendency of agents to sincerely *avow* or *endorse* *N* in unconstrained normative discussion. Gibbard has drawn attention to this feature in the context of norm-acceptance (see below). See Gibbard (1990: 74–82).

[6] As before, we are ignoring cases in which departure from *N* is due to guidance by another norm, taken to be weightier or more relevant. In such cases, deviation from *N* need not be accompanied by a sense that correction is called for, since relative normative priority explains and excuses the deviation. Notice, however, that even in cases of excused violation a felt need for correction can persist. For example, if attending to an urgent student need makes one late for a regular lunch engagement with a colleague, one will typically feel that explanation and apology are called for.

To distinguish norms from default plans, we'll try (5):

> (5) Agent *A*'s conduct *C* is guided by norm *N* only if *C* is a manifestation of *A*'s disposition to act in a way conducive to compliance with *N*, such that *N* plays a regulative role in *A*'s *C*-ing, where this involves some disposition on *A*'s part to notice failures to comply with *N*, to feel discomfort when this occurs, and to exert effort to establish conformity with *N* even when the departure from *N* is unsanctioned and non-consequential.[7]

Does (5) need supplementation because normative guidance involves a distinctive set of emotions, such as guilt, pride, shame, or reproach? While moral norms in particular are associated with such emotions, most norm-guidance is non-moral. An agent need not feel any guilt or shame when she discovers a typographical error while proof-reading a letter before mailing it off. She might feel annoyance, relief, or nothing at all beyond the familiar, minor dissatisfaction with the status quo that accompanies the discovery of one's lesser errors, and that typically persists until the errors are corrected or forgotten.

Mental Acts and Attitudes

Thus far, we have developed only a partial, largely functional characterization of the conditions a piece of behavior must meet to be norm-guided. (5) could stand a good deal of work, but perhaps it is sufficiently suggestive of the distinctive *role* of norm-guidance in an agent's psychology to enable us to move on to our next question: What mental act or state of mind on the part of an agent gives a norm this sort of role in her life? To revert to our original image: in a portrait of the agent 'from the inside out', what attitude on her part brings a norm to life in how she thinks and what she does and feels? As one might expect, this question has no single answer—a norm can play the role suggested in (5) for a variety of reasons. Let us consider two candidate answers that have figured in the recent literature: accepting *N* and endorsing *N*. To portray norm-guided agency accurately, we need to identify the distinctive place of each of these attitudes in the complex phenomenon of normative guidance, and to ask whether they suffice to give a comprehensive account 'from the inside out'. Let's consider them in turn.

[7] Not every case of normative guidance will display all the features in (5). For example, *A* might notice his departure from *N* and be moved straightway to make a correction, but experience no particular discomfort.

Acceptance

An agent's acceptance of *N* might seem to be the least restrictive answer to the question of how a norm comes to play an action-guiding role for him. It is intuitively plausible to say that Fred accepts 'Pay your way' as a norm and Martha accepts 'Respect departmental confidentiality' as a norm, that Fred accepts 'No snackless mornings' as a default plan rather than a norm, and that I do not accept 'Leave cupboards open' as either a norm or a plan. This difference in attitude would naturally translate into the relevant differences in thought and action. But what is it to accept or fail to accept a norm?[8]

We might simply point to a role-characterization such as (5), and treat it as supplying the 'job description' of norm-acceptance: for *A* to accept *N* just is a matter of *N*'s playing a (5)-like role in shaping *A*'s individual actions. However, inquiring minds will want more insight into *A*'s psyche than this affords. What does *A* do, think, or feel that brings this about?

In the paradigm of norm-acceptance, *A* reflectively considers norm *N* and freely decides to treat it as action-guiding or reason-giving. Explicit acceptance of this kind has the virtue of offering the beginning of an account of a norm *N*'s *authority* for the agent. She herself has decided to treat an act's conformity with *N* as counting in favor of performing that act—other things equal, as always. The source of the authority invoked here is liable to two readings. On a *voluntarist* reading, what matters is simply that *A* is the free author of the decision to hold herself to *N*, so that *N*'s action-guiding role for her is self-imposed. On a *judgmentalist* reading, *A* determines whether compliance with *N* is worthy, required, or otherwise appropriate on the basis of grounds she takes to be independent of her will—e.g. intrinsic values or rules of logic and evidence. For the judgmentalist, the source of *N*'s action-guiding authority for *A* is not rooted in her decision to accept *N* alone, but in the grounds of that decision.[9]

Since only a small portion of the norms of thought, language, behavior, and culture we have acquired since youth owe their regulative role in our conduct to reflective acceptance, we must appeal to hypothetical or tacit acceptance to account for the majority of cases of norm-guidance.

[8] Gibbard (1990: ch. 4) offers a characterization of norm-acceptance that differs somewhat from the account that follows.

[9] Does the judgmentalist account suffer the disadvantage of depending upon some further source of authority, namely, the grounds of the judgment, which cannot also be the upshot of judgment? However, the voluntarist account can equally be said to depend upon some other source of normative authority, since if one attributes no authority to oneself initially, one's acts of will could hardly confer such authority upon themselves or their outcomes. These questions will be discussed further below.

This requires, however, that we identify these forms of acceptance with actual states of mind capable of playing a regulative role in explaining an agent's conduct. It is not difficult to imagine how we might fill out the description of Fred's dispositions given above—we have, for example, said nothing about how Fred is disposed to view other sorts of situations, or the conduct of other individuals—in such a way that it is plausible to attribute to him tacit acceptance of 'Pay your way' as a norm, even if he has never formulated the norm as such, or given the matter reflective thought. Similarly, it seems plausible to say that Fred tacitly accepts 'Purchase and consume a mid-morning snack' as a default plan, even though he has never bothered to formulate any explicit plan to this effect.

Tacitly accepted norms and plans can regulate an individual's conduct in various ways. Whether we recognize them or not, the norms we hold and plans we make are reflected in the ways we *frame* our practical situations, much as our beliefs—including legions of tacit beliefs—function to frame our epistemic situations. Such framing is a matter of the expectations one brings to situations, the features of situations one tends to notice or ignore, the spontaneous interpretations of events one is primed to make, the possibilities for thought and action that come immediately to mind, and so on. If a 'Pay your way' norm frames how Fred thinks and acts when boarding a bus, then he will validate without giving the matter any thought. If a 'Preserve departmental confidentiality' norm frames Martha's conversations with Rick, then certain topics will or will not occur to her simply as a matter of course.

Frames do their job, of course, precisely because they function like a camera frame. They limit the otherwise unbounded and undelimited character of experience and restrict one's scope of attention—not because one *sees* the frame, but because what one sees is seen *through* it. Frames define a situation in a way that enables an agent to avoid distraction and focus selectively—Fred on finding a free seat or bit of hand-rail as he boards the bus, Martha on the content of what Rick is saying and what she herself wants to chip in.

Does this degree of 'automaticity' and lack of self-aware acceptance and application of a norm deny it the role of furnishing the agent's *reasons* for acting? Given what we know of Martha, it seems appropriate to say that she invites Kim out to dinner *out of concern to* express her gratitude to Kim for hosting her, or *for reasons of* reciprocity—not, for example, to curry favor with Kim or show off her newly acquired income. Similarly, it seems appropriate to say that she does not raise certain topics in talking with Rick *out of respect for* confidentiality and for a junior colleague's sensibilities—not out of distrust of Rick's discretion or fear of criticism by colleagues. We will not understand how Martha sees her situation until we understand the ways in which tacitly held norms shape her thought, experience, and initiatives, without being called to mind. Indeed, we will not understand how a small child sees his situation when saying 'go-ed' until we see that he acts *out of concern to* speak properly—not by simple mistake or owing to a mindless conditioned response.

None of us, presumably, would be able to formulate all the norms at work within us in a given situation, or give a detailed account of how they interact. We often discover what norms we hold only indirectly, from seeing how we react to another culture, changed life circumstances, personal emergencies, and even long-sought successes.

Acceptance and Belief

Unless we can say something more substantive about the 'interiority' of acceptance, however, our invocation of tacit acceptance runs the risk of identifying our *explanans* with our *explanadum*. We would short-circuit any effort to gain understanding of normative guidance 'from the inside out' by equating tacit acceptance with whatever-state-of-mind-it-is that underwrites a regulative role or practical framing effect.

Acceptance is, after all, a distinctive state of mind, often contrasted with belief. And yet nothing we have said about the manifestation of tacit acceptance would enable us to distinguish tacit acceptance that *p* from tacit belief that *p*. To help focus our thinking, let's turn briefly to uses made of the acceptance/belief distinction in other domains. In the philosophy of science, for example, Bas van Fraassen has drawn on this distinction to develop and defend a doctrine of Constructive Empiricism. Critical of the metaphysical *braggadocio* of the Realist, whom he sees as advocating outright belief that our going scientific theory is true right down to its latest claim about unobservable quarks, van Fraassen has developed an alternative. According to Constructive Empiricism, the appropriate attitude for scientists toward the dominant theory as a whole is acceptance rather than literal belief. Literal belief is to be reserved for the theory's claims about observables, while the remaining theoretical apparatus is to be *used* (not believed) for purposes of inference, hypothesis formation, experimental design and interpretation, explanation, and so on.[10] In another area, epistemology and

[10] Van Fraassen writes (1980: 88): 'While the only belief involved in acceptance, as I see it, is the belief that the theory is empirically adequate, *more than belief is involved*. To accept a theory is to make a commitment, a commitment to the further confrontation of new phenomena within the framework of that theory, a commitment to a research programme, and a wager that all relevant phenomena can be accounted for without giving up the theory.'

decision theory, some philosophers whose official doctrine holds that belief properly so-called is a matter of degrees of credence, have nonetheless found restricted uses for an attitude of acceptance (e.g. for statements that pass a contextually determined threshold degree of credence) to analyze cases in which everyday decision-making or informal reasoning call for a univocal up-or-down judgment.[11]

The doxastic attitude of acceptance is most commonly distinguished from belief in the following ways. (1) Although acceptance, like belief, can arise spontaneously, acceptance is much more amenable to volition and purpose, and hence more directly subject to decision. We do sometimes speak of *deciding whether to believe p*, but this is equivalent to *making up our mind whether p*. That is, the focus is on the question *whether p*—whether *p* is supported by the balance of evidence, intuitively plausible, etc.—while ignoring collateral effects attributable to the state of mind of *believing that p*. In contrast, *deciding whether to accept p* often is not equivalent to *making up our mind whether p*, and the decision typically focuses not only on *whether p*, but also on the costs and benefits of accepting or failing to accept *p* in the present context, many of which enjoy some independence from *p*'s truth. For example, it often is more important to have *some* answer to a question than to have *the* answer. To put an end to time-consuming quibbling over a small matter, such as who owes whom a few dollars, two friends might simply accept that things somehow have balanced out, and proceed accordingly. Other times, it is more important or efficient to accept someone's word at face value rather than dig around suspiciously to try to get at the truth oneself. Thus a manager faced with a damaged piece of office equipment can decide to accept an earnest new employee's rather elaborate explanation and carry on, since refusal to give the employee the benefit of the doubt would create an atmosphere of distrust. (2) Acceptance can be context-specific in ways that belief resists. A jury, having heard the testimony of a key witness for the defense and the prosecutor's feeble attempt to present disqualifying evidence, can unanimously decide to accept the witness's account as given—even though a number of the jurors sensed something odd in the witness's manner, and remain personally unconvinced about whether she is telling all she knows. Although these jurors can accept the witness's account of the facts as given for the purpose of reaching a verdict, they cannot similarly *believe* it for that purpose. Thus (3), acceptance is not subject to the same pressures of cross-contextual consistency and 'total evidence' as belief. The individuals who accept the witness's testimony as jurors deciding a case can reject it in

[11] For discussion of the nature of belief vs. acceptance, including disputes about the tenability of the distinction, see the contributions to Engel (2000).

whole or part as individuals offering their personal opinion as to what really happened.

These features reflect a fairly deep fact about belief vs. acceptance. Belief by its nature resists the self-aware instrumentalization and contextualization that acceptance freely permits. This is sometimes put, a bit misleadingly, by saying that 'Belief aims at truth'.[12] Acceptance, by contrast, tolerates quite diverse aims. It also tolerates quite diverse objects. Invitations, proposals, and commands can be accepted, but I'm not sure what it would be to believe them. Correspondingly, we resort to the vocabulary of belief when we wish to express emphatic trust or faith. The faithful *believe in* God and salvation, and the apt title for a *credo* is *This I Believe*, not *This I Accept*.

Although different from belief, acceptance normally depends upon belief in various ways. The manager deems acceptance of the employee's explanation appropriate because she believes that the equipment is genuinely broken but not sabotaged, that the employee is trying his best, and that a sign of trust on her part would be encouraging. The jurors deem acceptance of the witness's testimony appropriate in reaching their verdict of 'Not guilty' because they believe that the witness's testimony is inconsistent with locating the defendant at the scene of the crime, that the prosecutor clearly failed to discredit her testimony or otherwise meet the burden of proof, and that the judge instructed them to follow the rules of evidence and deliver a verdict accordingly, setting personal opinions or suspicions aside. Decisions to accept are like any other decision—they depend upon what one believes and seeks. Appeal to belief in justifying acceptance need not launch a regress, because we acquire most of our beliefs, as well as our evidence for and against them, from experience and inference, without need of any decision to accept them.[13]

If acceptance contrasts with belief in the domain of factual judgment, is there an attitude that similarly contrasts with acceptance—i.e. is 'belief-like'—in the normative realm?[14] If so, which attitude seems more appropriate for analyzing the examples of normative guidance discussed thus far? And does norm-acceptance, like doxastic acceptance, depend upon

[12] Interpreting this dictum is a complex matter. For some discussion see Humberstone (1992), Railton (1994), and Velleman (2000).

[13] It might be argued that belief does depend upon acceptance at a deep level: to believe, we must accept our own authority. While I agree that belief would not be possible if we rejected our own authority, I consider the attitude here to be *default trust* rather than acceptance: we trust our eyes, our memory, our reasoning. For discussion, see Railton (2004).

[14] E.g., Timmons and Horgan (forthcoming) introduce a form of normative belief (an 'ought-commitment') meant to parallel ordinary factual belief (an 'is-commitment') and capture the idea of a normatively engaged viewpoint. (Commitment, arguably, is different from acceptance.)

a contrasting belief-like attitude? Let us consider some cases in which the language of acceptance seems particularly appropriate, to help us identify what a contrasting state, if any, might be.

Felicity

Fred's boss Felicity comes from very modest southern Appalachian origins. She attended an expensive New England college on scholarship, and there she came to believe that being taken fully seriously and achieving the success of which she is capable depend upon her ability to overcome her twang and self-effacing rural manner, and generally learn to comport herself in accord with the Upper Middle Class Professional norms. She's been remarkably successful at this, and UMCP personal comportment has become second nature to her. She does not regard this as a betrayal of her own background, and happily reverts to many of her old ways when back home with family. To her, UMCP norms are just another way of comporting oneself—her second language, in effect, to be spoken in the community which is now her adopted home. UMCP norms are not at all contemptible in her eyes, nor do they seem to her incompatible with her own core values and family identity.

Josef

Josef is a conscientious utilitarian who has concluded that it would be much better for people to observe classic Lockean 'side constraints' rather than engage in case-by-case felicific calculation. He knows that from time to time acting in accord with these side constraints will yield non-optimal outcomes, but he doesn't think he or anyone else is particularly good at spotting such occasions, or at applying fully utilitarian reasoning properly when they try. So he has cultivated a strong disposition to follow and commend Lockean side constraints in virtually all situations, largely ignoring temptations toward utilitarian ways of thought. He has not abandoned or forgotten his underlying utilitarian convictions, and he will tell anyone who is interested that he thinks Lockeans miss a bet by failing to realize that the strongest argument for side constraints is based on utility, not specious 'natural rights'.

It seems to me accurate to say that Felicity *sincerely accepts* UMCP norms as action-guiding in most of her professional, public, and private life. Hers is no hypocritical or reluctant pretense, and she does not see UMCP comportment as essentially shallow or pointless. When in her professional milieu, she takes an act's conformity with such norms to be a perfectly good reason for her to perform it. Similarly, Josef *sincerely accepts* Lockean side constraints as action-guiding norms. He, too, is making no pretense. He believes following these norms to be a good way for him to be, and that he and others have ample reason to act as they require.

At the same time, it seems to me accurate to say that Felicity and Josef retain an attitude toward their acquired norms short of outright *belief* in them, unlike the attitude of a dyed-in-the-wool UMCP snob or a partisan Lockean. Although the vast bulk of their daily conduct is regulated by these norms directly, without detour through instrumental reasoning, still, their attitude toward these norms remains fundamentally instrumental. Purpose apart, they see no particular reason to comply with them. But precisely because their attitude toward these norms is one of acceptance rather than a personal *credo*, sincerity on their part is quite compatible with instrumentality and contextual limitation.

Norm-acceptance in these cases, as in the examples of doxastic acceptance discussed above, is underwritten by commitments with a belief-like character: Felicity's belief in herself—her commitment developing her talents to the fullest—and Josef's belief in a utilitarianism as the proper standard of right action. Felicity and Josef do not treat these underlying commitments as action-guiding for any *further* purpose, or with respect only to certain particular contexts.

Endorsement and Identification

To characterize the difference between norm-acceptance and a more 'belief-like' normative commitment, we must get closer to the center of the agent. It is natural, then, to look to what the agent *endorses* rather than merely accepts. Consider, for example, the difference in character between a statement issued by the losing side in a lawsuit that they *accept* the court's decision vs. a statement that they *endorse* the court's decision. The former conspicuously makes room for a certain distance between the views of the interested party on the merits and the view of the court, while the latter closes this gap considerably.

Although distinct from acceptance, endorsement does share with acceptance a potential defect as a candidate to be a belief-like attitude. For it too lends itself to contextualization and instrumentalization. I can endorse Smolenski as a candidate for State Senate, but not for Governor, or endorse her in the Republican primary for the purpose of keeping a Republican demagogue off the ballot, while endorsing her Democratic opponent in the general election. To identify a more belief-like attitude, we should focus on endorsement of a norm in itself, and not 'as a means alone'. To my ear, a difference remains between acceptance and endorsement even when we compare acceptance of a norm as such or without further purpose with endorsement of a norm as such or without further purpose. This difference seems to me to count in favor of endorsement's claim to be more belief-like. I am free, for example, simply to accept my basic aesthetic tastes as such, without having any sense that I am qualified to endorse them or that they possess any particular warrant or credibility. Endorsing my basic tastes, by contrast, is a *judgment*, and in pronouncing it I take myself to have some claim to evaluative standing. Similarly, if an Alpine guide for a mountain range unknown to me indicates which path leads to the shelter for the night, I can readily accept her selection as such, though it would seem odd or presumptuous for me to say that I endorse it.

I might, however, without presumptuousness, *endorse accepting* what the Alpine guide selects. I am in a position to judge that she is more informed on the matter than I am. Thus, endorsement might help make sense of acceptance, and thereby help us build up a normative portrait of an agent 'from the inside out'. Josef, for example, has given the foundations of ethics much thought, and can be said to endorse a utilitarian norm of conduct in its own right, while being critical of natural rights theory. This endorsement is for no further purpose, and so the utilitarian norm lies close to his center as an agent. He also endorses purposes consonant with this norm, and thus *endorses accepting* Lockean norms insofar as these promise to serve such purposes, even though he certainly does not endorse them in themselves. Such norm-guided acceptance of norms places these latter norms at one remove from his center as an agent, and yet his acceptance of them need not be alienated or insincere. It has a secure, albeit delimited, place in his normative scheme. But can we say more explicitly what this attitude of endorsing a norm as such is like?

As with acceptance, a paradigm is afforded by *reflective* endorsement. Recall that reflective acceptance involved an agent considering a norm *N*, and then freely deciding to treat it as action-guiding or reason-giving. We noted that voluntarist and judgmentalist readings of 'deciding' in this formula were equally viable. For reflective endorsement, however, the judgmentalist reading seems to be favored. There is a difference between endorsing *p* and simply fixing on *p* by an act of will, much like the well-known difference between choosing and picking. Endorsement is ordinarily understood to be an evaluative judgment, and thus to involve reasons or grounds, so that the source of normative authority of reflective endorsement lies in part with these reasons or grounds. On pain of circularity or vacuity, however, these reasons or grounds cannot be brought into being by the agent's endorsement. Endorsement, then, cannot be the sole occupant of the agent's normative center. It requires an environment of grounds, reasons, or values in order to come into being and guide action. The resulting picture need not be foundationalist. An agent can start with certain presumptive grounds, reasons, or values taken for granted, and then begin judging, endorsing, choosing, acting, and, over time, revise her starting point according to what she thereby has learned from life thus far. An agent conceived as beginning such a learning process with no presumptive normative 'priors' of this kind would, however, be at a loss to make endorsements or choose paths other than by simple plumping.

Of course, Kantians and neo-Kantians might be right that certain norms are rationally necessary a priori in the sense that they are a condition for any sort of agency, and would win the endorsement of any reflective rational agent as such, regardless of her starting point. What might underwrite such 'presuppositionless' endorsement, such that it can have the character of a genuinely evaluative judgment? Here are two possibilities. (1) Perhaps, as Kant suggests, it is grounded in something self-evident: our inability even to conceive anything that is good without qualification other than a good will. Here we have found a ground, but one that commands respect directly, without need of judgment.[15] If so, then a conviction of the unqualified goodness of a good will is a necessary normative *credo* lying at the center of rational agency. It explains, rather than being explained by, our endorsing judgments.[16] (2) Alternatively, this endorsement might be claimed to arise from the fact that the agent *identifies* with her rational nature as such. But, as Harry Frankfurt (1988) has observed, identification is not a form of endorsement, and need not have as its condition any endorsing attitudes.[17] Felicity identifies with the norms of personal comportment and sociability of her rural Southern family home rather than UMCP norms, but not because she deems them in any way more choiceworthy—rather, simply because this particular Appalachian setting is where she grew up, her social

[15] Kant writes: 'Respect (*reverentia*) is, again, something merely subjective, a feeling of a special kind, not a judgment about an object that it would be a duty to bring about or promote' (1996: 6. 402).

[16] Contrast Christine Korsgaard's remark: 'In the end, nothing can be normative unless we endorse our own nature, unless we place a value upon ourselves' (1996: 165). An agent who did not already have some *ground* for endorsement or some *sense* of his own value—e.g. in light of sensing the unqualified value of the humanity and moral law he finds within himself, as Kant puts it—would be unable to make an endorsing judgment or *confer* value.

[17] Might *identification with N* furnish the true core of the agent's normative structure? The question is too large to discuss here, but we might note in passing that even identification seems to require a prior answer to the question *who* is doing the identifying, and *how*. Fundamental as it certainly is, identification cannot, it seems, stand alone at the center. A realistic portrayal of the agent as a whole suggests the same conclusion. As the examples of Hal and Ed, below, will suggest, there can be elements that contribute importantly to who I am and what reasons I will recognize or respond to—my real practical identity, so to speak—with which I do not identify.

equivalent of a primary language. If these norms are closer to her center as an agent, that is not to be accounted for by an attitude of endorsement.

Endorsement also faces another difficulty as an account of the center-point of an agent's normative structure. For judgments do not necessarily motivate, and motivation is required for norm-guidance. It is, however, plausible to maintain that judgments of *endorsement* belong to a class of judgments that do have an 'internal' connection to motivation. Even so, other motivational forces at work within the agent can limit the role of endorsement in her overall psyche and conduct. These other springs of motivation might well be closer to her psychic center, and able to operate without benefit of her endorsement or awareness. We thus have an apparent conflict between an individual's judgmental center and her psychological center as a human being. It might seem easy to identify which center is critical for norm-guided *agency*, or action for a reason: the upper, judgmental center. But this easy answer turns out to depend upon a limited conception of rationality, norm-guidance, and autonomy.

Rationality, Norm-Guidance, and Autonomy

On one conception, rationality is a capacity for *reasoned decision* and *judgment.* To find rationality in action we look for agents engaged in practical deliberation, treating considerations as reasons to act and setting themselves to act accordingly. On another, broader conception, rationality is a capacity to be *aptly responsive to reasons*, which may involve a large variety of non-deliberative processes.

In the domain of theoretical reason, for example, individuals can be aptly responsive to sensory evidence by directly trusting their eyes and non-inferentially forming perceptual representations—much as animals do. Calling this process non-inferential by no means denies the visual system's great computational complexity. It simply registers that such computation is sub-agential, and not of the sort we ordinarily identify as reasoning. Similarly, the non-inferential, self-evident 'intuitions' that figure in the foundations of logical and mathematical thought involve highly complex cognitive representations and associations, but are thought to underwrite, rather than require, inferential reasoning.

In the domain of practical reason, individuals can be aptly responsive to risk through arousal of fear, even when the fear is not recognized as such by the agent.[18] More broadly, individuals can be aptly responsive to moral and

[18] See Bechara *et al.* (1997).

prudential considerations, even in the face of contrary self-aware judgments, through the emotional impact of sub-personal, empathetic simulations of the internal states of others or of one's future self.[19] Again, such processes are highly complex computationally and cognitively, but are not forms of reasoning in the canonical sense. Rational agency, conceived broadly, is not located exclusively in the judgmental core, but distributed over the larger psyche and physiology of the human individual. To understand rational agency in the broad sense 'from the inside out' we must start not at the seat of reasoning, but at the center of mass of the person as a whole—the center of a constellation of desires, drives, emotions, moods, experiences, images, thoughts, values, expectations, associations, dispositions, sensibilities, and commitments that take shape over a lifetime. Taken together, they comprise the many ways in which the agent's psyche and its embodiment equip him to be responsive to reasons, with or without the blessing of his judgmental or reasoning self.[20]

Once we thus broaden our optic on rationality, we can see that there is a certain falseness to the familiar contrast between action guided by norm-based judgment and action guided by feeling or emotion. Norms can exert regulative influence on thought and action only through the attitudes we hold toward them, the dispositions and feelings they shape, and the motivations they engage; and complex emotions find their distinctive character and expression thanks to an agent's acquired concepts, norms, and cultural understandings. Kant, for example, tells us that respect for humanity is a 'subjective feeling' indispensable for proper responsiveness to, and incentive toward, claims of duty; at the same time, this 'moral feeling' has as its defining object and form of expression action of a normative character: self-imposition of the moral law (Kant, 1996: 6. 399–402).

Let us, then, look at two examples in which norm-guidance and reason-responsiveness occur, but which encourage us to think in terms of the broad conception of rational agency.

Hal

Hal is the chairman of large department in the humanities. Two colleagues, an anthropologist and a historian, are surveying the scene at a crowded college gathering with mild interest when they notice that their friend Hal is showing exceptional bustle, circulating briskly among his department's junior faculty, who are scattered around the room in various tight conversational

[19] See e.g. Gordon (1995) and, on the experimental side, Ruby and Decety (2001).

[20] For recent work on other ways the emotions and other non-deliberative phenomena contribute to our responsiveness to reasons, see various essays in Hatzimoysis (2003) and Solomon (2004), and also Arpaly (2003) and Railton (1997, 2004).

clusters. Hal comes up from behind abruptly, places an unannounced hand on his junior colleague's shoulder, says a few quick words, and departs, barely pausing to catch any reply. 'Looks like a bumble bee gathering nectar,' comments the historian. 'No, it's too social for that,' the anthropologist replies, 'See he how lays his hand on their shoulders, interrupting their conversation and commanding immediate attention? It's a sign that he's the head-man, their superior in the tribe. And notice the recognition he gets—they turn to him right away, flash their eyebrows, and give him their full attention.'

The historian knows Hal to be a fairly unreconstructed 1960s progressive who often argues for egalitarian power-sharing at faculty meetings. She has just launched into her alternative explanation—'Hal's just a very tactile, gregarious guy and . . .'—when the Dean, a notoriously aloof woman, strides up behind Hal and places a firm hand on his shoulder. Not missing a beat, Hal pivots away from his conversation to face her, eyebrows raised, attention fixed. She concedes her anthropologist colleague's point without further argument.

The agents observed in this little drama, one could say, have internalized a distinctive norm concerning hierarchy, physical contact, and the ability to interrupt with impunity and command attention. Would Hal, the Dean, or his junior colleagues have endorsed this norm if the question had been put to them in another context? Very likely not. Yet a shared norm of this kind played an indispensable role in supporting the smooth choreography of their motions, the efficiency and rapidity of their exchanges, and the absence of any ruffled feathers. If Hal had tried striding up behind the Dean unannounced, and placing his hand firmly on her shoulder to interrupt her conversation, or if an undergraduate had likewise accosted one of Hal's junior colleagues, the reception would have been startled and decidedly cool—feathers definitely ruffled and eyebrows narrowed, not flashed.

Shared, internalized norms, and the expectations, motivations, and feelings they shape, govern many aspects of our social interactions. They resolve countless questions of comportment and conduct that would otherwise be unsettled, and impede the nearly automatic functioning of our lives together. Like norms of language, they serve not only for coordination, but also communication. They make it possible for particular actions to carry certain meanings rather than others, even when the norms in question would not be accepted by those involved. Hal, a long-time egalitarian, would not endorse a hierarchical norm governing contact and interruption. Yet in virtue of having internalized it, and belonging to a community where others have internalized it as well, he has been able in the various stages of his career not only to comport himself in ways appropriate for his position in the hierarchy at the time, but also in ways that communicate messages quite apart from relative hierarchical standing: feelings of respect for those he admires (by comporting himself toward them as he would someone higher in social ranking), and fellow-feelings for those he views as equals (by placing a hand on the back rather than the shoulder, waiting for the right moment rather than presuming to interrupt at will). Indeed, thanks to this widely internalized norm, Hal in his radical days was able to communicate actively anti-hierarchical sentiments by counter-normative accosting of figures in authority. Given shared norms and relative standing, Hal's placing a hand on a junior colleague's shoulder carries no 'news value' and shows no special intent, good or ill. It is what anthropologists call 'unmarked behavior'. Contrast the 'marked' character of a similar gesture made by a stranger or by an inferior in the hierarchy—it would be 'news', and immediately pose the question of what the intent might be.

Still, one might say, norms that have merely been internalized, and are not recognized or accepted by those who follow them, do not enable us to see the action from the standpoint of the agent, to grasp *her* reasons for action. The operation of internalized norms can provide a third-personal or anthropological explanation of conduct, but what has this to do with recovering the agent's point of view?

Just as we can think of an agent's rationality broadly, we can think of her agency broadly—thereby adding greater psychological depth and realism to our portrayal of her 'lived world' and its meanings. Consider Hal's conduct at the gathering:

(*a*) Are his touchings of shoulders to interrupt junior colleagues, or his pirouette to hear what the Dean has to say, intentional? Yes.

(*b*) Does an internalized hierarchical norm of contact and interruption play a role in a regulative explanation of this behavior? Yes.

(*c*) Does Hal, or would he if asked, endorse the hierarchical norm? No.

(*d*) Does the hierarchical norm help us to grasp how he saw his situation, and what he saw in it that recommended acting as he did? No, *and* yes.

Hal certainly did not see himself as following such a norm, and would be surprised to be told that hierarchical standing was even at issue ('I was simply trying to get some important news to my junior colleagues right away with a minimum of fuss'). But we won't understand his practical framing of the situation until we recognize how it was structured by relations of hierarchy, among many others. Like most of us, Hal *reads* social situations in complex ways, which assign an important place to hierarchy. In culturally familiar settings, he does not do this expressly and deliberatively, but tacitly and with remarkable speed. Hierarchical relations are among the objects of his immediate attention—though seldom his conscious thought—and begin to shape his behavioral dispositions, expectations of others, and sense of what is appropriate as soon as he enters a room. He is, moreover, highly responsive to this reading in his intentional conduct—it plays a significant role in explaining his actions, their manner, and the reasons for which they were performed.[21]

The demand to read situations in hierarchical and relational terms can arise even from language itself. Many languages, unlike English, require the speaker to elect either the formal or familiar form of the second-person pronoun whenever addressing someone. And a sociolinguist would have no difficulty identifying comparable formal/familiar markers in spoken English—matters of vocabulary, intonation, and form of address. For competent speakers, these linguistic shifts are as unthinking as the shift between 'he' and 'she' in response to differences in gender. Although conversational fluency requires that speakers at some level be attentive to hierarchy, familiarity, and gender, this does not preclude their using the language to criticize age or status hierarchies, or to challenge distinctions of gender. Still, even critics, or those who simply wish to take no side, often find it impossible to control fully the messages carried by their words, or to find words that convey their communicative intentions without carrying unwanted meanings. (Witness the instability of whether to elect 'he', 'she' or 'they' when speaking in the impersonal third-person singular in English.) Without shared norms, linguistic communication would not be possible; but for this reason we cannot by an act of personal will extract from our words and deeds all implicit messages with which might we disagree.

We will not faithfully represent the phenomenology of Hal's 'lived world' if we strip it of the framing effect of hierarchy. Of course, we might know independently that Hal is the sort of person to whom hierarchy and status matter greatly, much more than he himself acknowledges. Were this the case, we might have no qualm about saying that he acts in the manner he does 'for reasons of hierarchy', even if, when queried, he would sincerely disavow this. But we need not imagine Hal to be especially attentive or deferential to hierarchy in order to see how responsiveness to relations of hierarchy figures in his reasons for acting, partly explaining the favorable

[21] Compare: Hal's actions also served to give the anthropologist a good example of his theory. Was this one of his reasons for acting as he did? 'Providing a good anthropological example' in no way structured Hal's practical framing of his situation—he did not read the situation in such terms, nor did they shape which actions struck him in a favorable light, which actions he did not consider, why his action felt comfortable and had no special significance to him as he performed it, or what he expected his junior colleagues' response to be.

light in which he sees certain acts, and the unfavorable light in which he sees others. If we wish to give a portrait of Hal 'from the inside out' as he flits from colleague to colleague that afternoon, or turns on his heel to face the Dean, we must enter more deeply into his perspective and agency than the level at which he self-narrates his actions. Indeed, to understand his own reasons, Hal himself would need to do this.

And yet. However things might be with Hal's responsiveness to reasons, is it not detrimental to his autonomous agency that his actions and their meaning can be significantly shaped by norms he does not acknowledge, and might neither understand nor accept?[22] Don't Hal's actions exhibit a 'heteronomy of norms' akin to the more familiar heteronomy of appetite? 'Unendorsed' norms might prevent an agent from being a creature of appetite—there is all the difference civilization makes between being driven by sheer appetite and being driven by norms not of one's own making. But heteronomy is heteronomy, preventing the agent from piloting herself as she sees fit. Consider, however, our last example, Ed.

Ed

Inspired by a recent conversion experience to join a strict religious community, Ed has earnestly declared allegiance to the community's rules and practices. Important among these is an honor code, according which each is to hold everyone accountable alike, and to tolerate no rule-violation by oneself or others. Since the community's mechanisms of enforcement are informal—public confession and chastisement, denunciation of rule-breakers, shunning, etc.—each is expected to take an active part. Ed is convinced, like the others, that this sense of shared responsibility, including responsibility to mete out as well as accept social punishment, is important for the community's character and health. Ed does not flinch when called upon to confess a minor infraction he has committed, and takes his dose of public chastisement accordingly. Nor does he feel any animus toward the individual who denounced him. Yet Ed finds himself feeling ashamed and apologetic the first and only time he reports someone's rule-breaking, and he lacks the spirit for full-throated participation when occasions arise for him to join in meting out social chastisement. When his path crosses that of a member who is being shunned, he struggles awkwardly to avoid eye-contact, and cannot help but give some sign of recognition of the other's existence through his hesitant, confused manner. Ed views all this as deplorable weakness on his part, and feels guilty and ashamed of that, too.

[22] Compare the discussion of being 'in the grip' of a norm in Gibbard (1990: 58–61).

What explains Ed's behavior and feelings, despite the strength of his commitment to the community and to its rules? Ed grew up in a family in which the parents had a strong principle against punishing or berating children in front of their siblings or friends. Punishment should be respectful, he was taught, not a public spectacle and never an occasion for humiliation. Certainly, punishment was not something that the other children were encouraged to join, and they were actively discouraged from telling on one another. Once a parent reprimanded a child, usually in privacy, that was to be the end of it. This norm he and his siblings thoroughly internalized, acquiring a full suite of associated beliefs, evaluative attitudes, aversions, sensibilities, and feelings.

But as a young adult, Ed ran into problems with self-discipline and alcohol. For these, he blamed his parents and his upbringing. In his eyes, their restraint, bordering on secrecy, in matters of discipline gave the children no clear sense of limits. It encouraged the children to think that upholding the rules was not their responsibility, even when unsupervised. It sufficed, it seemed, not to tell and not to get caught. Recently, Ed had an entirely unexpected and powerful religious experience while attending a funeral for a classmate who had died of a drug overdose. He began to attend church, and soon became a fervent convert. After his long months of lonely struggle against alcohol, he was strongly attracted to a newly founded community of the faithful, which strictly prohibits alcohol and insists upon each member's responsibility for himself and for the whole.

Still, as a member he finds he cannot fully overcome his inhibitions against playing his proper part in social policing and discipline, nor can he escape his feelings of cruelty and shame when he does inflict punishment. What are we to say about the Ed's conduct?

(*a*) Is Ed's behavior in failing to denounce others and in refraining from full engagement in social punishment intentional conduct? Yes, albeit the upshot of conflicting motives.

(*b*) Do internalized norms against reporting on others and inflicting public punishment play a regulative role in explaining his behavior? Yes, and he realizes this.

(*c*) Does he endorse these norms? No.

(*d*) Do the internalized norms help us to grasp how he sees his situation, and what he saw in the actions he performed? No, *and* yes.

Ed judges his behavior wrong and unwarranted. A norm he no longer endorses or accepts still structures how he sees his situations, what meanings and valence various actions carry, and what he feels toward others and himself. He views this as weak and contemptible, and distances himself judgmentally from his own conduct and feelings. And yet, one might want to say, Ed's inability to overcome alienation from the tasks of policing the behavior of others or inflicting humiliating punishment are (very likely) more *Ed* than his current religious zeal and intense involvement with the community and its rules. The resistance he feels might, that is, come from nearer to Ed's 'center of mass' as an agent and moral being—closer to the core of his durable and basic values, norms, and sensibilities.

'But surely this sense of Ed's center does not capture anything like autonomous agency. This is a straightforward case of weakness of will, is it not?' Ed certainly hates the fact that he feels insufficiently in charge of himself. But it is a genuine question how deeply today's judgmental Ed—a recently converted religious enthusiast under the influence of a guilt-displacing narrative about hard-to-control personal failings—embodies the whole person that Ed has been, is now, and will in the long run be.

Character, especially moral character, is a controversial notion.[23] But many have found attractive an idea of moral character quite different from conscientiousness and steely self-control. Instead, they see moral character in terms of long-developing training and habituation in ways of thought, sensibility, and action—deeply internalizing certain norms of conduct, not as rigid rules of conduct, but as guides for what to be sensitive to in situations and how to think about choices. On this view, an agent's ability to be appropriately responsive to moral considerations—to act 'for the right reasons, in the right way, at the right time' as Aristotle would put it—does not reside in governance of his conduct by the operation of judgment and will alone. Such a view might locate Ed's moral character further down in his psyche than today's hyper-judgmental self.

One of the functions of autonomy, as well as moral character in this broad sense, is to enable an agent to respond directly to morally significant reasons for action that contravene current desires or enthusiasms. It thus helps equip us to resist the seduction of insistent appetites and passions. *But what helps us to resist the enthusiasms and seductions of insistent judgment?* How to prevent our moral selves from being hijacked by the peculiar allure of high-minded principles and causes, which can win over our 'better judgment', but which often are radically out of touch with the actual nature of the lives affected and the values at stake—including our own? Resistance to arbitrary judgment and willful rigidity are as important in moral life as resistance to arbitrary desire and whim. 'Weakness of will', which can frustrate an individual's attempt to impose a principle upon own his conduct, can thus be an important part of our moral endowment, opening the space for

[23] See Doris (2003) and Vranas (2005). For alternative views, see Kamtekar (2004) and Sabini and Silver (2005).

sentiments, sensibilities, and deeply internalized norms and values to exert a shaping force on our actions in their own right without our permission.

To be sure, there is no guarantee that one will have been well brought up by family or society—that one's more deeply internalized norms and most durable sensibilities will ballast the moral self in a commendable way. Prejudice characteristically is instilled and nurtured early, often running too deeply in the personality to be fully overcome by later-acquired norms of equal respect and regard. And internalized norms governing social comportment, and the associated feelings of shame and embarrassment, too often overpower what an agent recognizes to be much weightier moral concerns—as when we fail to insist to a friend that he is too drunk to drive, or brush past a situation of urgent need in order to avoid being a few conspicuous minutes late at a gathering.[24] The point for our purposes is not to decide when, or how frequently, these sources of resistance to judgment's sway work to the good. Rather, the point is that judgment is but one component of our capacity to be aptly responsive to reasons—a component that, like the others, can be insistent yet unreliable. Judgment by its nature is no less vulnerable than sentiment to uninformed, narrow-minded, or overenthusiastic tendencies. *Mature* autonomy—a *fully developed* moral personality—requires diverse counterbalances, so that no one channel of receptiveness to reasons, and no one locus for responding to these reasons, enjoys hegemony.[25] Norms recalcitrant in the face of judgment, along with their associated attitudes and feelings, can add substance to our personality, lowering our center of mass as agents and enhancing our stability when judgment has become benighted or enthralled. It is not impossible to imagine Ed, some years after he has dried out, left the religious community, and successfully regained his footing in the larger world, looking back and thinking himself fortunate to have been saved by his inhibitions from committing cruelties in the name of righteousness that he now would regret—saved by his own norms, yet in spite of his 'better judgment'.

What emerges at the end of our brief search for belief-like attitudes underlying normative guidance? A sense of the diversity of ways in which norms can become integral to agency. No privileged attitude—of endorsement, acceptance, or identification—accounts for the role of norms in shaping our lived world and contributing to the reasons for which we act. Humble

[24] Cf. Darley and Batson's well-known experiment (1973), in which the likelihood that seminary students would stop to assist an obviously needful individual was dramatically affected by whether the student believed himself to be slightly behind schedule in giving a lecture—even on the topic of the Good Samaritan.

[25] For a psychodynamic perspective, see Shapiro (1981).

internalization of norms without the self's permission, approval, or identification, like humble acquisition of beliefs without the benefit of judgment or reflection, provides much of our substance as agents. And the critical assessment and revision of norms that saves us from mere conformity and inertia, like the critical assessment and revision of what we believe, proceeds more often by trial-and-error feedback and unselfconscious readjustment over the course of experience than by spontaneous higher-order acts of endorsement or self-definition. Both, however, play a crucial role in making us candidates for rational agency and moral accomplishment. Threatening as this might be to our autonomy in the narrow sense, it makes possible our autonomy in the broad sense—agents bearing distinctive histories, sensibilities, and limitations of the kind that a life-portrait drawn 'from the inside out' seeks to capture.

REFERENCES

Arpaly, Nomy, *Unprincipled Virtue: An Inquiry into Moral Agency* (Cambridge: Cambridge University Press, 2003).

Bargh, J. A., and Chartrand, T. L., 'The Unbearable Automaticity of Being', *American Psychologist*, 54 (1999), 462–79.

Bechara, Antoine, Damasio, Helen, Tranel, David, and Damasio, Antonio R., 'Deciding Advantageously Before Knowing the Advantageous Strategy', *Science*, 275 (1997), 1293–5.

Bratman, Michael, 'Reflection, Planning, and Temporally Extended Agency', *Philosophical Review*, 109 (2000), 35–61.

Darley, J. M., and Batson, C. D, 'From Jerusalem to Jericho: A Study of Situational and Dispositional Variables in Helping Behavior', *Journal of Personality and Social Psychology*, 27 (1973), 100–19.

Davis, M. (ed.), *Interaction Rhythms: Periodicity in Communicative Behavior* (New York: Human Sciences Press, 1982).

Doris, John, *Lack of Character: Personality and Moral Behavior* (Cambridge: Cambridge University Press, 2003).

Engel, Pascal (ed.), *Believing and Accepting* (Dordrecht: Kluwer, 2000).

Frankfurt, Harry, *The Importance of What We Care About* (Cambridge: Cambridge University Press, 1988).

Gibbard, Allan, *Wise Choices, Apt Feelings* (Cambridge, Mass.: Harvard University Press, 1990).

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

Gordon, Robert M., 'Sympathy, Simulation, and the Impartial Spectator', *Ethics*, 105 (1995), 727–42.

Hatzimoysis, A. (ed.), *Philosophy and the Emotions*, Royal Institute of Philosophy Supplement, 54 (Cambridge: Cambridge University Press, 2003).

Humberstone, I. L., 'Direction of Fit', *Mind*, 101 (1992), 59–83.

Kamtekar, Rachana, 'Situationism and Virtue Ethics on the Content of our Character', *Ethics*, 114 (2004), 458–71.

Kant, Immanuel, *Metaphysics of Morals*, tr. Mary J. Gregor (Cambridge: Cambridge University Press, 1996).

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

Railton, Peter, 'Truth, Reason, and the Regulation of Belief', *Philosophical Issues*, 5 (1994), 71–93.

____ 'Mark Twain's "Sound Heart and Deformed Conscience": Rational Beings Being Rational', John Dewey Lecture, University of Vermont (unpublished, 1997).

____ 'How to Engage Reason: The Problem of Regress', in R. J. Wallace, P. Pettit, S. Scheffler, and M. Smith (eds.), *Reason and Value* (New York: Oxford University Press, 2004).

Ruby, Perrine, and Decety, Jean, 'Effect of Perspective-Taking during Simulation of Action: A PET Investigation of Agency', *Nature Neuroscience*, 4 (2001), 546–50.

Sabini, John, and Silver, Maury, 'Lack of Character? Situationalism Critiqued', *Ethics*, 115 (2005), 535–62.

Shapiro, David, *Autonomy and Rigid Character* (New York: Basic Books, 1981).

Solomon, R. C. (ed.), *Thinking about Feeling: Contemporary Philosophers on Emotions* (New York: Oxford, 2004).

Timmons, Mark, and Horgan, Terry, 'Morality without Moral Facts', in J. Dreier (ed.), *Contemporary Debates in Moral Theory* (Oxford: Blackwell, forthcoming).

Van Fraassen, B. C., *The Scientific Image* (Oxford: Clarendon Press, 1980).

Velleman, J. David, 'On the Aim of Belief', in *The Possibility of Practical Reason* (Oxford: Clarendon, 2000).

Vranas, Peter, 'The Indeterminacy Paradox: Character Evaluations and Human Psychology', *Noûs*, 39 (2005), 1–42.

This page intentionally left blank

2

Saying what we Mean: An Argument against Expressivism

Terence Cuneo

Some years ago I heard a well-known professor tell an audience that the best response he had witnessed to expressivism was that of a colleague who pointed to an expressivist paraphrase of a moral sentence on a blackboard and exclaimed, while rapping his knuckles on the board, that this paraphrase did not capture what he meant to say when engaging in moral discourse. On that occasion, I recall feeling that this was a simple-minded response to a very subtle position. I think now, however, that this response is in its fundamentals correct, and I shall endeavor in this essay to explain why. The argument I will develop is not a variant of what is now the standard objection to expressivism, namely, that it cannot explain the phenomenon of so-called embedded contexts.[1] Rather, it is one that draws upon contemporary speech act theory and maintains that expressivism is false on account of its being unable to accommodate properly the illocutionary act intentions of agents who engage in ordinary moral discourse. More precisely, the objection hinges on the question of whether, when agents engage in ordinary moral discourse, they intend to assert moral propositions. I argue that both an affirmative and a negative answer to this question yield unacceptable results for the expressivist. Given several plausible assumptions about the nature of illocutionary acts, an affirmative answer implies that expressivism is

I have two groups of scholars to thank for their help with this essay. Among the philosophers, I thank Chris Eberle, Phil Goggans, John Greco, Steve Layman, Pat McDonald, Mark Murphy, Ryan Nichols, Sarah Stroud, Russ Shafer-Landau, and C. J. Van Dyke for their comments. Jimmy Lenman and Christian Miller, in particular, saved me from committing some egregious errors. Among the sociologists, I thank Michael Emerson, Jennifer McKinney, Dave Sikkink, Christian Smith, and participants in the workshop 'Morality, Culture, and the Power of Religion in Social Life' held at Calvin College in February 2003. Finally, thanks to Luke Reinsma for editorial suggestions.

[1] See Geach (1960, 1965) for the classic formulation of the objection.

false, while a negative one (among other things) falls afoul of our best empirical evidence for what at least many agents intend to say when engaging in ordinary moral discourse. In what follows, I shall present what I will call the 'Core Argument' for this conclusion in schematic form, commenting on each premise, and then consider four replies to the argument.

I. Preliminaries

By a 'moral sentence' I shall mean an atomic sentence that has the surface form of predicating a moral property of an entity. 'Smith's assassination of Jones is wrong' and 'Sam is compassionate', according to the present view, are examples of moral sentences.[2] By 'moral discourse' I shall mean discourse that consists only in the sincere utterance of moral sentences. By a 'moral proposition' I shall mean the content (or, if you like, the object) of a moral sentence that (in at least some cases) purports (robustly) to represent a moral fact.[3] And by a 'moral fact' I shall mean a feature of the world that makes the content of moral sentences true (or perhaps makes the sentences themselves true), and that can be represented by the content of such sentences (or perhaps the sentences themselves).[4] *That Smith's assassination of Jones is wrong* and *that Sam is compassionate* are examples of moral facts as I am thinking of them.

[2] I will not take a view about whether sentences such as 'murder is wrong', which are naturally thought of as being universal generalizations, are moral sentences. By stipulating that the sentences in question are atomic, I mean to exclude from the class of moral sentences so-called molecular sentences that express disjunctive moral propositions, negative moral propositions, and so forth.

[3] Three clarifications are in order. First, in what follows, I adopt the simplifying assumption that the logical form of moral propositions corresponds to the semantic components of the moral sentences that express them. Thus, if a sentence such as 'Sam is compassionate' expresses a moral proposition, it expresses the proposition *that Sam is compassionate*. Second, as I suggest above, I will use the term 'represent' in a robust, non-deflationary sense (although I make no suggestion about what this relation consists in). Third, I say that moral propositions represent moral facts 'in at least some cases' only because I wish to leave open the possibility that in some cases true moral propositions *are* moral facts and, thus, are what is represented in moral discourse and not what represents such facts in moral discourse.

[4] What I say about moral facts is intended to be largely neutral regarding their nature. It is e.g. compatible with robustly realist and constructivist accounts of such facts. However, I assume that the present account of moral facts is not compatible with so-called deflationary accounts of moral facts according to which such facts cannot be robustly represented by (the content of) moral discourse. I explore this issue at more length in Cuneo (n.d.: ch. 6).

By 'expressivism' I mean any view that embraces the following two theses:

> **Moral Nihilism:** There are no moral facts.
>
> **The Expressivist's Speech Act Thesis:** When an agent sincerely utters a moral sentence, that agent does not thereby assert a moral proposition, but rather (at least) expresses an attitude of endorsement, approval, condemnation, disapproval, or the like toward a non-moral state of affairs or object.[5]

Common to both expressivism thus understood and its cognitivist rivals is an affirmation. Both positions maintain that ordinary moral discourse *looks* as if it were discourse wherein agents assert moral propositions. For example, ordinary moral discourse appears truth-apt (e.g. 'It is true that you ought to x') and embeds in conditionals (e.g. 'If you ought to x, then you ought to y') and propositional attitude ascriptions (e.g. 'I believe that I ought to x'). Unique to expressivism, however, is a denial. The expressivist denies that the surface form of moral discourse gives us very good reason to believe that it is genuinely assertoric moral discourse. We cannot, says the expressivist, read off the linguistic function of some area of discourse simply by gazing at its surface syntax.[6].

Thus described, expressivism is a very general position that includes among its members such diverse views as emotivism, prescriptivism, norm-expressivism, quasi-realism, and assertoric non-descriptivism.[7] While the differences between these views are not unimportant, I am going to focus primarily not on what divides, but on what unites these positions. And what unites these positions, I suggest, is not simply the two theses that I have just identified, but a common rationale for accepting them. Simon Blackburn states this rationale when he writes that the very 'essence' of

[5] I understand by a 'sincere' utterance of a moral sentence an utterance that is not intended for the purpose of dissembling. I use the locution 'expressing an attitude' to stand for those illocutionary acts that Alston (2000: ch. 4) calls 'expressives'. An act of expressing an attitude is an act of expressing a non-assertoric attitude devoid of moral propositional content. Or to use the terminology employed by Horgan and Timmons (2000), these attitudes are devoid of moral 'descriptive' content. Finally, the qualifications I introduce in the previous footnotes regarding moral facts are intended to apply to moral states of affairs and objects.

[6] See Blackburn (1993: 57; 1998: 50)

[7] See Ayer (1936), Stevenson (1963), Hare (1981), Gibbard (1990), Blackburn (1984, 1993, 1998), and Timmons (1999). Gibbard (2003) calls the view he develops an expressivist position. However, since Gibbard intends to be noncommittal on the issue of whether there are moral facts (see p. x) and takes his view merely to depict a way in which moral discourse might work (see pp. 6, 8), the view he develops in this book isn't clearly an expressivist view in the sense I am using the term. Accordingly, I will not assume that the argument I develop against expressivism applies to it.

expressivism is 'to protect ... against the descent into error theory'. [8] Central to expressivism, then, is the conviction that any view that implies that ordinary folk are massively in error about morality is unacceptable. [9] Accordingly, the expressivist urges that any acceptable moral theory should satisfy the following injunction:

> **The Expressivist's Guiding Rationale:** Avoid an error-theoretic account of ordinary moral thought and discourse.

The Expressivist's Guiding Rationale is crucial in two respects for understanding expressivism. In the first place, it explains why expressivists reject cognitivist views of moral discourse in favor of The Expressivist's Speech Act Thesis. For suppose we were to accept Moral Nihilism or the claim that there are no moral facts. And suppose also we were to accept moral cognitivism or the view that moral discourse expresses moral propositions. Having accepted these two views, we would find it impossible to satisfy the guiding rationale, for then we would be committed to a view according to which moral discourse purports to represent moral reality, but fails to do so because there is no such reality to represent. If expressivists are right, the only plausible way by which we can at once accept Moral Nihilism and satisfy the guiding rationale is by embracing the claim that moral discourse does not even purport to represent moral reality. Then (and perhaps only then) there is—to use Blackburn's words—'no real *mismatch* between the truth about the nature of [moral] ... claims, and their content'.[10]

The second respect in which the guiding rationale is crucial for understanding expressivism is that it helps us to see that we have good reason to understand The Expressivist's Speech Act Thesis not as a proposal for how we ought to use moral discourse, but as a descriptive claim about how we actually use ordinary moral discourse. For notice, if the speech act thesis were simply a proposal for how we ought to engage in moral discourse,

[8] Blackburn (2002: 167).

[9] At least this is true of expressivism in its most powerful and sophisticated guises. See e.g. Blackburn (1993: ch. 8; 1984: 171). Like Blackburn, Gibbard (1990: 8) claims that his account does not leave 'normative language defective or second-rate'. Timmons (1999: 175) says something similar when he writes that, in his view, assertoric nondescriptivism is not at odds with our ordinary moral practice because 'I just do not think that ordinary moral discourse presupposes that ethics is objective *in the sense that realism and some versions of constructivism attempt to capture*'. See also Hare (1981: 81–3).

[10] Blackburn (1993: 56). In a somewhat different context, Blackburn (1999: 214) writes that for the expressivist there 'is no problem of relativism because there is no problem of moral truth. Since moral opinion is not in the business of *representing* the world, but of assessing choices and actions and attitudes in the world ...'

it would be compatible with the view that ordinary moral discourse is massively in error and, thus, incompatible with satisfying The Expressivist's Guiding Rationale. In saying this, I don't mean to claim that expressivists have been entirely clear about the way in which the speech act thesis should be understood. They haven't.[11] I suggest, however, that interpreting the speech act thesis as a descriptive claim is the best way to make sense of the reasons expressivists offer in favor of their view.

I now want to identify a tension between The Expressivist's Speech Act Thesis and The Expressivist's Guiding Rationale with which I will be concerned in the remainder of this essay. What opens the conceptual space for the speech act thesis is the broadly Wittgensteinian insight that the surface form of an area of discourse can mask the genuine linguistic function of that area of discourse. However, it is widely accepted that Wittgenstein taught us another lesson, namely, if we want to find out what the linguistic function of some area of discourse is, we should pay close attention to the ways in which we *use* the sentences that comprise that area of discourse. The thesis I wish to defend is that, while expressivists have taken the first Wittgensteinian lesson to heart, they have ordinarily not done so with the second. To put the matter in the jargon of contemporary speech act theory: while expressivists have paid a great deal of attention to the manners in which the sentences used to express attitudes can mimic the syntactic properties of sentences that express moral propositions, they have paid comparatively little attention to what it is for the sentential act of uttering a moral sentence to count as the illocutionary act of 'expressing an attitude' or asserting a moral proposition. In particular, they have paid comparatively little attention to the role that illocutionary act intentions play in the performance of speech acts such as expressing an attitude and asserting. Getting clear on this issue, I suggest, will help us see why we should reject The Expressivist's Speech Act Thesis.

[11] Timmons (1999: 154) apparently takes the speech act thesis to be a descriptive claim: 'Moral statements, in their primary use, do not purport to make such ontological claims; rather, their primary function is to evaluate, not to describe'. Joyce (2001: 201 n. 38) and Dreier (1999) interpret Blackburn's quasi-realism as a project that seeks to protect ordinary moral discourse. Hare (1981: 86), moreover, writes that 'ordinary people when they use these [moral] words are not intending to ascribe objective prescriptive properties to actions'. Allan Gibbard (1990: 154) is more ambivalent, claiming, on the one hand, that 'norm-expressivism is meant to capture whatever there is to ordinary notions of rationality if Platonism is excluded'. On the other hand, Gibbard makes it clear that his account of rationality is not so much intended to capture what people ordinarily mean by the term 'rational', but is a proposal about how we can plausibly reconstruct normative language (ibid. 30–4). In light of The Expressivist's Guiding Rationale, I shall take the former strain of Gibbard's thought as more nearly approximating his considered view in Gibbard (1990). For a different interpretation of Gibbard, see Sturgeon (1995).

II. The Core Argument

The Core Argument I wish to develop is predicated on the assumption that we perform illocutionary acts such as asserting, insisting, promising, commanding, expressing contempt, and the like. Let me now say something about how I shall think about such acts.

I assume, first of all, that we standardly perform illocutionary acts by way of uttering sentences of certain types. So, for example, a standard way of asserting *that the car won't start* is by uttering the sentence 'The car won't start.' I do not assume that performing so-called sentential acts is the only way by which we can perform illocutionary acts.[12] I can also perform the act of asserting *that the car won't start* by, say, signing or sending a smoke signal. However, on this occasion, I shall be exclusively concerned with those illocutionary acts that we perform by way of uttering sentences.

Furthermore, I will assume that illocutionary acts have *content*. By the content of an illocutionary act I mean (roughly) what a person who performs that act seeks to communicate by the performance of that act—what the hearer must grasp to understand what the speaker is saying by the performance of that act.[13] Thus understood, the content of an illocutionary act should not be identified with *propositional* content since what an agent can express by the performance of an illocutionary act can include non-propositional elements (such as feelings of disapproval).

Finally, and most importantly, I will assume that an agent's performing an illocutionary act is something that she does deliberately or intentionally. Performing an illocutionary act isn't something that merely *happens* to an agent; it is something an agent does. In saying this, I don't mean to suggest that when a speaker performs an illocutionary act of a certain type, her intention to perform that act is always explicit to her. Rather, I will assume that a speaker's intentions can come in varying degrees of explicitness and precision. In certain cases, it may be perfectly evident to a speaker what illocutionary act she intends to perform by way of uttering a given sentence. In other cases, the intention in question may be elicited only by skillful questioning. After I have uttered the sentence, 'Eat the leftovers!' you may ask me: 'Are you commanding or merely exhorting me

[12] Thus understood, 'sentential acts' are a subset of what J. L. Austin (1962) called 'locutionary acts'.

[13] See Alston (2000: 15). Of course an agent can seek to communicate (say) displeasure with someone by using a certain kind of tone of voice or facial expression when performing an illocutionary act. But this is not what I have in mind by the content of an illocutionary act. The content of an agent's illocutionary act concerns *what* he says and not *how* he says it (ibid. 108).

to eat the leftovers?' If I truthfully tell you that I was merely exhorting you, this settles the issue of the illocutionary act type I performed.[14] Modulo certain exceptional cases, my intention to—or better, my endeavoring to—perform an illocutionary act of a given type by uttering a given sentence determines whether I in fact performed an illocutionary act of that type by uttering that sentence.[15]

With these distinctions in hand, I can now introduce the concept of 'ordinary optimal linguistic conditions'—or 'ordinary optimal conditions', for short.[16] Ordinary optimal conditions are ones in which a speaker and his audience have competence with a given language L, and the speaker performs an illocutionary act of a certain kind by way of performing some appropriate sentential act that conforms to the norms of L. Ordinary optimal conditions, then, are ones in which the speaker's audience has sufficient clues that he intends to perform a speech act of a given type by way of performing a sentential act of a certain kind (e.g. he does not use a secret code with which his audience is unfamiliar), the speaker doesn't misuse language or engage in a slip of the tongue in performing the sentential act in question, or the like. Keeping in mind these qualifications, here is the first premise of the Core Argument I want to develop:

(1) In ordinary optimal conditions, an agent performs an illocutionary act of Φing by way of performing a sentential act if and only if that agent intends to Φ by way of performing that sentential act.[17]

[14] To say that my illocutionary act intentions can become evident to me upon reflection or skillful questioning is not to claim, however, that they are always very precise. Upon reflection, I may be aware that when I utter the sentence 'The dinner was delicious', I intend to say, of the dinner, that it was delicious. But I may have no view about what deliciousness consists in, or whether each course of the dinner was delicious, and so forth. Vagueness of this variety, I shall suggest later, is not something that affects the main lines of the argument I am developing.

[15] Two points: first, philosophers such as Bratman (1987: 133) distinguish between (standing) 'intentions', on the one hand, and 'tryings' or 'endeavorings', on the other—the latter being (roughly) a certain way of expressing intentions. In deference to common usage, I have spoken of (and will continue to speak of) 'illocutionary act intentions' when referring to those intentions that determine the character of an illocutionary act. But, unless the context reveals otherwise, these mental states are best thought of as being endeavorings. Second, Bratman (ch. 8) distinguishes between intentions and intentional actions, denying that intending to act in a certain way is the expression of an intention to act in that way. Bratman may be right about this. In the interest of simplicity, however, I have spoken as if intending to act in a certain way expresses the intention to act in that way.

[16] I borrow the term from Rosati (1996), but not the details of her way of understanding it.

[17] The scope of the relevant intention concerns only the illocutionary act in question. Thus it should read as follows. In ordinary optimal conditions, an agent performs an

(1), I submit, states a truism about illocutionary acts: given that the proper conditions hold, necessary and sufficient for performing a speech act of a certain kind by way of performing a sentential act of a certain type is that agent's intending to perform a speech act of that kind by way of performing that sentential act. But it also raises a question. What exactly is involved in a speaker's intending to perform an illocutionary act of a given type?

I propose to remain neutral on this question. One might hold, along with 'perlocutionary intention' theorists such as H. P. Grice and Stephen Schiffer, that the relevant intention in question consists in a speaker's getting his audience in a certain state of mind on account of his uttering a sentence of a certain type.[18] One might believe, for example, that what makes it the case that Sam performs the illocutionary act of asserting *that the car won't start* by sincerely uttering the sentence 'The car won't start' is that Sam intends his audience to believe that the car won't start and intends them to believe this on the basis of his uttering this sentence. Alternatively, one might believe, along with 'illocutionary intention' theorists such as John Searle, William Alston, and Nicholas Wolterstorff, that the intention in question is one in which a speaker intends to take responsibility for a certain state of affairs.[19] So, for example, according to the illocutionary intention account, for me to assert *that the car won't start* by way of uttering the sentence 'The car won't start' is for me to leave myself open to appropriate correction, blame, reproach, or the like in case it is false that the car won't start. It is my deliberately taking responsibility in this fashion for the fact *that the car won't start* that brings it about that my uttering this sentence counts as an assertion. Although I myself find the latter view considerably more plausible than the former, the argument I shall develop can be understood in terms of either position.[20]

The second premise of the Core Argument I am propounding is entailed by The Expressivist's Speech Act thesis and the claim that this thesis concerns ordinary optimal conditions. It says:

> (2) If expressivism is true, then, in ordinary optimal conditions, when an agent sincerely utters a moral sentence, that agent does

illocutionary act of Φing by way of performing a sentential act if and only if that agent intends [to Φ] by way of performing that sentential act. Moreover, I understand the qualifier 'ordinary optimal conditions' in such a way that it qualifies both the performance of and the intention to perform the illocutionary act in question.

[18] See Grice (1957) and Schiffer (1972) for a defense of the perlocutionary view. Schiffer has subsequently abandoned this view.

[19] See Alston (2000), Searle (1969), and Wolterstorff (1995: ch. 5). Williamson (2000: ch. 11) briefly defends something like the illocutionary intention view, while Brandom (1994) has some affinities with it.

[20] See Alston (2000: ch. 2) for an argument against the perlocutionary view.

not thereby assert a moral proposition, but expresses an attitude toward a non-moral state of affairs or object.[21]

As I've already indicated, I am using the locutions 'expresses an attitude' and 'assert' to denote different illocutionary act types.[22] So, I assume that

(3) Expressing an attitude and asserting are distinct illocutionary act types.

Let's now combine the first three premises of the argument. When we combine (1), (2), and (3) we get this:

(4) So, if expressivism is true, then, in ordinary optimal conditions, when an agent performs the sentential act of sincerely uttering a moral sentence, that agent does not thereby intend to assert a moral proposition, but intends to express an attitude toward a non-moral state of affairs or object.

I want now to draw attention to what expressivists say about what we purport to do when we utter moral sentences. At the beginning of his book *Wise Choices, Apt Feelings*, Allan Gibbard writes that

> normative talk is part of nature, but it does not describe nature. In particular, a person who calls something rational or irrational is not describing his own state of mind; he is expressing it. To call something rational is not to attribute some particular property to that thing—not even the property of being permitted by accepted norms.[23]

Strictly speaking, in this passage Gibbard is elaborating upon an expressivist account of rationality, and not morality. However, that shouldn't matter for our purposes since Gibbard's expressivist account of moral discourse is simply an extension of his expressivist view of rationality. In Gibbard's view, to say that a person's act is, say, wrong is (roughly) to say that it is rational to blame her for performing that act.[24] In any case, the important point to notice is that Gibbard doesn't say that in uttering the sentence 'Sam is rational' a person doesn't *purport* to call Sam rational. He merely says that in uttering this sentence a person does not thereby assert that Sam is rational. But the first premise of our argument tells us that, in ordinary optimal conditions, what we intend to do by uttering a sentence determines

[21] Strictly speaking, this premise should say that, if expressivism were true, then when uttering a moral sentence an agent *at least* expresses an attitude toward a non-moral state of affairs or object. I'll leave this qualification implicit.

[22] In so doing, I follow both Searle (1969) and Alston (2000).

[23] Gibbard (1990: 7–8).

[24] At least this is true of his view in Gibbard (1990: see esp. ch. 3).

what illocutionary act(s) we perform by uttering that sentence. So, our question is this: according to the expressivist, when an agent sincerely utters a sentence such as 'Sam is compassionate', does that person thereby intend to assert the proposition *that Sam is compassionate*?

The response to which the expressivist appears committed is, No. In sincerely uttering a sentence such as 'Sam is compassionate,' a speaker does not thereby intend to assert a moral proposition—and for two reasons. To begin with, this denial is entailed by the three premises we've considered thus far—premises that expressivists themselves would appear to accept.[25] Second, on the assumption that (1) is true (and ordinary optimal conditions hold), if the expressivist were to say that, when we engage in moral discourse, we thereby intend to assert moral propositions, she would commit herself to an error theory of morality similar to the kind J. L. Mackie defended.[26] But as I've already emphasized, expressivists themselves indicate that this is an unacceptable consequence, for it is precisely this kind of view that expressivist positions are designed to avoid.

So, the view to which the expressivist appears committed is what we should expect: by sincerely uttering a moral sentence, an agent does not thereby intend to assert a moral proposition, but rather (at least) intends to express an attitude. But now consider what expressivists say about what we do when we engage in moral discourse.

In *Spreading the Word*, Simon Blackburn suggests that engaging in moral discourse is a matter of 'projecting' attitudes: 'we *project* an attitude or habit or other commitment which is not descriptive onto the world, when we speak and think as though there were a property of things which our sayings describe which we can reason about, know about, be wrong about, and so on'.[27] And in *Essays in Quasi-Realism*, Blackburn hypothesizes that projections thus understood are the upshot of a 'mechanism whereby what starts life as a non-descriptive psychological state ends up expressed, thought about, and considered in propositional form'.[28]

In *Wise Choices, Apt Feelings*, Gibbard says something striking concerning discourse about what is rational:

> When a person calls something rational, he seems to be doing more than simply expressing his own acceptance of a system of norms . . . he claims to *recognize and report* something that is true independently of what he himself happens to accept or reject. . . . Any account of his language that ignores this claim must be defective. It

[25] That Gibbard e.g. accepts the centrality of intentions to the performance of speech acts is evident in Gibbard (1990: 84–6; 2003: 76).

[26] Mackie (1977: ch. 1). For a defense of a view similar to Mackie's, see Joyce (2001).

[27] Blackburn (1984: 170–1).

[28] Blackburn (1993: 5).

> may capture all that the speaker could claim without illusion, but it will not capture all that he is *in fact claiming*.[29]

Finally, in a more recent article on expressivism, Justin D'Arms and Daniel Jacobson write that

> philosophers such as Blackburn see little reason for noncognitivists to forgo property talk and are even willing to speak of the truth of evaluative claims. All [i.e. expressivists such as Gibbard and Blackburn and realists such as John McDowell] grant, too, that the phenomenology of valuing is such that sentiments purport to be sensitivities to features of the world—that is, to evaluative properties. Hence, we agree with McDowell that the only way to understand our responses as they do and must seem to us, whatever our metaethics, is to be prepared 'to attribute, to at least some possible objects of the responses, properties that would validate these responses'.[30]

If D'Arms and Jacobson are right, expressivists don't believe that 'valuing' simply consists in registering a vague 'evaluative response' to a state of affairs. Rather, valuing purports to be a 'sensitivity' to evaluative features of the world. As such, the states of mind expressed in valuing—what D'Arms and Jacobson call 'sentiments'—purport to be *about* evaluative features of the world. If we add the plausible assumption that, by uttering a moral sentence an agent thereby intends to express a 'sentiment', then what D'Arms and Jacobson tell us is that expressivists themselves hold that, when we engage in moral discourse, we purport to say something about moral reality.

A very natural way to read these quotations is in light of what Gibbard says in the second passage I have quoted from him. In this passage, Gibbard says that when an agent calls something 'rational', that agent 'claims to recognize and report something that is true independently of what he himself happens to accept or reject'.[31] We could, suggests Gibbard, offer a different gloss on what this agent is claiming, but it would not capture 'all that he is in fact claiming'.[32] And what Gibbard appears to mean by this is that an alternate gloss on what this agent is saying wouldn't capture everything that this agent *purports* to claim.

I believe that what Gibbard says here is true. In the ordinary case, when an ordinary agent calls a particular action 'rational', 'wrong', or 'compassionate', he means to predicate of that action the property of being rational, wrong, or compassionate respectively. Nonetheless, what Gibbard

[29] Gibbard (1990: 153; italics mine).

[30] D'Arms and Jacobson (2000*b*: 730). The quote is from McDowell (1985: 207). Compare, also, what D'Arms and Jacobson (2000*a*) say at p. 69 n. 9.

[31] Gibbard (1990: 153).

[32] Ibid.

says here is not something an expressivist should say. For one thing, if premise (1) of the Core Argument is true, what Gibbard says here is in direct contradiction with what he says in the first passage I have quoted from him ('To call something rational is not to attribute some particular property to that thing . . .'). More importantly, if we combine what Gibbard says in this passage with (1) and (3), it implies what I shall call the 'strong result':

> (5) It is false that, in ordinary optimal conditions, when an agent performs the sentential act of sincerely uttering a moral sentence, that agent does not thereby intend to assert a moral proposition, but intends to express an attitude toward a non-moral state of affairs or object.

And this entails that

> (6) Expressivism is false.

III. Four Expressivist Responses

What I have called a 'natural' reading of the passages quoted from Blackburn, Gibbard, and company is not, of course, the only way to read them. In what follows, I want to explore four expressivist responses to the Core Argument, all of which attempt to avoid the conclusion that The Expressivist's Speech Act Thesis is false. The first position, which I shall call 'perspectivalist' expressivism, responds to the Core Argument by employing a distinction, popular of late among expressivists, between different perspectives regarding ordinary moral discourse. The second position, which I call 'illusionist' expressivism, maintains that, although we may *think* we express moral propositions when engaging in ordinary moral discourse, we do not actually do so, but rather express attitudes of various sorts. The third position, 'agnostic' expressivism as I call it, flatly denies that we can gain a cognitive grip on our illocutionary act intentions. And the fourth position, what we can call 'sophisticated' expressivism, contends that although we can ascertain our illocutionary act intentions, our moral discourse consists primarily in our intending to perform assertion-like acts that mimic genuine acts of assertion. Of these four responses, it is the last that will occupy most of my attention, for two reasons. In the first place, this position is arguably the most promising of the expressivist views we'll consider and, as such, deserves the most attention. Second, to this point I've said rather little about why we should accept premise (5) of the Core Argument, merely pointing out that expressivists themselves at times appear to accept it. But I think there is more to be said in favor of premise (5). In particular, in response to sophisticated expressivism, I wish to suggest that discerning the character of our illocutionary act intentions is largely an empirical matter and that there is good empirical evidence to marshal in favor of the claim that, when engaging in ordinary moral discourse, many agents intend to express moral propositions and not merely express attitudes.

The first response: perspectivalist expressivism

The first expressivist response I wish to consider—what I call 'perspectivalist' expressivism—takes its inspiration from an interpretation of the passages quoted earlier from Simon Blackburn. Perhaps the best way to understand what Blackburn says in these passages is first to introduce a concept central to Blackburn's 'quasi-realist' expressivism and then clarify another.

The concept I'd like to introduce is that of a 'perspective' on what we are doing when we engage in moral discourse. In several places, Blackburn suggests that we can distinguish between an 'internal' and 'external' perspective of what we are doing when we 'moralize'.[33] The internal moral perspective, in Blackburn's view, is one of an engaged participant in ordinary moral practice. From the internal perspective, we say that we respond to moral features, claim that moral judgments are true by virtue of representing those features, and say that moral truths obtain independently of our feelings. Indeed, in some places Blackburn is willing to say that, from the internal perspective, some of our moral judgments *are* true—though Blackburn denies that we should see this as having any metaphysical import.[34] The 'external' perspective, by contrast, is that of the observer not engaged in moralizing—in Blackburn's case, the perspective of the philosopher committed to a robust version of naturalism who is diagnosing what agents do when they moralize. From the external perspective, the philosopher denies that there are moral facts, and interprets what happens from the internal perspective as simply the 'adjusting, improving, weighing, and rejecting [of] different sentiments or attitudes'.[35] From the external perspective, there are no 'moral properties ... made for or by sensibilities', and 'the only things in this world are the attitudes of people'.[36] A perspectivalist expressivist view, then, 'deserves to be called anti-realist because it avoids the view that when we

[33] See esp Blackburn (1993: ch. 9, ch. 8, p. 157; 1984: 247; 1998: 50). See also Timmons (1999: 150–2).

[34] See Blackburn (1998: appendix).

[35] Blackburn (1993: 173–4).

[36] Ibid. 174.

moralize we respond to, and describe, an independent [moral] aspect of reality'.[37]

Now let me turn to the concept I wish to clarify—that of 'projecting an attitude'.[38] What Blackburn says in the passages quoted earlier is that for an agent to project an attitude that is not descriptive onto the world is for that agent to speak and think as though there were a property that her sayings and thoughts describe. Accordingly, to project a *moral* attitude is for an agent to project an attitude that is not descriptive onto the world in such a way that she speaks and thinks as though there were a moral property of things that her moral sayings and thoughts describe. So, definitive of Blackburn's view is the thesis that we can distinguish between the surface 'propositional' form of an attitude that is expressed and the attitude itself (or what the attitude expresses). Of course in claiming this, Blackburn is not suggesting that to project an attitude is thereby to 'play act'—to act as if something is there that we know isn't. Nor does Blackburn maintain that ordinary folk engaging in ordinary moral discourse operate with the distinction between the propositional form of an attitude expressed and the attitude itself (or the content thereof). Rather, what Blackburn suggests is that, as far as ordinary moral discourse goes, to express an attitude in propositional form is simply to think of that attitude (or better: what that attitude expresses) *as* a proposition or claim—something that can be true or false.[39]

The distinctions between different perspectives on what we are doing when we moralize and the notion of projecting an attitude are supposed to do a great deal of work in Blackburn's project. They are supposed to be the materials by which a quasi-realist can at once show that we make true moral claims *and* that expressivism is true. The idea is that, from the internal perspective, we make moral claims, some of which are true (in some sufficiently thin deflationary sense). However, from the external perspective, these claims are interpreted as projections or expressions of attitude that don't in any sense represent moral facts.

There is a problem with all this. For suppose we assume with Blackburn that the internal perspective of what we are doing when we engage in moral discourse is supposed to capture what it is like to engage in ordinary moral discourse and practice. As Blackburn himself emphasizes, however, from

[37] Blackburn (1993), p. 157. Blackburn goes on to suggest that a realist view of obligation is 'unintelligible or marks a mistake about explanation' (ibid. *n*. 9). See also Blackburn (1981: 164–5).

[38] Blackburn (1998: 77) expresses some reservation about using this locution. But, as Blackburn emphasizes, this is not because he rejects the notion of projecting an attitude, but because he realizes that using this term 'can make it sound as if projecting attitudes involves some kind of mistake' and this 'is emphatically' not what he intends.

[39] See Blackburn (1998: 317–19).

the internal perspective we don't think *as though* Smith's assassination of Jones is wrong; we think of it *as being* wrong. Moreover, from the internal perspective, when Sam sincerely utters the sentence 'Smith's assassination of Jones is wrong', he doesn't think of what he says *as though it is about* the wrongness of Smith's action; he thinks of it as *being about* the wrongness of what Smith did, and intends to express this thought by way of uttering the sentence in question. We can grant, then, that Blackburn is entirely correct to claim that we can offer an external reading of what Sam says—a reading according to which Sam expresses an attitude toward Smith's killing without asserting a moral proposition. And we can even grant that this maneuver can assuage certain types of worry one might have regarding the expressivist enterprise. But offering an external reading of this type is not a particularly helpful response to the Core Argument developed thus far. This is because whether we can give an external reading of what we are doing when we engage in moral discourse has no bearing upon what we actually intend to do when we engage in moral discourse. For if premise (1) of the Core Argument is true, in ordinary optimal conditions, necessary and sufficient for determining what type of speech act an agent performs by way of uttering some moral sentence is what that agent *intends* to do by way of uttering that moral sentence. It follows from this that, if we assume from the internal perspective that we intend to assert moral propositions, perspectivalist expressivism is also committed to the strong result, or the claim that

(5) It is false that, in ordinary optimal conditions, when an agent performs the sentential act of sincerely uttering a moral sentence, that agent does not thereby intend to assert a moral proposition, but intends to express an attitude toward a non-moral state of affairs or object.

And the strong result, we've seen, yields

(6) Expressivism is false.

The second response: illusionist expressivism

Earlier I considered the following passage from Allan Gibbard:

> When a person calls something rational, he seems to be doing more than simply expressing his own acceptance of a system of norms . . . he claims to *recognize and report* something that is true independently of what he himself happens to accept or reject. . . . Any account of his language that ignores this claim must be defective. It may capture all that the speaker could claim without illusion, but it will not capture all that he is *in fact claiming*.[40]

[40] Gibbard (1990: 153; italics mine).

In the last section, I claimed that a natural reading of this passage entails that expressivism is false. But there is another reading of what Gibbard says in this passage. What Gibbard says is that a position that rejects a cognitivist account of moral discourse may capture all that an agent could be doing *without illusion*. This suggests that there may be available to the expressivist a distinction between what it *seems* to an agent he is doing and what he is *actually doing* when he utters a moral sentence, which is different from the internal/external distinction that Blackburn employs. According to this alternative, the expressivist can readily admit that when an agent utters a moral sentence, it seems to him that he thereby intends to assert a moral proposition. However, this doesn't imply that by uttering a moral sentence the agent in question thereby intends to assert a moral proposition. It may be the case that we are ordinarily confused or misled about what we intend to do when we engage in moral discourse. So, it is available to the expressivist to say that what an agent really intends to do by uttering a moral sentence is not assert a moral proposition, but to express an attitude toward a non-moral state of affairs. This position is what I have called 'illusionist' expressivism.

I doubt that an expressivist should be very happy with the claim that we are massively confused or misled about what we intend to do when engaging in moral discourse. In this regard, it is helpful to distinguish between a first-order and a second-order error theory. A first-order error theory is of the sort that Blackburn and Gibbard want to avoid. It says that the content of all our moral statements is mistaken inasmuch as it purports, but fails, to express true moral propositions. (This may be because either these propositions are false or are neither true nor false, since the subject terms of the sentences that express them fail to refer.) A second-order error theory, by contrast, says that the content of what we take for granted or believe *about* what we intend to do by uttering moral sentences is mistaken. (This may be because either the contents of these attitudes are false or neither true nor false, since the subject terms of the sentences that express them fail to refer.) Now it should be admitted that the expressivist view under consideration avoids a first-order error theory of moral discourse. But the view doesn't avoid a second-order error theory. Indeed, given a plausible assumption, the position implies it. The plausible assumption is that, if it appears to an ordinary person in ordinary circumstances that she intends to Φ at t, then (in the absence of relevant defeaters) she takes it for granted or believes that she intends to Φ at t. According to the present view, then, since it appears to an ordinary agent in ordinary circumstances that she intends to assert a moral proposition at t, then (in the absence of relevant defeaters) she takes it for granted or believes that she intends to assert a moral proposition at t. But the present expressivist view also says that we are confused or misled about what we are doing when we utter moral sentences, and don't really intend to assert moral propositions by uttering moral sentences. It follows that the content of all these second-order taking for granteds or beliefs is mistaken. And, thus, it follows that an expressivist who adopts this position is committed to a second-order error theory.

Now, strictly speaking, a second-order error theory is consistent with The Expressivist's Guiding Rationale, as the contents of our second-order beliefs are not themselves moral propositions, but propositions that concern our illocutionary act intentions. Still, I think that an expressivist should be no more enthusiastic about a second-order error theory than a first-order one. We can look at the matter this way. According to a first-order error theory, the content of all our moral statements is mistaken. The content of our second-order taking for granteds and beliefs about our illocutionary act intentions, however, is true: we correctly take it for granted or believe that when we engage in moral discourse we intend to say things about moral reality. According to the expressivist view under consideration, by contrast, the content of our moral discourse is not mistaken. But the content of all our second-order taking for granteds and beliefs about what we intend to do when we engage in moral discourse is: we incorrectly take it for granted or believe that when we engage in moral discourse we intend to say things about moral reality. So, according to a first-order error theorist, we have a set of first-order statements whose content is mistaken and a set of second-order attitudes with respect to those first-order statements whose content is not. According to the expressivist view under consideration, by contrast, we have a set of first-order attitudes whose content is not mistaken and a set of second-order attitudes with respect to those first-order attitudes whose content is. Each view, then, countenances one set of attitudes whose content is mistaken and one whose content is not. Given this type of parity between the two views, however, it is difficult to see why illusionist expressivism ensures that moral discourse and belief are somehow in better shape than if a first-order error theory were true. Claiming that expressivism is in better shape because it guarantees that the content of our first-order attitudes is not mistaken, and that these attitudes are somehow more important or fundamental than our second-order ones, isn't very promising. As Harry Frankfurt and Charles Taylor have taught us, our second-order attitudes often have enormous practical and theoretical importance.[41]

The third response: agnostic expressivism

The obvious way to address the problems with both perspectivalist and illusionist expressivism is to make two moves: first, reject the claim that

[41] See Frankfurt (1971) and Taylor (1985).

from the internal perspective we intend to assert moral propositions and, second, reject the idea that we are systematically deceived about the nature of our illocutionary act intentions. Rejecting the former claim can itself take one of two forms. One could claim either that we cannot get a cognitive grip on what ordinary agents intend to do when sincerely uttering moral sentences, or that agents who participate in ordinary moral discourse intend to express entities that resemble genuine moral propositions—call them ‘moral quasi-propositions’—by way of performing speech acts that closely resemble acts of assertion. There is evidence that expressivists have endorsed both of these positions. Let’s consider them in turn.

To gain a better grip on the third type of expressivist response to the argument I have developed—what I am calling ‘agnostic’ expressivism—let me quote a lengthy passage from Simon Blackburn’s essay, ‘Errors and the Phenomenology of Value’:

> It is in principle possible that we should observe the practice of some subjects as closely as we wish, and know as much as there is to know about their ways of thinking, commending, approving, deliberating, worrying, and so on, yet be unable to tell from all that which theory they hold. The practice could be clipped on to either metaphysic. . . . To use a close analogy, there are different theories about the nature of arithmetical concepts. Hence a holist may claim that a subject will give a different total meaning to numerals depending on which theory he accepts, and this difference will apply just as much when the subject is counting as when he is doing metamathematics. All that may be true, yet it would not follow that any practice of counting embodies error. That would be so only if one could tell just by observing it which of the competing metamathematical theories the subject accepts. In the arithmetical case this would not be true. Similarly, I maintain, in the moral case one ought not to be able to tell from the way in which someone conducts the activity of moralizing whether he has committed the ’objectivist’ mistake or not; hence, any such mistake is better thought of as accidental to the practice.[42]

About this line of argument let me make several comments.

Begin with the ‘close analogy’ between mathematics and ethics that Blackburn employs. Suppose we extend the analogy a bit by assuming that our two best theories about the nature of numbers and mathematical thought and discourse are Platonism and expressivism (with regard to mathematical discourse) respectively. Suppose also that both do a fairly nice job of accounting for various features of mathematical discourse and thought, even if many philosophers have found it natural to understand ordinary mathematical discourse and thought as being implicitly committed to Platonism. Suppose, moreover, that (for reasons that most ordinary people

[42] Blackburn (1993: 151). Blackburn (1998: 51 and 121) seems to say something very different, however. I am unsure how to reconcile these different passages.

are unaware of) Platonism is false and, hence, understanding mathematical discourse in expressivist fashion saves participants in mathematical discourse from an 'objectivist' mistake. Suppose, finally, that after having paid close attention to mathematical discourse, we cannot tell which theory ordinary participants in the discourse embrace: either metaphysic can be 'clipped on' to their discourse and practice. The thrust of Blackburn's thought appears to be that, in this situation, we ought not to interpret what participants in the discourse say as being committed to Platonism. Rather, we ought to interpret what they say along expressivist lines since only the latter will give us a plausible account of various features of mathematical discourse and also save its participants from error.[43]

This line of thought seems to me correct in one respect and mistaken in another. What's correct is the claim that, if we can't tell what theory ordinary folks are committed to and, thus, cannot get at their relevant illocutionary act intentions, we ought not to interpret what they say as being committed to Platonism. What's mistaken, however, is that this gives us any reason to construe what they say in expressivist terms. If we are genuinely trying to settle the empirical issue of what the participants in the discourse are actually trying to say by way of engaging in that discourse, and what they are trying to say about the nature of numbers remains inscrutable on account of our having no idea of what theory they are committed to, then I submit that it is evident what we should conclude: we should either (i) be agnostic about whether they intend to assert propositions about numbers Platonistically understood or whether they intend to express attitudes toward non-mathematical reality or (ii) conclude that there is no fact of the matter about what they intend to do. Either way, we should not attribute to them intentions to express attitudes toward non-mathematical reality. (I do not claim, incidentally, that if we had some sort of access to the theoretical commitments of participants in ordinary mathematical discourse, and these fit poorly with Platonism, then we should remain agnostic about what they intend to say. In that case, I concur that charity suggests that no 'objectivist' mistake is being made. Rather, the suggestion is that, in the absence of such access, agnosticism or the belief that there is no fact of the matter about what agents intend to assert is the appropriate stance. I shall have more to say about this in the next section.)

Now turn to the moral case that Blackburn suggests parallels the mathematical. Suppose all the relevant parallels hold: moral realism and

[43] Once again, for reasons cited earlier, I assume that Blackburn does not wish to offer expressivism as a mere recommendation for how we might transform moral and mathematical discourse, but as an account of how to 'protect' ordinary moral and mathematical discourse.

expressivism are our best candidates for understanding moral thought and discourse, there are no moral facts (but most ordinary folk are not aware of the arguments for this), and so on. According to Blackburn, in the moral case we ought not to be able to tell from the way in which we moralize whether we are committed to the existence of moral facts or not. But if our commitments concerning whether or not there are moral facts are genuinely inscrutable and, hence, the relevant class of illocutionary act intentions are as well, then our conclusion ought not to be that The Expressivist's Speech Act Thesis is true—even if adopting it establishes that the folk are not in error. Rather, it should be either (i) agnosticism about whether participants in ordinary moral discourse intend to assert moral propositions about moral facts or express attitudes toward non-moral states of affairs or objects or (ii) admission that there is no fact of the matter about this issue.[44] We cannot settle the empirical question of what persons who engage in ordinary moral discourse intend to say by way of *superimposing* an interpretation upon what they claim. (I grant, however, that if we were to discover that the theoretical commitments of ordinary folk fit poorly with realism about morals, that would give us reason to believe that no 'objectivist' mistake is being made in moral discourse.)

So, what follows? If the line of argument I have developed is correct, it doesn't follow that agnostic expressivism is false. Rather, what follows is what I shall call the 'weak result'. The weak result says that we have strong (objective) reason not to believe The Expressivist's Speech Act Thesis.

Recall that agnostic expressivism tells us that the way to save agents who engage in ordinary moral discourse from the 'objectivist' mistake is to interpret what they say as the expression or 'projection' of attitudes. Now, if the interpretation I've offered of Blackburn's argument is correct, then either the relevant illocutionary act intentions of agents who engage in ordinary moral discourse are inscrutable or there is no fact of the matter about whether they intend to assert moral propositions or express attitudes toward non-moral states of affairs or objects. It follows—to employ the terminology of the Core Argument—that the following disjunctive claim is true: it is either false or inscrutable that, in ordinary optimal conditions, when an agent sincerely utters a moral sentence, that agent thereby intends to express an attitude, and does not intend to assert a moral proposition by way of uttering that sentence.

[44] It is worth noting that what Blackburn says here about the inscrutable nature of moral discourse sits uneasily with his repeated claims that moral discourse is non-representational. See Blackburn (2001*a*: 31; 1999: 214; 1996: 83–6); as well as Gibbard (1990: 107); and Timmons (1999: 139–47).

Consider the first half of this disjunction. If this half of the disjunction is true, then it is false that, in ordinary optimal conditions, when an agent sincerely utters a moral sentence, that agent thereby intends to express an attitude, and does not intend to assert a moral proposition by way of uttering that sentence. Since this is incompatible with The Expressivist's Speech Act Thesis (when interpreted as a claim about moral discourse in ordinary optimal conditions), it follows that expressivism is false. Now consider the second half of the aforementioned disjunction. If this half of the disjunction is true, then it is inscrutable whether in ordinary optimal conditions an agent who sincerely utters a moral sentence thereby intends to express an attitude, and does not intend to assert a moral proposition. It follows from this that (as a thesis concerning moral discourse in ordinary optimal conditions) The Expressivist's Speech Act Thesis is either false or inscrutable. Either option gives us strong (objective) reason not to believe that expressivism is true.

As we might put it, thorough-going agnosticism about our illocutionary act intentions functions as an 'undercutting' defeater for accepting expressivism.[45]

The fourth response: sophisticated expressivism

At the outset of the last section I noted that remedying the problems with both perspectivalist and illusionist forms of expressivism can take either of two forms. One might claim, on the one hand, that we cannot get a cognitive grip on what ordinary agents intend to do when sincerely uttering moral sentences or maintain, on the other, that agents who participate in ordinary moral discourse intend to express entities that resemble genuine moral propositions—call them 'moral quasi-propositions'—by way of performing speech acts that closely resemble acts of assertion. I have argued that the former route is not one that should appeal to expressivists. In the remainder of this essay, I want to consider the latter approach—the position I call 'sophisticated' expressivism—as I suspect it best captures what expressivists should say about our illocutionary act intentions.

Sophisticated expressivism hinges upon a distinction between performing acts of asserting and performing assertion-like acts. Acts of assertion, as I indicated earlier, are such that their content aims to represent reality; to assert that p is to purport to represent the fact *that p*. Thus understood, assertions explicitly express propositions, where propositions are understood to be the content of assertions—entities whose job description (at least

[45] Of course it also functions as an undercutting defeater for cognitivism!

in a wide array of cases) includes representing the world. Assertion-like acts resemble acts of assertion insofar as they manifest certain features characteristic of assertions. For example, assertion-like acts are such that their content embeds in propositional attitude ascriptions and conditionals, is truth-apt, and is irreducible to what is expressed in non-declarative sentences such as imperatives and questions.[46] However, what is 'asserted' does not purport to, nor does it, represent what the world is like. To perform the assertion-like act of uttering p, then, is not thereby to purport to represent the fact *that p*; it is rather to do something different such as 'evaluate' a state of affairs.[47] As we might put it, moral assertion-like acts express moral quasi-propositions, where moral quasi-propositions are understood to be the content of such acts—entities whose job description includes mimicking moral propositions in certain important respects, but that do not in any sense purport to represent moral reality.[48]

While the concepts of an assertion-like act and a quasi-proposition call for more elaboration, I am going to assume for present purposes that we have a sufficient understanding of them to see their importance for the expressivist project.[49] Their importance, I judge, is threefold.

First, by employing such concepts, the sophisticated expressivist can take full account of the illocutionary act intentions operative in ordinary moral discourse. More specifically, the sophisticated expressivist can say that we should view ordinary moral discourse as that in which agents intend to perform not acts of asserting that express moral propositions, but assertion-like acts that express moral quasi-propositions. If this suggestion is right, then it is not the case that sophisticated expressivism is open to the objections raised earlier against perspectivalist, illusionist, and agnostic

[46] It is this last feature of assertion-like acts, I take it, that distinguishes them from Blackburnian projections of attitude.

[47] Timmons (1999: 139) puts the matter thus regarding moral judgments: 'moral judgments are not aimed at representing or describing a world of facts. Their content is not representational but evaluative—aimed at choice and guidance of action.'

[48] The distinction between propositions and quasi-propositions parallels the distinction that Horgan and Timmons (2000) make between the cognitive and the descriptive content of a belief. The main difference between my and Horgan and Timmons's way of putting things is that they believe that something can count as a genuine (predicative) belief or assertion even if it or its content does not purport to represent the world. I, by contrast, deny this. In my view, essential to something's being a predicative assertion or belief is its being such that it or its content purports to represent reality.

[49] In fact, those who defend the idea that we perform assertion-like acts whose content is comprised of quasi-propositions have rather little to say about their nature. Probably the most detailed account of their nature is found in Horgan and Timmons (2000), although their terminology differs from mine. See, however, Horgan (2002: 330), in which he expresses sympathy for the idea that quasi-propositions do not exist.

expressivism. Contra perspectivalist expressivism, it is not true that when agents engage in ordinary moral discourse they intend to assert moral propositions. Rather, they intend to perform assertion-like acts that express moral quasi-propositions. And contra illusionist expressivism, such agents are not systematically mistaken about what speech acts they intend to perform when engaging in ordinary moral discourse. Once again, such agents intend to perform assertion-like acts that express moral quasi-propositions and do not suppose they are doing otherwise. And contra agnostic expressivism, it is not true that we cannot discern what agents intend to do when engaging in ordinary moral discourse. Rather, ascertaining what agents intend to do when engaging in ordinary moral discourse reveals that such agents intend to perform assertion-like acts and not acts of asserting moral propositions. The sophisticated expressivist can say all these things (in part) because she denies that the syntactic trappings of moral sentences function to *mask* their genuine content. As the sophisticated expressivists sees things, the syntactic features of moral discourse reveal discourse of this sort for what it is: discourse wherein agents express moral quasi-propositions.

Second, by giving illocutionary act intentions their due, sophisticated expressivism avoids commitment to any form of error theory—whether first- or second-order in character. When we maintain that agents who participate in ordinary moral discourse intend to perform assertion-like acts that express moral quasi-propositions and not acts of asserting that express moral propositions, we guarantee that there is both no mismatch between the content of moral discourse and reality and no mistake about what we intend to do when engaging in such discourse.

Third, and finally, once we grant that moral discourse expresses moral quasi-propositions, the sophisticated expressivist can help herself to two realist-looking claims. In the first place, she can say that moral quasi-propositions (or the sentences that express them) are true. And, second, on the assumption that the content of a claim that p is true if and only if it is a fact that p, then she can say that there are moral facts. Granted, these realist-seeming doctrines need to be interpreted aright. An agent's ascribing truth to a moral quasi-proposition that p (or to the sentences that express it) is not thereby to claim that that quasi-proposition represents moral reality (or even that it possesses the property of *being true*). Rather, if expressivists such as Blackburn and Gibbard are right, it is to do something else such as merely endorse that quasi-proposition.[50] Likewise, to say that

[50] See Blackburn (1993: 129; 1998: appendix; 2002: 128). See also Gibbard (2003: 18) and Lenman (2003). I should note that, while in numerous places Blackburn says

it is a moral fact that p is not to claim that there is some entity that is the intentional object of our moral claims. Rather, it is to do something else such as simply repeat the claim that p.[51] As far as sophisticated expressivism is concerned, however, this is as it should be, for as Blackburn says in one place, deflationary views of this kind are so minimalist in character that an expressivist can toss them 'in for free, in the end'.[52]

At this point, however, sophisticated expressivism needs to take another step, for simply insisting that ordinary moral discourse consists in intending to perform assertion-like acts is not itself a satisfactory reply to the Core Argument. We should also want some reason to believe that we perform acts of this kind and not acts of asserting moral propositions when we engage in ordinary moral discourse. (Simply claiming that the reasons for affirming moral cognitivism and sophisticated expressivism are on par won't do; this would yield the so-called weak result.) It is worth stressing, however, that if we grant that there are such things as assertion-like acts and that there are no moral facts, there is a powerful-looking argument to affirm sophisticated expressivism. After all, if we grant these assumptions and also hold that a charitable interpretation of moral discourse demands that we avoid an error theoretic account of moral thought and discourse, then sophisticated expressivism looks to furnish the best explanation of ordinary moral discourse for which we could hope. Given the aforementioned assumptions, sophisticated expressivism, unlike moral cognitivism and other versions of expressivism, can both honor The Expressivist's Guiding Rationale and capture everything we could plausibly mean by the sincere use of moral sentences.

As I read some contemporary expressivists, something like the argument just offered on behalf of sophisticated expressivism constitutes a central rationale put forward in favor of their view.[53] Moral realists resist it along two fronts. They reject either Moral Nihilism, claiming that the sorts of considerations furnished by expressivists in favor of this claim fail to hit

that moral 'opinion' is not in the business of representing moral reality, he sometimes indicates that the quasi-realist can deflate representation too. According to his (1998: 79) take on deflationary views of representation, ' "represents the facts" ' means no more than ' "is true" '—where we understand "is true" in a deflationary way. Given this identification in meaning, I shall assume that what I say about deflationary views of truth holds *mutatis mutandis* for deflationary accounts of representation.

[51] See Gibbard (2003: 18).

[52] Blackburn (1998: 80). Elsewhere, Blackburn (1993: 5) writes, 'It teaches us a great deal about representation and description to learn that they are so cheap to purchase that even the Humean [i.e., the quasi-realist] can have them, along with truth, fact, knowledge, and the rest'. My own judgment is that matters aren't as straightforward as this. See Cuneo (n.d.: chs 5 and 6).

[53] See e.g. Blackburn (1993: 4).

the mark, or the claim that there are assertion-like acts that express moral quasi-propositions, arguing that any sophisticated expressivist position that genuinely captures what we mean when engaging in moral discourse collapses into a form of moral cognitivism.[54] While I have some sympathy with these lines of response, I do not on this occasion wish to pursue them. Rather, I propose that we simply concede for the sake of argument that there is conceptual space for there being assertion-like acts that mimic genuinely assertoric moral discourse (and, hence, for there being deflationary moral 'truths' and 'facts'). Moreover, I propose that we concede for argument's sake that there are no moral facts. I want to suggest that, even if we concede these two assumptions, sophisticated expressivism should be rejected, for the rationale offered on its behalf rests on a further implausible assumption. Identifying this assumption is the first step toward assembling a positive case for premise (5) of the Core Argument.

The assumption I have in mind is that a charitable interpretation of a given range of discourse demands that we interpret the content of that discourse in such a way that it does not turn out to be systematically mistaken.[55] For consider: a charitable interpretation of Euclid's views about geometry is not one that attempts to guarantee that Euclid's views about geometry do not come out false. Likewise, a charitable interpretation of Anselm's views about God is not one that tries to ensure that Anselm's views about God do not come out false. And while I do not propose to offer anything like a developed account of what a charitable interpretation of a speaker's discourse consists in, I suggest that a more nearly adequate account of the *aim* of charitable interpretation is this: the aim of a charitable interpretation of a speaker's discourse is to get at what that speaker is trying to say by way of that speaker's engaging in that discourse. The aim of a charitable interpretation of Anselm's views about God in the *Proslogion* is to try to get at what Anselm was actually trying to say about God by way of his having engaged in theological discourse of a certain kind. To be sure, this will typically involve a certain amount of reconstructing what Anselm says. It may, for example, involve disregarding slips of the pen, ambiguity in expression, infelicitous examples, and the like. And, all other things being equal, it will dictate that we interpret what Anselm says as not being plainly obtuse or blatantly confused. But it is not such that it attempts to guarantee that what Anselm says is *not false*.

[54] See Hale (1993), Rosen (1998), and Dworkin (1996) for examples of this type of strategy.

[55] This assumption is, of course, associated with the work of Donald Davidson. I think Davidson's more careful formulations of the principle of charity do not imply it, however.

Let's now add to this the following point. Getting at what an agent intends to say by way of engaging in discourse of a certain kind is not a matter of mere guesswork; it requires taking into account certain facts about that agent, among which are that agent's convictions about the nature of certain features of reality. For example, when Euclid propounds the parallel postulate, we do not interpret him as trying to say something about two-dimensional, positively curved Riemannian space; rather, we read him as trying to say something about three-dimensional 'flat' space. That's because Euclid explicitly speaks of three-dimensional flat space, and was entirely ignorant of the concept of Riemannian space. Similarly, when Anselm claims in the *Proslogion* that God is a being of which a greater cannot be conceived, we do not interpret him as intending to say something about the pantheon of gods worshipped by the ancient Egyptians; rather, we interpret him as intending to say something about God as God was understood by traditional medieval theists. The reason for this is that Anselm was a traditional medieval theist who rejected Egyptian polytheism.

If this is right, then the aim of a charitable interpretation of what we are doing when we engage in moral discourse is not to guarantee that the content of moral discourse is not false. Rather, it is (roughly) to get at what agents who engage in moral discourse are trying to say by way of their engaging in that discourse. And, thus, it is a matter of attempting to get at what agents who engage in moral discourse intend to say by way of their engaging in it. Moreover, getting at what agents intend to say by way of engaging in moral discourse is not mere speculation; it requires taking into account their commitments about the nature of reality and interpreting what they say in light of those commitments.

Why we should reject sophisticated expressivism

Suppose, then, we accept the assumption that (in a wide range of cases at least) we can discern the relevant illocutionary act intentions of participants in ordinary moral discourse. And suppose also that we agree that a charitable interpretation of moral discourse requires that we do our best to get at what agents are trying to say by way of engaging in such discourse. I now want to suggest that these assumptions generate a difficult type of case for sophisticated expressivism. The type of case on which I have my eye is one in which an agent both sincerely engages in ordinary moral discourse and clearly rejects Moral Nihilism or the claim that there are no moral facts. About putative cases of this type, I want to raise two questions. First, are there cases of this type to be found? And, second, how should we interpret the moral discourse of an agent who figures in such a case?

Let me take these questions in turn. In response to the first question, I submit that there are many cases of this type to be found. In what follows, I want to describe what is, in my estimation, the most vivid example of such a case. Having this case before us will allow us to address the second question I have raised.

Consider a figure whom we can call 'the traditional religious believer'. As I think of her, the traditional religious believer is a traditional Jewish, Christian, or Muslim theist, a person who believes such things as: that a personal God exists; that God has various characteristics such as being the creator of the world, being perfectly good, all-powerful, and all-knowing; that God acts in human history and has revealed God's self in various ways to human beings; that a sacred text (or texts) such as the Bible or the Koran or a particular religious tradition is authoritative on matters of faith and morals; and so forth. Thus described, the traditional religious believer is a theological realist; she rejects all 'naturalistic' accounts of the nature of reality. She is also a moral realist. The traditional religious believer is someone who believes that there are both 'divine' and 'ordinary' moral facts—facts that, on the one hand, concern God and God's activity such as *that God is just, that God is merciful*, and *that God has exercised compassion toward the outcast* and, on the other, concern human beings and their activity such as *that Mother Teresa was compassionate, that Smith's assassination of Jones is wrong,* and *that one ought to give to the poor*. Let's note that the traditional believer needn't have a very well worked-out account of the nature of these facts; she usually believes that they in some way depend on God's nature or will. Nevertheless, the traditional religious believer does not hesitate to invoke such facts to explain states of affairs and events in the world. The traditional Muslim believer, for example, appeals to Allah's mercy when explaining the goodness of creation. The traditional Christian believer, similarly, appeals to God's love and our having wronged God to explain why God became incarnate. Furthermore, the traditional religious believer holds that we experience moral reality in various fashions. She sometimes speaks of being presented with God's goodness in mystical experience and of experiencing the moral goodness of those who have dedicated themselves to lives of obedience to God and the pursuit of charity.[56]

[56] See Alston (1991: ch. 1) for a catalog of experiences of the first kind. One person Alston cites says 'all at once I . . . felt the presence of God—I tell of the thing just as I was conscious of it—as if his goodness and his power were penetrating me altogether' (p. 12). An example of the second kind of experience surfaces in Linda Zagzebski's defense of what she calls a 'pure' virtue theory. Zagzebski (1996: 83) writes, 'Many of us have known persons whose goodness shines forth from the depths of their being. . . . I

I take the foregoing to be a fairly uncontroversial characterization of what a traditional religious believer is. I also take it to be evident that considerations of charity dictate that we should not interpret the traditional believer's moral discourse as the sophisticated expressivist suggests. Given her rejection of Moral Nihilism, we should no more interpret the traditional religious believer's moral discourse as the performance of assertion-like acts in which she explicitly presents moral quasi-propositions anymore than we should interpret Anselm's discourse in the *Proslogion* as the performance of assertion-like acts in which he explicitly presents various 'theological' quasi-propositions. Rather, we should interpret the traditional religious believer's ordinary moral discourse in light of her realist commitments and, thus, as being what it seems: discourse in which she intends to predicate moral features of various kinds of persons, their intentions, actions, and so on. Granted, if this account of the traditional religious believer's moral discourse is correct and naturalism is true, then the propositional content of the traditional religious believer's ordinary moral discourse is systematically mistaken. In that case, her moral and religious utterances are of a piece.

The figure of the traditional religious believer is of heuristic value because she provides a vivid example of someone who both engages in ordinary moral discourse and rejects Moral Nihilism. What I should now like to add is that the heuristic value of the figure of the traditional religious believer extends beyond this, for unlike other characters familiar to moral philosophers such as the amoralist, the radical skeptic, or the ideal observer, the traditional religious believer is not a philosopher's fiction.

Earlier I said that the case I would make in favor of premise (5) of the Core Argument would be empirical in nature. I am now suggesting that the description I have offered of the traditional religious believer is an empirical claim. It is not, I think, a terribly controversial empirical claim to make. Blackburn, in one place, affirms that there are people who suffer from 'such defects' as believing that 'things really matter only in so far as God cares about them'.[57] In our own discipline, numerous philosophers acquainted with the debates surrounding expressivism and

believe it is possible that we can see the goodness of a person in this rather direct way. She may simply exude a "glow" of nobility or fineness of character, or as I have occasionally seen in a longtime member of a contemplative religious order, there may be an inner peace that can be perceived to be good directly . . .'

[57] Blackburn (1993: 156–7). (See, also, Blackburn 2001*a*: part one, in which Blackburn seems to admit that there are those who mistakenly believe that morality in some interesting sense depends on the existence of God.) See also Timmons (2002: 23), in which he states that 'in the minds of many people, there is a deep connection between morality and religion'.

moral realism claim to be traditional religious believers. In so doing, they take themselves not to espouse a highly stylized philosophical position divorced from those of ordinary religious believers, but to adopt a view that ordinary theists have defended and espoused for a very long time.[58] I suppose, however, if one wanted a better feel for the sorts of conviction harbored by ordinary religious believers, the natural place to turn is not to the philosophers, but to the sociologists, for sociologists have paid a great deal of attention to the ordinary religious believer. If what the sociologists tell us is true, the vast majority of the adult population of the United States—some 85 per cent—identifies itself as religious, indeed as theists of some variety.[59] Most relevant for our purposes is that the percentage of what I have called 'traditional religious believers' among those who identify themselves as religious appears to be very high. While there are several ways to measure for whether a person who identifies herself as religious is a traditional believer, the standard way of doing so among sociologists is to identify the manner in which this person claims to interpret sacred texts. According to what sociologists tell us, nearly a third of the surveyed adult American population claims that 'the Bible is the actual word of God and is to be taken literally, word for word'. And, predictably, these numbers soar when we consider the largest subset of religious Americans, namely, regular church-going Protestants. Nearly three-quarters of such folk claim to be literalists, while over 20 per cent claim that the Bible is divinely inspired.[60] What is more, the evidence strongly suggests that the moral views of religious believers do not float free from their theological convictions, but are deeply affected by them. To cite just two examples: the empirical evidence indicates that because religious people have theological convictions about taking care of the poor, they are much more likely (in fact, twice as likely) to give to the poor than non-religious people;[61] the evidence also strongly supports the claim

[58] Mitchell (1980) and natural law theorists such as MacIntyre (1992) are nice examples of this.

[59] Unless noted otherwise, the data I shall use are gleaned from the 1998 General Social Survey, a national, full-probability sample drawn from non-institutional English-speaking persons 18 years of age or older. More precisely, the data tell us that about 85% of the US population identifies itself as Christian, Jewish, or Muslim. Of that 85%, some 82% identifies itself as Christian, with 52% identifying itself as 'practicing Christian'.

[60] More exactly, 31% of the surveyed adult American population claims that 'the Bible is the actual word of God and is to be taken literally, word for word'. 70% of regular church-going Protestants affirm this, while 21% claim that the Bible is divinely inspired. Note also that, in a survey taken in 1996, 75% of Christian evangelicals, who comprise roughly 10% of the Christian population, and 62.5% of so-called fundamentalists said that morals are based on an absolute, unchanging standard. See Smith (1996).

[61] See Regnerus *et al.* (1998).

that religious convictions and the belief that there are absolute standards of morality are the most dominant variables in determining conservative attitudes toward abortion.[62]

For those who are not hard-nosed skeptics about whether sociological data give us any reliable information about what people believe, data of this sort can be a helpful reminder that there is often less than a comfortable fit between the convictions of ordinary folk and those of most philosophers. In any event, I will assume in what follows that these sociological claims are sufficiently well established, and that our best empirical evidence supports the claim that there are millions of traditional religious believers in the United States alone. In light of these data, I suggest that we can simply grant that the sophisticated expressivist is correct to say that there are moral quasi-propositions that nicely mimic genuine moral propositions. I suggest, moreover, that we can simply grant that the sophisticated expressivist is correct to say that agents can say of such quasi-propositions that they are true, provided that the sense of 'true' is sufficiently deflationary and consists in no more than doing something like repeating or endorsing such quasi-propositions. I submit, however, that considerations of charity dictate that it would be mistaken to interpret what traditionally religious folk say as consisting in the expression of such quasi-propositions. Rather, we should interpret what traditionally religious folk are saying when they engage in ordinary moral discourse in light of their theological and moral commitments. Doing so, I suggest, gives us decisive reason to believe that when these agents engage in ordinary moral discourse they (among other things) intend to refer to moral features of God and the world, and do not intend to perform assertion-like acts that, according to the sophisticated expressivist, mimic such acts.

Suppose, then, for argument's sake, that we grant both that there are no moral facts and that there are assertion-like acts that express moral quasi-propositions. Now ask yourself the following question: Are there traditional religious believers? If the answer is 'Yes, there are lots of them', then, I am suggesting, we should also believe that there are many ordinary folks who reject Moral Nihilism. (If you believe that this implies that the account I have offered of the traditional believer's moral discourse is 'too metaphysically loaded', ask yourself whether you believe the foregoing account I have offered of the traditional believer's religious discourse is also too metaphysically loaded.) But if we believe that there are many ordinary folks who reject Moral Nihilism, then, I have been contending, we should also believe that (with respect to a large subsection of such people)

[62] See Emerson (1996). See also Hamil-Luker and Smith (1998).

(5) It is false that, in ordinary optimal conditions, when an agent performs the sentential act of sincerely uttering a moral sentence, that agent does not thereby intend to assert a moral proposition, but intends to express an attitude toward a non-moral state of affairs or object.

And this implies that

(6) Expressivism is false.

An objection

Let me conclude this section by canvassing an objection to a central premise of the argument just offered against sophisticated expressivism. The premise in question is the claim that there are traditional religious believers, and the objection is one that flatly denies this. Perhaps the most obvious way of running the objection is to offer a Wittgensteinian-expressivist interpretation of religious discourse according to which religious discourse is also non-assertoric and entirely consists in the expression of religious quasi-propositions.[63] According to this view, by sincerely presenting quasi-propositions of this sort when engaging in religious discourse, an agent does not thereby intend to say anything about God, but intends to express attitudes toward some non-divine object or state of affairs.

My reply to this objection is brief: I deny that such a view accurately represents what ordinary religious believers intend to say when engaging in religious discourse. It is of some comfort that an expressivist such as Simon Blackburn appears to agree. Here is what Blackburn says on the matter:

> To suppose, for instance, that the world exists as it does because it ought to do so might be the privilege of the moral realist. To suppose that the world exists because God made it is the privilege of the theological realist. If this kind of belief

[63] I won't consider a slightly different case in which religious believers are best understood as being cognitivists and realists about theological discourse, but expressivists about moral discourse. I ignore such a case because it seems to me a very strange hybrid. It asks us to imagine that traditional religious believers believe in such things as God, angels, demons, and so on, but do not believe that there are genuine moral features of these and other entities. It is difficult to see, however, why they would accept the former and not the latter sorts of thing—it can't be because the latter are more 'queer' than the former! Moreover, it is difficult to see, according to such a view, what could be made of the putative experience of qualities such as God's goodness. The official projectivist stance is that in some sense we project our attitudes—in this case, upon God—and read them off that on which we project them (see Blackburn, 1984: 181). But I take it to be fairly clear that traditional religious folk don't think of putative perception of God or God's qualities in this fashion. See e.g. Alston (1991). So, I don't see how the position could avoid being an error theory of some kind.

is intrinsic to first-order theorizing (as in the theological case), the kind of diagnosis of the commitments offered by a projectivist will indeed find error in the everyday practice, as well as in various interpretations of it; this is why a 'Wittgensteinian' protection of religious belief is a kind of cheat. Ordinary religious belief, thought of in an expressive way, involves the mismatch referred to above.[64]

The mismatch of which Blackburn speaks is that which is involved in what I called a 'first-order error theory'. The mismatch obtains because it is intrinsic to both first-order theological theorizing of the sort engaged in by figures such as Augustine, Maimonides, al-Ghazali, and the religious discourse of ordinary religious folk, that persons involved in such theorizing and discourse genuinely believe that God exists, loves the poor, will exercise justice on the part of the oppressed, and so forth, and regularly express these propositions in ordinary religious discourse. To which I add the point that some of our most prominent contemporary philosophers of religion who are well aware of the issues that divide expressivist accounts of religious discourse from ordinary cognitivist ones make it clear that when *they* use theological discourse, they intend to say things about God and God's activity and not just express attitudes of various kinds. According to these philosophers, their use of such language is not idiosyncratic, but entirely in keeping with the religious traditions of which they are a part.[65]

IV. Conclusion

I close by recapitulating the Core Argument I have defended. In its basic form, the Core Argument runs as follows:

[64] Blackburn (1993: 58).

[65] See Alston (1989: 6–7), Plantinga (1983: 19), and Wolterstorff (n.d.). On a different note, Jimmy Lenman has suggested to me that perhaps the moral and theological cases are not on all fours. Imagine an agent, Peter, who decides that J. L. Mackie is right about both religion and morality: both are irredeemably riddled with error. It is plausible to conjecture that Peter would continue to moralize by way of expressing attitudes because the things that concern him in the moral realm would continue to do so, and he would need a language to express these concerns. But arguably he would cease to theologize because it would appear silly and pointless.

In response, I do not think it is obvious that moral concerns would continue to concern Peter while he would view theologizing as silly and pointless. The two may be too intertwined in Peter's life for him simply to abandon theologizing. More importantly, it seems to me that even if the two cases were disanalogous in the sense that the objection specifies, this is irrelevant to the argument I am making. My argument concerns the actual illocutionary act intentions of traditional religious believers who moralize and theologize. Whether they would become expressivists after discovering that traditional religion or morality is false is an interesting empirical question that doesn't bear upon this issue.

(1) In ordinary optimal conditions, an agent performs an illocutionary act of Φing by way of performing a sentential act if and only if that agent intends to Φ by way of performing that sentential act. (Assumption)

(2) If expressivism is true, then, in ordinary optimal conditions, when an agent sincerely utters a moral sentence, that agent does not thereby assert a moral proposition, but expresses an attitude toward a non-moral state of affairs or object. (From The Expressivist's Speech Act Thesis)

(3) Expressing an attitude and asserting are distinct illocutionary act types. (Assumption)

(4) So, if expressivism is true, then, in ordinary optimal conditions, when an agent performs the sentential act of sincerely uttering a moral sentence, that agent does not thereby intend to assert a moral proposition, but intends to express an attitude toward a non-moral state of affairs or object. (From (1), (2), and (3))

(5) It is false that, in ordinary optimal conditions, when an agent performs the sentential act of sincerely uttering a moral sentence, that agent does not thereby intend to assert a moral proposition, but intends to express an attitude toward a non-moral state of affairs or object. (As argued in sections II and III)

So,

(6) Expressivism is false. (From (4) and (5), MT)

I have claimed that premises (1) and (3) are platitudes about illocutionary acts that all philosophers ought to accept. Premise (2) follows (when conjoined with a claim about ordinary optimal conditions) from The Expressivist's Speech Act Thesis while (4) follows from premises (1)–(3). If that is right, this leaves (5) as the only vulnerable premise. I have suggested that, according to a fairly natural reading of some passages, expressivists such as Gibbard actually accept (5). I have also suggested that the most obvious manners in which an expressivist might reject (5) are unpromising. Perspectivalist expressivism fails to take into account the role that intentions play in the performance of illocutionary acts; illusionist expressivism commits the expressivist to an error theory of a certain kind, thus violating at least the spirit of The Expressivist's Guiding Rationale; agnostic expressivism yields the result that it is inscrutable what illocutionary acts we perform when we engage in ordinary moral discourse;[66]

[66] Here I oversimplify. Claiming that it is inscrutable what illocutionary acts we perform when we engage in ordinary moral discourse implies not (6), but the claim that we have strong reason not to believe that expressivism is true.

and sophisticated expressivism falls afoul of our best empirical evidence of what at least many agents intend to do when engaging in ordinary moral discourse. (In fact, if this last line of argument is correct, it suffices also as a reply to both illusionist and agnostic expressivism.)

In closing, let me emphasize two points. First, if sound, the argument I have defended does not directly count against viewing expressivism as a proposal for how we might reconstruct moral discourse. Frankly, I have my doubts about why we should adopt an expressivist account of moral discourse if there were no moral facts. But this is not an issue I will attempt to settle here.[67] Second, I have not argued that, when engaging in moral discourse, agents *never* intend to express attitudes and not assert moral propositions. Perhaps they do. But if they do, I judge that this does not affect the main lines of the argument I have offered. Indeed, it may make expressivism a less attractive thesis than it might seem otherwise. For if expressivism were understood merely as a claim about the manner in which some subset of agents engages in ordinary discourse, it would follow that persons who do not intend to express moral propositions in moral discourse and those who do are saying very different things when engaging in such discourse.[68] In such a scenario, we would have reason to believe that we are often quite literally talking past each other when engaging in ordinary moral discourse. And while one probably cannot rule out that

[67] The view doesn't seem to me obviously preferable to a type of moral fictionalism wherein we assert moral propositions in our ordinary lives and think and act as if there were moral facts while admitting in our more reflective moments there are none. See Joyce (2001: chs. 7 and 8) for a defense of a similar position.

[68] Is the view I defend subject to the same complaint? Well, suppose moral cognitivism is true and that a theist were to claim that eating meat is wrong. Suppose also that in saying this she is supposing that eating meat is wrong because it is forbidden by God's law. Now suppose an atheist were to claim that eating meat is permissible. In doing so, he does not think that the moral propriety of eating meat has anything to do with God. According to the view I've defended, won't one agent intend to say something about an action being contrary to the commands of God while the other won't? If so, how could these two figures possibly disagree with one another?

The problem, I think, is only apparent. After all, if cognitivism is true, these agents disagree about this much: whether eating meat has the property of being wrong. These agents are, accordingly, not talking past each other by predicating different properties of different things, but are saying different things about the same thing (eating meat). To be sure, each agent disagrees about whether the property of wrongness depends on God's commands or not. But, once again, this does not preclude genuine moral disagreement. For suppose we assume that both agents share certain assumptions about the nature of wrongness—e.g. that if an option were right and others wrong, she ought to take the right one, that a right option that is chosen because it is right is always morally justifiable, and so forth. Then the disagreement on this level is also a matter of saying different things about the same thing, namely, the nature of wrongness. The theist contends that the property of wrongness depends on God's will, while the atheist denies this.

this in fact is what happens in ordinary moral discourse—one thinks of Alasdair MacIntyre's claim in *After Virtue* that this is precisely what happens in moral discourse—it is nonetheless a result that most cognitivists and expressivists have been eager to avoid.

REFERENCES

Alston, William, *Divine Nature and Human Language* (Ithaca, NY: Cornell University Press, 1989).

____ *Perceiving God* (Ithaca, NY: Cornell University Press, 1991).

____ *Illocutionary Acts and Sentence Meaning* (Ithaca, NY: Cornell University Press, 2000).

Austin, J. L., *How to Do Things with Words* (Oxford: Clarendon Press, 1962).

Ayer, A. J., *Language, Truth and Logic* (London: Gollancz, 1936).

Bennett, Jonathan, 'The Necessity of Moral Judgments', *Ethics*, 103 (1993), 458–72.

Blackburn, Simon, 'Reply: Rule-Following and Moral Realism,' in Steven Holtzman and Christopher Leich (eds.), *Wittgenstein: To Follow a Rule* (London: Routledge & Kegan Paul, 1981).

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

____ *Essays in Quasi-Realism* (Oxford: Oxford University Press, 1993).

____ *Ruling Passions* (Oxford: Oxford University Press, 1998).

____ 'Is Objective Moral Justification Possible on a Quasi-Realist Foundation?', *Inquiry*, 42 (1999), 213–28.

____ 'Reply by Simon Blackburn', *Philosophical Books* [symposium on Blackburn's *Ruling Passions*] 42 (2001*b*), 1–32.

____ *Being Good* (Oxford: Oxford University Press, 2001*a*).

____ 'Replies', *Philosophy and Phenomenological Research*, 65 (2002), 164–76.

Brandom, Robert, *Making it Explicit* (Cambridge, Mass.: Harvard University Press, 1994).

Bratman, Michael, *Intention, Plans, and Practical Reason* (Cambridge, Mass.: Harvard University Press, 1987).

Cuneo, Terence, *The Normative Web: An Argument for Moral Realism* (MS, n.d.).

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

____ and ____ 'The Moralistic Fallacy: On the "Appropriateness" of Emotions', *Philosophy and Phenomenological Research*, 61 (2000*b*), 65–89.

Dreier, James, 'Transforming Expressivism', *Noûs*, 33 (1999), 558–72.

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

Emerson, Michael, 'Through Tinted Glasses: Religion, Worldviews, and Abortion Attitudes', *Journal for the Scientific Study of Religion*, 35 (1996), 41–55.

Frankfurt, Harry, 'Freedom of the Will and the Concept of the Person', *Journal of Philosophy*, 68 (1971), 5–20.

Geach, P. T., 'Ascriptivism', *Philosophical Review*, 69 (1960), 221–5.

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

Gibbard, Allan, *Wise Choices, Apt Feelings* (Cambridge, Mass.: Harvard University Press, 1990).

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

Grice, Paul, 'Meaning', *Philosophical Review*, 66 (1957), 377–88.

Hale, Bob, 'Can there be a Logic of Attitudes?', in John Haldane and Crispin Wright (eds.), *Reality, Representation and Projection* (Oxford: Oxford University Press, 1993).

Hamil-Luker, Jenifer and Smith, Christian, 'Religious Authority and Public Opinion on the Right to Die', *Sociology of Religion*, 59 (1998), 373–91.

Hare, Richard, *Moral Thinking* (Oxford: Oxford University Press, 1981).

Horgan, Terry, 'Replies to Papers', *Grazer Philosophische Studien*, 62 (2002), 303–41.

____ and Timmons, Mark, 'Nondescriptivist Cognitivism: Framework for a New Metaethic', *Philosophical Papers*, 29 (2000), 121–53.

Joyce, Richard, *The Myth of Morality* (Cambridge: Cambridge University Press, 2001).

Lenman, James, 'Disciplined Syntacticism and Moral Expressivism', *Philosophy and Phenomenological Research*, 67 (2003), 32–57.

McDowell, John, 'Values and Secondary Qualities', in Ted Honderich (ed.), *Morality and Objectivity* (London: Routledge & Kegan Paul, 1985).

MacIntyre, Alasdair, 'Plain Persons and Moral Philosophy: Rules, Virtues and Goods', *American Catholic Philosophical Quarterly*, 66 (1992), 3–19.

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

Mitchell, Basil, *Morality: Religious and Secular* (Oxford: Oxford University Press, 1980).

Plantinga, Alvin, 'Reason and Belief in God', in Alvin Plantinga and Nicholas Wolterstorff (eds.), *Faith and Rationality* (Notre Dame: University of Notre Dame Press, 1983).

Regnerus, Mark, Smith, Christian, and Sikkink, David, 'Who Gives to the Poor? The Influence of Religious Tradition and Political Location on the Personal Generosity of Americans toward the Poor', *Journal for the Scientific Study of Religion*, 37 (1998), 481–93.

Rosati, Connie, 'Internalism and the Good for a Person', *Ethics*, 106 (1996), 297–326.

Rosen, Gideon, 'Blackburn's *Essays in Quasi-Realism*', *Noûs*, 32 (1998), 386–405.

Schiffer, Stephen, *Meaning* (Oxford: Clarendon Press, 1972).

Searle, John, *Speech Acts* (Cambridge: Cambridge University Press, 1969).

Smith, Christian, *Religious Identity and Influence Survey* (Chapel Hill, NC: University of North Carolina at Chapel Hill, 1996).

Stevenson, C. L., *Fact and Value* (New Haven: Yale University Press, 1963).

Sturgeon, Nicholas, 'Critical Study: Gibbard's *Wise Choices, Apt Feelings*', *Noûs*, 29 (1995), 402–24.

Taylor, Charles, *Human Agency and Language* (Cambridge: Cambridge University Press, 1985).

Timmons, Mark, *Morality without Foundations* (Oxford: Oxford University Press, 1999).

—— *Moral Theory* (Lanham, Md.: Rowman & Littlefield, 2002).

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

Wolterstorff, Nicholas, *Divine Discourse* (Cambridge: Cambridge University Press, 1995).

—— 'Philosophy of Religion after Foundationalism III: Locating the Issues' (MS, n.d.).

Zagzebski, Linda, *Virtues of the Mind* (Cambridge: Cambridge University Press, 1996).

This page intentionally left blank

3

Expressivism, Yes! Relativism, No!

Terry Horgan and Mark Timmons

> In an important sense of words, then, the so-called noncognitive view defends neither an ordinary relativism nor a methodological relativism. It is an *answer* to relativism; and it can explain, in part at least, why the errors of relativism are tempting ones.
>
> (Stevenson, 1963*a*: 93)

> The emotivist or prescriptivist is committed to some form of relativism, however little he may like the label. Stevenson, who claimed to have refuted moral relativism, turns out to be himself a kind of moral relativist.
>
> (Foot, 1978: 189)

Expressivism in ethics is a metaethical view according to which (roughly) a typical moral judgment functions to express some psychological state other than a descriptive belief, such as some desire, intention, or other motivating state with regard to some object of evaluation. This metaethical view is, of course, a close cousin (if not identical to) what used to be called 'noncognitivism', and noncognitivists like C. L. Stevenson (as our epigraph indicates) were opposed to moral relativism. Contemporary expressivists like Simon Blackburn (1996, 1998) follow Stevenson in being opposed to relativism. Our own evolving metaethical view is a version of what we call *cognitivist expressivism*, and we, too, are opposed to moral relativism. (see Horgan and Timmons, 2006).

However, some contemporary critics of expressivist views, including Paul Bloomfield (2003) and Russ Shafer-Landau (2003), have followed Philippa

A version of this paper was presented at the *Truth and Judgment* conference, Turin, Italy, 3–4 June 2003, and at the Wisconsin Metaethics Workshop, Madison, 14–16 Oct. 2004. To Paul Bloomfield, Keith Lehrer, Michael Lynch, Russ Shafer-Landau, and audiences at both conferences: thank you!

Foot in claiming that expressivists *are* committed to some form or another of moral relativism despite what such expressivists say. And we suppose Bloomfield and Shafer-Landau would say the same thing about Stevenson's version of noncognitivism. This is peculiar, given the careful attention that Stevenson, Blackburn, and we ourselves have devoted to explaining why expressivist and related views *are not* committed to moral relativism.

Our aim in this paper is to get to the bottom of this issue. We will argue that expressivism has the resources to make good on Stevenson's claim: specifically, we will argue that this view, when properly developed, is not a kind of moral relativism nor does it entail relativism and, indeed, it is an *answer* to moral relativism. Sections 1 through 4 are set up. In them we describe expressivism in ethics (1), and then, after explaining Stevenson's rather narrow conception of moral relativism (2), we move on to what we take to be a proper understanding of the kinds of moral relativism that an expressivist wants to avoid (3). In section 4, we present what we'll call the *relativism objection* to this view. As we understand this objection, the heart of the matter, as you might well expect, has to do with how expressivists are to understand ascriptions of truth (and falsity) to moral thoughts and utterances. And so, in section 5, we sketch an expressivist account of the notion of moral truth and then proceed in section 6 to respond to the relativism objection by explaining why the kind of expressivist view we favor is not a version of any objectionable kind of moral relativism. In section 7 we complete our response by diagnosing one main source of misunderstanding upon which, we think, the relativism objection depends. Although we claim that the relativism objection misfires, we also think that critics of expressivism ought to refocus the kind of worry about expressivism that they mistakenly pose as the relativism objection. In section 8, we briefly consider what this worry may come to and how we would answer it. Obviously, we will not here be able to defend expressivism against all legitimate philosophical worries. But enough will have been accomplished if we can put to rest the relativism objection.

1. What is Expressivism?

The short answer to this question is that 'expressivism' (for reasons that we will get to) refers to a species of metaethical view which includes but is not restricted to noncognitivism.[1] And we will, for the most part, be discussing the relativism objection to expressivism, whose main tenets we are about

[1] For attempts to develop a version of expressivism that rejects noncognitivism, see Timmons (1999), and Horgan and Timmons (2000, 2006). See also Blackburn

to present. However, when discussing the views of Stevenson, we will use the older term 'noncognitivism' (and its cognates), since we'll be quoting Stevenson who uses this label in referring to his and kindred views.

The metaphysical component of expressivism—the component that perhaps largely motivates the semantic component—is moral irrealism:

> **IR:** There are no *moral* properties or relations to which moral terms (and the concepts they express) might be used to refer and, relatedly, there are no moral facts that moral judgments might describe or report.

It should be noted that this thesis is meant to deny not only the sort of moral ontology favored by moral realists, but also the kind of constructed moral properties, relations, and facts that figure in both relativist and nonrelativist versions of moral constructivism.[2] The expressivist view, then, represents a robust form of moral irrealism.

The semantic part of the expressivist's package has both a psychological and a corresponding linguistic component. On the psychological side of things, the central claim is that those psychological states that moral thoughts express are not primarily representational,[3] rather they are a certain kind of evaluative action-guiding state. We can make this idea a bit more precise by introducing a certain kind of content for psychological states that are in the business of representing or describing the world. Let us call such descriptive content, *way-the-world-might-be* content. And let us contrast descriptive content with what we will call evaluative content, that is, *way-the-world-ought-to-be* content. Then, we can put the semantic idea just mentioned as follows:

> **PE:** Moral judgments express psychological states whose primary role is not representational and hence whose intentional contents are not descriptive, way-the-world-might-be contents. Rather, such

(1999: 82–8), who claims that since his form of expressivism ('quasi-realism') allows for moral truth and knowledge, it is not a form of noncognitivism, and R. M. Hare who embraces a nondescriptivist account of moral thought and discourse, but rejects the label 'noncognitivism' because 'this term would seem to imply that I recognize no rational procedure for deciding moral questions' (1985: 96).

[2] Michael Smith's (1994) metaethical view counts as a version of nonrelativist constructivism, and on one plausible reading so does Scanlon's contractualism (1998). For a critical evaluation of these views as versions of moral constructivism, see Horgan and Timmons (1996) on Smith, and Timmons (2004) on Scanlon.

[3] Sophisticated versions of noncognitivism defended by Stevenson (1944), Hare (1952), Nowell-Smith (1954), and Edwards (1955) offer accounts of moral thought and language which allow that moral judgments express both a descriptive belief and some nonbelief psychological state. These hybrid views are properly classified as noncognitivist because they take the noncognitive component of moral judgments to be primary for purposes of understanding their meaning.

> states play some nonrepresentational role (typically a reason guided, action oriented role) and thus their intentional contents are not overall descriptive.

Focusing on the linguistic side of things, the expressivist makes the following parallel claim:

> **LE:** Moral sentences, assertions, and utterances are not primarily representational and hence do not have descriptive, way-the-world-might-be contents. Rather, such linguistic items play some nonrepresentational role (typically a reason guided action oriented role) and thus their intentional contents are not overall descriptive.

We are not sure that this characterization of expressivism is generic enough to capture every metaethical view that it should, but we think it comes close. In any case, let us proceed to make a few remarks about expressivism so characterized.

First, expressivism is not committed to any kind of reductionism about moral thought and discourse. A reductionist version of expressivism would attempt to paraphrase and thus in some sense 'reduce' moral thought and discourse to some familiar type of nondescriptive thought and discourse such as the giving of commands, the expression of affective states such as feelings, or the expressing of desires, or perhaps a complex combination of these types of states. Reductionism so understood is certainly not part of the projects of recent expressivists and it was not the project of either Stevenson or Hare (though these older versions of expressivism have sometimes been misunderstood as committed to reductionism).[4]

Our second point is this. You might be surprised that our characterization avoids all mention of beliefs and of cognitive versus noncognitive psychological states. Our avoidance is due to the fact that we think there are two distinct sorts of belief—descriptive and nondescriptive—and that an expressivist can allow that moral judgments express genuine beliefs, although such beliefs are not descriptive. Thus, we believe one can be a cognitivist in ethics—someone who thinks that moral judgments are beliefs—but also be a nondescriptivist in ethics, claiming that the content of such beliefs is not (or not primarily) descriptive, how-the-world-might-be-content.[5]

[4] The nonreductionism in Stevenson's writing should be clear to anyone who reads his 1944 book. Also, see Stevenson, 1963*b*: 214. Hare claims that his form of universal prescriptivism is not reductive: 'it is no part of my purpose to "reduce" moral language to imperatives' (1952: 2, see also 180). Ewing, 1959: 33–4, calls attention to this point as well.

[5] So as we are defining these terms, although moral cognitivism entails descriptivism, it is possible to reject descriptivism and remain a cognitivist. Some philosophers will

These ideas are central to our 'cognitivist expressivism' (2006)—a story for another occasion.

Third, expressivists are not committed to denying the various deeply embedded features of ordinary moral thought and discourse. One does make, and takes oneself to make, moral assertions: one sometimes thinks that the moral claims of others are mistaken, and one sometimes thinks that changing one's mind about some moral issue is not just a matter of mere taste; and so forth.[6] Many contemporary expressivists want to preserve and account for such deeply embedded features of moral thought and discourse from within an expressivist metaethical view. Expressivists, then, can be preservativists, but they can also be reformists. On one reading of his expressivist view A. J. Ayer (1946) was a reformist. Ayer advocated a version of expressivism (typically called emotivism), according to which ordinary moral judgments are grammatically and logically misleading: although they would appear to be assertions capable of embedding in logically compound constructions and so forth, this surface behavior is misleading. Such judgments, on this view, are really, 'deep down' simply expressions of feeling. So, on Ayer's view, a proper metaethical understanding of moral thought and discourse would require that we radically reform how we think about morality. Reformist brands of expressivism are not popular these days.

Fourth and finally, expressivism is not committed to metaphysical naturalism; one could be a nonnaturalist realist about, say, epistemic properties and facts but deny that there are any such moral facts and properties. But typically it is commitment to a naturalist world view that is behind the scenes motivating the irrealist metaphysical claim (IR) accepted by expressivists.

Enough about expressivism, let us now turn to moral relativism.

2. Stevenson on Moral Relativism

There are many theses in ethics that are called 'relativist' or 'relativistic', and here is not the place to survey them all.[7] In arriving at a characterization of

think that combining cognitivism with nondescriptivism is oxymoronic since cognitivism and descriptivism are often simply conflated. Unfortunately, conflating these ideas has led to an unnecessarily limited menu of metaethical options. We won't defend this claim here, but see Timmons (1999), and Horgan and Timmons (2000, 2006). Hare also distinguishes cognitivism/noncognitivism from descriptivism/nondescriptivism and notes that 'one could very well call them [moral convictions] beliefs yet maintain that they were radically different from ordinary factual beliefs' (1985: 96–7).

[6] See Timmons (1999: ch. 4, 2006) for a more complete list of such features of commonsense morality.

[7] It is typical for philosophers to distinguish moral relativism in the sense under consideration here from *descriptive relativism*—an empirically testable thesis about the

moral relativism that seems to be featured in the relativism objection, let us begin with Stevenson's understanding of relativism in ethics. As we shall see, Stevenson was working with an overly narrow understanding of moral relativism and, moreover, a version which allowed Foot to claim that the noncognitivist is a relativist *malgré soi*.

Stevenson distinguishes what he calls 'ordinary relativism' from 'methodological relativism'. As we will proceed to explain, both types of relativism offer relativist analyses of evaluative concepts and so are properly understood as versions of *conceptual relativism.* Stevenson takes relativism with respect to some area of thought and discourse to represent a kind of meaning analysis of the terms and concepts featured in that area. The main ingredient, then, in a relativist treatment of some area of thought and discourse is what Stevenson calls a 'relative term'. He gives as examples of non-evaluative relative terms 'X is tall' and 'X is moving' which are to be analyzed respectively as 'X is taller than ____' and 'X is changing its distance from ____', where the blanks represent the fact that the terms in question are 'inexplicit', as Stevenson says, with regard to one of the relata involved in the relations expressed by the terms 'tall' and 'moving'. This in turn allows for variation in the truth values of judgments containing relative terms so that two apparently contradictory judgments containing such terms can both be true. The judgments, 'John is very tall' and 'John is not very tall' (where these judgments refer to the same person) can both be true if the (unmentioned) relatum that represents the basis of comparison for judgments about someone's being tall differs in each of these judgments. John might be very tall for a high school student, but not very tall for a high school basketball player.

What Stevenson calls ordinary relativism in ethics represents a philosophical theory about the meanings of such terms as 'right', 'wrong', 'ought', 'good', 'bad', and 'evil', according to which they are all relative terms. More precisely, a statement of the form 'X is good' is to be analyzed as equivalent in meaning to a statement of the form 'X is (approved of) by ____', where the blank is to be filled by reference to some actual or ideal

moral attitudes of individuals and groups according to which there is in fact deep variation in moral standards accepted by individuals or groups. Moral relativism is also to be distinguished from what is often called *circumstantial relativism*, the relatively uncontroversial idea that what is right or wrong for an individual to do in some set of circumstances depends importantly on the morally relevant nonmoral facts that obtain in those circumstances. For a discussion of these various relativistic theses in ethics see Timmons (2002: ch. 3). For further distinctions pertaining to relativism, see nn. 12 and 18 below.

individual or group.[8] Moreover, because 'good' is a relative term, there are alternative permissible ways of specifying the blanks so that the apparently contradictory claims 'X is good' and 'X is not good' may both be true. So, for example, if Joe's claim 'Welfare spending is good' is properly analyzed as meaning 'Welfare spending is approved of by a majority of citizens in Mexico' and Jill's apparently contradictory claim 'Welfare spending is not good' is properly analyzed as meaning 'Welfare spending is not approved of by a majority of citizens in the United States', then the statements of Joe and Jill may both be true.

Methodological relativism, as Stevenson understands it, is the view that terms like 'is justified' and 'is a reason for' are also relative terms. The methodological relativist, then, claims that statements of the form 'A justifies B' where 'A' is some nonmoral factual claim and 'B' is some moral claim, are to be analyzed as equivalent in meaning to statements of the form, 'The belief A will in fact cause people of sort ____ to be more inclined to accept B', where the blank is to be filled by reference to some (actual or hypothetical) individual or group.

Thus, as we've said, both ordinary and methodological relativism, as Stevenson understands them, are species of conceptual relativism. However, as we will proceed to explain, Stevenson's characterization of moral relativism is too narrow. To see why this is so, let us consider some of the philosophical features of Stevenson's moral relativism. First, moral relativism so understood is a *reductive* account of moral terms and concepts—such terms and concepts are analyzed as equivalent in meaning to terms (and concepts) that contain no moral terms or concepts. Second, such reductive analyses are *descriptivist* in that statements containing such terms have purely descriptive content—they purport to describe some state of affairs in the world. Third, such analyses are *naturalistic* in that the state of affairs moral statements purport to describe—states of affairs having to do with the approvals, likings, and in general certain psychological reactions of individuals or groups—are part of the natural world that science purports to describe. Finally, and most importantly for our immediate purposes, Stevenson's way of characterizing moral relativism represents what we call a *relativized content* version of conceptual moral relativism. When someone says, for example, 'Welfare spending is wrong' and her judgment is made in relation to the attitudes of, say, the members of her church, then in her mouth the term 'wrong' *just means* 'is disapproved

[8] Strictly speaking, on Stevenson's view, statements of the form, 'X is good' would appear to be equivalent in meaning to, 'X is ____ by ____', where the first blank is filled by some term of approval (e.g. 'liked', 'favored', 'esteemed', 'commended', etc.).

of by members of my church' and so the content of her claim, when properly spelled out is 'Welfare spending is disapproved of by members of my church.'

So on the basis of this relativized content understanding of moral relativism, Stevenson makes two claims. First, noncognitivism construes moral judgments as expressing, not reporting or describing, certain noncognitive attitudes of some individual or group. Thus, since moral relativism construes moral judgments as purporting to describe the reactions of some individual or group, noncognitivism is not a form of, nor does it entail, moral relativism. Second, Stevenson claims that

> Accordingly, the so-called noncognitive view not only rejects relativism but also locates its error: it claims that relativism blurs the distinction between the direct discourse of 'X is good' and the indirect discourse of 'X is considered good,' and that it thereafter proceeds to mislead us by handling the former expression as though it were the latter. (Stevenson, 1963*a*: 91)

The idea here is that advocates of moral relativism are often guilty of a subtle conflation. Those guilty of it appreciate the undoubted fact that moral judgments are associated with noncognitive feelings and attitudes of the speaker, but then go on to suppose that moral judgments are in the business of simply reporting the attitudes in question. Once the difference between expressing and reporting is sorted out, we can see, so Stevenson claims, that the typical defender of moral relativism is just confused.

Let us leave Stevenson for a moment and consider how we ought to understand moral relativism, given the range of metaethical views that are worthy of the label.

3. Moral Relativism Properly Conceived

Given a very common understanding of moral relativism, this kind of view (most generically speaking) represents a philosophical interpretation of moral thought and discourse that features the idea of 'relative truth'. We note that sometimes, the terms, 'correctness' and 'incorrectness' are used instead of 'true' and 'false' in this context presumably because some philosophers reserve the latter pair of terms for use in connection with descriptive judgments, but allow that nondescriptive judgments can have 'correctness' conditions. In what follows, we will largely ignore such nuanced matters of terminology for they don't bear on the metaethical issues we are considering.

We can perhaps best clarify this central idea by means of the following three theses. In presenting them, let us distinguish two levels of moral commitment possessed by an individual or group. First, there is the level of more or less particular moral judgments about actions, persons, institutions, and perhaps states of affairs. Second, there is the level of basic moral standards (principles) that (ideally at least) are the basis for arriving at more particular moral judgments. (This characterization is rough—for instance we haven't explained what we mean by 'basic'—but we believe this two-levels distinction is clear enough for present purposes. [9]) With this two-levels model in mind, here are the three theses in question.

According to what we will call the *dependence thesis*, the truth of a particular moral judgment of a certain type (e.g. about the deontic status of an action) depends on some relevant set of basic moral standards governing the type of moral evaluation in question (e.g. basic standards of deontic evaluation). (More precisely, the truth of a particular moral judgment will depend on some relevant set of moral standards together with certain nonmoral facts about the object of evaluation.[10]) According to the *constitution thesis*, we may think of a set of basic moral standards as true (in connection with basic standards, some philosophers prefer to talk of correctness), but the truth of a set of basic moral standards is to be understood as constituted by certain specified attitudes of some (actual or hypothetical) individual or group.[11] Finally, according to the *variation thesis*, there is no one set of true basic moral standards governing the truth of all other moral judgments of a certain type. Rather, there are multiple and incompatible sets of true (so understood) basic moral standards. Putting these ideas together and focusing on nonbasic moral judgments, the rough idea is that particular moral judgments (by which we simply mean the 'nonbasic' ones) are properly judged as being true or false, but their truth or falsity is relative to some specified set of basic moral standards (governing the type of moral judgment in question), and there is no one set of basic moral standards that governs the truth and falsity of all moral judgments of a certain type. Let us

[9] In any case, this distinction is very common in philosophical discussions of moral relativism. See, for instance, Brandt (1984).

[10] Note that a moral rationalist might hold this thesis by claiming that there is a *single* set of a priori knowable basic moral principles in relation to which more particular moral judgments are either true or false. No relativism here.

[11] Again, this thesis (together with the dependence thesis) is not sufficient for being a relativist. Some versions of moral constructivism maintain that the truth or correctness of a set of basic moral standards is constituted by the attitudes of some actual or hypothetical group, but go on to claim (contrary to the third, variation thesis) that given the constraints on group membership and the attitudes in question, there is only one correct set of basic moral standards. This kind of view is defended by Rawls (1980).

encapsulate all this with a simple statement of the basic idea of moral relativism:

> **MR:** Particular moral judgments have relativized truth or correctness conditions.[12]

Finally, let us contrast moral relativism with moral objectivism:

> **MO:** Particular moral judgments have objective, nonrelative truth or correctness conditions.

So understood, it is clear that Stevenson's semantic construal of moral relativism as involving relativized content represents but one way in which moral judgments may be understood to have relative truth conditions, but not the only way and surely not (for the relativist) the most plausible way. A moral relativist need not, and should not, offer a relativized content account of the meanings of moral judgments.

Now one might suppose that even if Stevenson's particular semantic construal of moral relativism is too narrow to adequately accommodate all forms of the generic view, he might nevertheless be right in supposing that moral relativism is committed to an essentially descriptivist analysis of moral judgments—moral judgments make descriptive claims which are made true (in part) by some set of moral standards that are implicitly invoked when making particular moral evaluations. The rough idea would be that particular moral judgments about actions, say, purport to attribute some moral property (e.g. rightness) to the action and whether the action possesses this property depends on some set of basic standards that are

[12] Often, moral relativism is characterized as committed to what we will call the 'no genuine conflicts' thesis according to which apparently conflicting moral judgments about, say, some particular concrete doing by an agent at a time can both be true and hence (so far as truth goes) are not genuinely conflicting. Put this way, however, not all forms of moral relativism are committed to this thesis. For example, according to what David Lyons (1978) calls 'agent's-group relativism', an action is right if and only if it accords with the basic moral standards of the *agent's* group (rather than, say, the basic moral standards of the *appraiser's* group). If the group in question subscribes to a consistent set of basic moral standards concerning some type of moral evaluation (e.g. deontic evaluation), then there is only one true or correct moral judgment about the actions of members of that group—the judgment that is implied, or at least accords with, this one set of standards. We do, however, think that some carefully formulated 'no conflicts' thesis will cover the case of agent's-group relativism. The idea is that with respect to what is intuitively the same type of action (say, abortion) performed in what is intuitively the same set of circumstances (in the first trimester of pregnancy for economic reasons), a judgment about the morality of an act of abortion in one group might be true (relative to that group's basic moral standards), but a similar judgment about an act of abortion in another group might be false (relative to this second group's basic moral standards).

implicitly invoked by the judgment. Moral properties, on this picture, are deeply 'moral-scheme' dependent. But, at least given the metaethical territory that moral relativism is taken to cover these days, even this claim would be too narrow. As we are about to explain, moral discourse can be given a relativist interpretation even if one construes moral judgments as psychological states other than descriptive beliefs. Here, we think it will be instructive to consider Foot's main complaint against Stevenson's construal of moral relativism, which will help set up the relativism objection which is the main focus of this paper.

4. The Relativism Objection

Foot objects to Stevenson's characterization of relativism and she claims that, once properly characterized, noncognitivism turns out to be a form of relativism. And, of course, if noncognitivism is a form of moral relativism, it cannot represent a diagnosis of the (alleged) underlying error committed by at least some relativists in ethics.

As we understand the way in which Foot presses the relativism objection against Stevenson's version of noncognitivism, she makes two moves. First, she points out that certain judgments of taste are properly treated in a relativistic fashion, but that such judgments are not plausibly understood as *describing* the reactions of some individual or group. Rather, a relativistic treatment of judgments about a person's being good-looking, for example, does require that we appeal to a relevant set of standards for such claims, and (according to relativism about these judgments) there is no single true (or correct) set of standards in relation to which this class of judgments is objectively true or false. Rather, we say that judgments of the form 'X is good-looking' have relative truth conditions—they are true or false relative to some set of standards. But, to say all this is not to say, along with Stevenson, that these judgments of taste are *about* reactions that are represented by some set of standards. Stevenson is correct in thinking that to understand value judgments as equivalent in meaning to judgments about the reactions or standards of some individual or group is to confuse direct discourse with indirect discourse. But this only shows that what we have called relativized content versions of conceptual moral relativism should be rejected.

Foot's example of judgments of taste leaves it open whether such judgments are primarily descriptive. But (and this is what we take to be her second move) even on an expressivist construal of aesthetic and moral judgments, a kind of relativism is possible—one that mimics the notion of 'relative truth' with respect to a realm of descriptive claims.

The idea would be this: assuming for the moment that moral judgments are properly construed as nondescriptive in nature, such judgments can nevertheless be properly evaluated as 'true' (using this term very broadly) or 'correct' (if one would rather reserve 'truth' talk for descriptive judgments) *relative to* some set of standards.[13] This is relative truth or correctness—a kind of relativism about the proper semantic or semantic-like evaluation of moral judgments—the essential ingredient in moral relativism.

So, if we work with a conception of moral relativism that rejects the idea of relativizing the content of moral judgments in the way Stevenson does and instead we rely on the idea of 'relative truth', and if we recognize that relativists in ethics may also be expressivists, we have a more generic characterization of relativism in ethics that seems to cover the relevant metaethical territory. And according to Foot, once this more accurate characterization is in place, Stevenson's own noncognitivist version of expressivism is committed to a form of moral relativism. Here is how she defends this claim:

> According to the argument just presented relativism is true in a given area if in that area all *substantial* truth is truth relative to one or another of a set of possible standards. And it is now possible to see that individualistic subjectivism may itself be a form of relativism. . . . For even if the truth of moral judgments is not relative to local community standards of the individual it (the substantial truth[14]) could still be relative to the standards of the individual. This is how it is, in effect, in emotivist and prescriptivist theories, since these theories deny the presence of objective criteria, or any objective method by which differences between individuals with radically different basic moral principles could in principle be resolved. If these theories are correct, anyone who queries the truth of a moral judgment, and still possesses the resource of testing it by his more basic moral principles, uses 'true' substantially; but beyond this point he does not. It follows that the emotivist or prescriptivist is committed to a form of relativism. (Foot, 1978: 189.)

Foot's argument here goes a bit fast, partly because it is not clear that a noncognitivist theory can (or should) countenance any *substantial* use of 'true' and 'false'. We will come back to this point later in the paper. But for the time being, let us simply formulate the objection under scrutiny.

[13] Quiz: J. J. C. Smart was an ethical emotivist and he advocated a version of act utilitarianism. Explain how this combination of views could be wedded to an essentially relativist account of the 'correctness' of moral judgments.

[14] Foot borrows the term 'substantial truth' from Bernard Williams (1966, 1974–5), and admits that at least her use of it in this context is not explained sufficiently.

The relativism objection

(1) Versions of moral expressivism deny the objective truth or correctness (and the objective falsity or incorrectness) of moral judgments, and so they deny that moral judgments (including the basic moral standards in relation to which an individual makes moral judgments) can be objectively true or objectively false. (An implication of the irrealist component of expressivism, IR.)

(2) However, expressivism can allow judgments of truth and falsity, or if truth and falsity are only possible in relation to descriptive judgments, then expressivism can allow for some distinction between the 'correctness' and 'incorrectness' of moral judgments.

(3) The only basis for an expressivist account of the truth or correctness of a moral judgment (of a speaker) is the set of basic moral standards of some specified individual or group. And so a moral judgment is true or correct just in case it is entailed by, or in some looser sense 'follows from', some set of basic moral standards together perhaps with relevant nonmoral information about the object of evaluation. This is relativized truth or correctness.

Thus,

(4) On an expressivist view, truth (or correctness) is *relative* truth (or correctness).

(5) Relative truth or correctness in ethics commits one to moral relativism (by definition).

Thus,

(6) At least insofar as an expressivist can make sense of judgments of truth or correctness in ethics, the view is committed to moral relativism.

We believe that this is more or less the objection that is behind all versions of what we are calling the relativism objection to expressivism. And, as we mentioned at the outset, this objection is alive (if not, in our opinion, well) in the recent writings of Paul Bloomfield (2003) and Russ Shafer-Landau (2003: 30–3). But before responding directly to these critics, let us first sketch a conception of truth in ethics that we have presented elsewhere and which we claim is a basis for adequately responding to the relativism objection.

5. What Expressivists should Say about Moral Truth

In section 1, we characterized expressivism as committed to a robust form of moral irrealism. By 'robust' in this connection we had in mind a denial of the sorts of properties and facts featured in naturalist and nonnaturalist versions of moral realism, as well as constructed moral facts featured in both relativist and nonrelativist versions of moral constructivism. In connection with the issue of moral truth, the claim is that there are no 'substantive' truth-makers for moral judgments.

However, these negative ontological and semantic claims are not the end of the story. An expressivist, as we know from the writings of Allan Gibbard (1990) and Simon Blackburn (1998), can make a place for 'true' and 'false' as predicated of moral judgments; specifically, the expressivist can advocate a so-called minimalist use of these terms (and concepts). Indeed, Stevenson, in his 1944 book, was already doing so, though perhaps this feature of his brand of expressivism has not been sufficiently appreciated. (see Stevenson, 1944: 169–73; 1963*b*: 214–20).

We agree with Stevenson (and Gibbard and Blackburn) about the legitimacy of a minimalist use of the truth predicate in ethics, but we have a way of developing this minimalist theme in connection with moral thought and discourse that helps make clear why an expressivist is not committed to moral relativism. Our view involves the following four theses.

1. The concepts of truth and falsity are governed by implicit, contextually variable semantic parameters which allow that judgments employing these concepts (and the terms that express them) may themselves vary from one context to another.
2. In morally engaged contexts, the truth predicate is used to categorically affirm the first order moral judgment of which it is predicated; this (disquotational) usage is not relativistic at all.
3. In morally detached contexts, the truth predicate may be properly used in either a nonrelativistic manner or in an explicitly relativized manner; but the latter usage does not commit one to the idea that morally engaged, categorical, truth (and falsity) ascriptions to moral judgments are implicitly relativistic.
4. Finally, the morally engaged, categorical usage of 'true' and 'false' is semantically primary; the morally detached usage is secondary.

Elaborating and combining these theses provides the expressivist with a convincing response to the relativist's objection. So, let us now proceed to consider (in order) these claims in more detail.

1. Generally speaking, we maintain that ‘true’ and ‘false’, and the concepts they express, are terms and concepts governed by what we call ‘contextually variable parameters’. They are not unique in this respect; the kind of contextual variability we have in mind is ubiquitous in thought and language. In fact, in connection with terms like ‘tall’, we have already noted how a concept (and the term expressing that concept) can be semantically governed by a variable parameter that has to do with relevant comparison classes. This kind of parameter is, of course, a *semantic* parameter because its operation, as it were, bears on the meaning and truth of judgments that employ the concept of tallness. Of course, just saying this much does not represent a philosophical defense of a contextual treatment of the semantics of concepts and terms generally or of truth and falsity in particular. Here is not the place to launch into a defense of such claims; rather, we will proceed to explain how this kind of contextualist treatment of ‘true’ and ‘false’ works out in connection with moral judgments.[15]

2. So, let us suppose that there is contextual variability of the sort in question governing the concepts of truth and falsity (and the terms expressing them). If so, then it is plausible to suppose that there is a kind of contextual variability in the correct uses of ‘true’ and ‘false’ regarding the manner in which these terms are applied to moral judgments. To help explain this rather abstract suggestion, let us suppose that when it comes to moral judgments, there are two main ‘perspectives’ in relation to which predications of truth and falsity may properly take place: a *morally engaged* perspective and a *morally detached* perspective.[16] Let us focus on each one in turn.

Within what we call a morally engaged perspective, one makes categorical, nonrelative moral judgments. For example, one judges that apartheid is wrong, period. Now typical uses of ‘true’ and ‘false’ as predicated of moral judgments take place within a morally engaged perspective as well, where one’s metalinguistic judgment is governed by schema T. So, in thinking or uttering, ‘ “Apartheid ought to be abolished” is true’, one is, in effect, affirming the named first-order moral judgment from a metalinguistic stance. And the crucial point to notice is that one’s higher-order truth predication is itself expressive of a psychological state whose overall content is not descriptive, and thus the overall content of the truth predication is *also* not descriptive. What we have here is a ‘fused’ semantic/moral evaluation

[15] For more on the engaged/detached distinction, see Timmons (1999: ch. 4) and Horgan and Timmons (2002, 2006).

[16] We don’t mean to suggest that there are *only* these two perspectives.

evaluative judgment or claim. This usage of the truth predicate, we claim, is *primary* in connection with moral judgments. (This point will emerge more clearly below as we describe two secondary uses of the truth predicate in ethics.)

Another way of putting this point is to note that from within a morally engaged perspective or context, one uses 'true' and 'false' *disquotationally*; that is, in predicating truth of a first-order moral judgment of the form, say, ' "X is good" is true', one is using 'true' in such a way that this metalinguistic claim entails that X is good. And conversely, from within a morally engaged stance, in affirming that X is good, one is committed to affirming that 'X is good' is true. Again, there is nothing relativistic about the uses of 'true' and 'false' in these contexts.

So, here is our second thesis: the primary use of the truth predicate in moral thought and discourse mirrors first-order moral thought and discourse: such higher-order judgments involve a categorical, nonrelativistic, morally engaged, use of 'truth' and 'false'. Let us call this usage the *disquotational usage.*[17]

3. We now turn to our third thesis, having to do with truth ascriptions from a morally detached perspective. Here is where Bloomfield in particular thinks moral relativism creeps into our view.[18] Let's see.

From a morally detached perspective—a perspective in which one is not thinking and judging in a morally engaged manner—there are (at least)

[17] Notice that saying all this is quite compatible with commitment to robust moral irrealism. On our view and on expressivist views generally, moral judgments are not in the business of describing or representing the world—they are not to be understood as way-the-world-might-be judgments. So, metaphysically speaking, there is nothing to say about what *makes* a moral judgment true or false, where the expectation is to specify some substantive truth (and falsity)-makers for moral judgments. So when we combine our moral irrealism with our view of the proper use of the truth predicate in morally engaged contexts of semantic assessment (assessment in which one is explicitly thinking or saying of some moral judgment that it is true or false), our view represents a kind of 'metaphysical minimalism' about the proper understanding of the primary usage truth predicate in connection with moral discourse. (We will return to the topic of 'substantive' truth below in section 7.)

[18] Bloomfield (2003) distinguishes between what he calls 'normative relativism' and 'metaethical relativism'. As he uses these terms, normative relativism seems to be much like, if not identical to, what we described as a relativized content version of relativism in ethics (what Stevenson was calling 'ordinary relativism'). And Bloomfield admits that our view is not committed to this kind of relativism. He characterizes metaethical relativism as the view that from a morally detached perspective, there can be conflicting but equally true or correct moral outlooks, which is roughly what we have in mind by what we are simply calling moral relativism. Or, more precisely, Bloomfield's characterization is in terms of the 'no genuine conflicts' thesis. See above n. 12. There are other ways in which the normative/metaethical relativism in ethics contrast has been drawn. See for instance Brandt (1976), Carson and Moser (2000), and Wong (1991).

two appropriate uses of the truth predicate that the expressivist can allow, but both uses are *nondisquotational* and neither is primary in ethics. One of these uses is nonrelativistic, the other is relativistic. Let us proceed to consider them in order.

First, the expressivist can recognize an appropriate usage of the truth predicate which is a nonrelativistic detached usage. Under this use, a statement or judgment counts as true or false only if its overall content is descriptive content. Truth, on this usage of the truth predicate, is correspondence; and falsity is noncorrespondence. Thus, on this *correspondence usage* as we will call it, a moral statement or judgment, in order to be (correspondence) true must (1) have overall descriptive content and thus be in the business of purporting to report or represent in-the-world moral facts, and (2) the statement or judgment in question must correctly represent (and thus correspond to) the relevant facts. Correlatively, for a moral statement or judgment to be false it must (1) have overall descriptive content and (2) fail to correctly represent (and thus correspond) to the relevant facts.

This correspondence usage of 'true' and 'false' is proper, we claim, and a common way of usage in metaethics, in connection with the metaphysics of morals. But, given that for the expressivist moral statements and judgments do not have overall descriptive content, they are neither (correspondence) true nor (correspondence) false. These judgments are neither true nor false, as semantically appraised from a morally detached perspective, because they lack overall descriptive, way-the-world-might-be content. Because, for an expressivist, using 'true' and 'false' in this nonrelativistic correspondence manner is to be understood as proper only when one is viewing things from a morally detached perspective, one is *not* therefore denying the relevant first-order moral statements and judgments under consideration when one says (under the morally detached usage) that they are not true; nor is one *affirming* the first-order statements when one says (under the detached usage) that they are not false. This detached usage is clearly different from the disquotational usage; and thus for an expressivist, such usage is secondary.[19]

Let us now turn to another detached, secondary usage of the truth predicate, one that is relativistic. We think it is in connection with this

[19] It is important to notice that our view is not a version of the error theory in ethics. An error theory results when one construes moral judgments as purporting to make claims that would require cooperation from the world in order to be true, but goes on to maintain that the world does not cooperate. This is not our view. And here, rather than digress, let us point out that the work we have done that is critical of competing metaethical views—critical of moral realism, nonrelativist moral constructivism, and moral relativism—*reveals* (we think) that moral judgments are best understood as primarily nondescriptive in nature.

usage that critics have been guilty of conflation, leading them to think that expressivism is committed to moral relativism.

As we have said, 'true' and 'false' are subject to contextually variable parameters. In morally detached contexts, these terms may be properly used in an overtly relativistic manner. But such use is nondisquotational and semantically secondary. This usage, like the primary use (the morally engaged use), is appropriate whether or not the overall content of moral statements and judgments is itself descriptive. The relativistic truth-ascription is descriptive in any case: it reports that a given moral statement or judgment by someone is a semantically appropriate one to make—that is, it is a statement or judgment that is indeed reflective of the person's own moral standards, that does indeed express a relevant psychological state the person is in. To understand this usage more clearly, let us consider an example. Let us suppose that someone utters the following claim:

> According to the moral outlook of group G, it is true that apartheid ought not to be abolished.

This is an explicitly relativized, noncategorical use of the term 'true' because as the remark makes clear, one who utters it is in effect making a claim about some first-order moral judgment *in relation to* the outlook of some group. In thinking or uttering this claim, one is not thereby *affirming* a moral judgment about apartheid as practiced by members of group G or anything of the sort. Rather, all one is doing is reporting a (descriptive) fact about a certain feature of G's moral outlook. To be more precise, according to our view, morally detached uses of 'true' as featured in the above claim are properly glossed as saying:

> The moral outlook of group G includes moral norms that permit (or require) the practice of apartheid.

But using 'true' in this morally detached manner does not commit one to moral relativism. As our gloss on the original claim makes clear, anyone (whether realist, constructivist, relativist, or expressivist) can agree that, as a matter of (descriptive) fact, G's moral outlook may include norms that imply that apartheid is permitted (or required). But thinking and talking in this manner is not affirming any moral judgment. In particular, one is not claiming that apartheid really *is* right for members of group G. Nor is one claiming that apartheid really *is* right for everyone—where one is categorically affirming a moral judgment using the standards accepted by members of G. One is, rather, making a sociological observation.

So, clearly, this sort of morally detached relativized usage of 'true' and 'false' is *not* disquotational: in remarking that according to the outlook of some group apartheid is not wrong, one would not normally be taken as thereby affirming that apartheid is not wrong. Similarly, in claiming that, according to the outlook of some group, the judgment 'Apartheid is not wrong' is true, one is *not* claiming that this moral judgment *is* true. In fact, having noted that the outlook in question implies that there is nothing wrong with apartheid, one may sensibly go on to categorically deny that apartheid is not wrong. Morally engaged uses of 'true' and 'false' are governed by schema T and hence are disquotational; morally detached uses of the sort we have been considering are not disquotational.

4. We come finally to our fourth thesis—the categorical usage of 'true' and 'false' in moral discourse is primary; the morally detached usages are secondary. This thesis really just highlights some differences we have already been making in contrasting morally engaged with morally detached uses of the truth predicate. One basic contrast we have been calling attention to is the fact that there are both disquotational and nondisquotational uses of the truth predicate. Disquotational uses are semantically primary: given the role of the concepts of truth and falsity generally—to *affirm* metalinguistically first-order judgments[20]—it is clear that such disquotational uses of 'true' and 'false' are semantically primary. One comes to understand truth talk in connection with ordinary affirmation and denials of various first-order claims. By contrast, nondisquotational uses—both correspondence uses and explicitly relativized uses of truth talk—are secondary: one understands relativized uses of 'true' and 'false' by detaching, as it were, from more primary uses as a way of semantically commenting on how certain things look (and hence what is thought to be true or false) from the vantage point of some other party, including our former selves.

So, if we are right and the four theses listed above are correct, then once one sorts out the various proper uses of the predicates 'true' and 'false' as understood by our kind of expressivism, one can see that an expressivist view in ethics need not be a version of, or otherwise committed to, moral relativism. Relativism, No!

6. A Reply to the Relativism Objection

In order to secure our anti-relativist position, let us first briefly return to the relativism objection against expressivism and then turn to a diagnosis of what we take to be a kind of confusion that prompts the relativism charge.

[20] Incidentally, this is not to say that this is all there is to truth ascriptions in relation to all types of discourse. Arguably, truth and falsity in relation to descriptive discourse involves metaphysically robust truth-makers.

The premise in the argument that we reject is, of course, the third premise which in effect claims that insofar as expressivists can make sense of moral truth (or correctness), their account must be one of internal consistency among a set of basic moral standards, relevant nonmoral information, and nonbasic moral judgments. More fundamentally, the assumption behind the argument is that an expressivist account of truth in ethics must involve appeal to some substantive 'truth-makers' for moral judgments. This is certainly what Foot was thinking when, in the passage we quoted earlier, she mentions using 'true' *substantially*.

But on our view, categorical morally engaged uses of the truth predicate are not being used 'substantially' in the manner that Foot and other fans of the relativism argument suppose. Nor is it being used in any substantial manner from a morally detached standpoint. Premise 3, we claim, is false.

7. Avoiding Metalinguistic Conflation

That, anyway, is the short of it. But we can say more to help diagnose the presumption, expressed in premise 3 of the argument, that expressivism is committed to relativism. There is a kind of conflation that may make moral relativism tempting and perhaps fuels the relativism objection to expressivism, though not the one Stevenson had in mind. Stevenson, recall, was rather narrowly focused on what we have called relativized content versions of moral relativism. And he pointed out that to understand moral judgments—judgments in which one *expresses* certain attitudes—as equivalent in meaning to descriptive judgments *about* one's attitudes (or the attitudes of some group) is to confuse direct, first-order moral thought and discourse with indirect thought and discourse. But as we have seen, because moral relativism need not be (and typically is not) understood as committed to any relativized content construal of the meanings of moral thought and discourse, Stevenson's diagnosis of relativism does not go to the heart of the matter.

Nevertheless, there is a possible source of conflation concerning the idea of truth-ascriptions in ethics that ought to be avoided, because otherwise the clear difference between expressivism and relativism will be missed. In order to make this clear, we will first offer a general diagnosis of the source of conflation and then we will turn to our critics and explain how, in particular, the conflation manifests itself.

As we've said, our reply to the relativism objection is to charge the critic with supposing that for the expressivist the relativistic usage of 'true' must be the only, or at least the primary, legitimate usage of the truth predicate. But, this charge ignores the contextual parameters that allow for a range of contextually appropriate, but importantly different, uses of the truth predicate in connection with moral statements and judgments. Moreover, the relativism charge ignores the fact that the relativistic usage is decidedly secondary—a disengaged, descriptive-reportive, usage. As we have explained, the semantically primary usage is the morally engaged one, which works in the manner set forth in section 5. It is disquotational, and categorical. In the primary-usage mode, the semantically correct thing to do is to ascribe truth to, and only to, those first-order moral statements that express one's own relevant psychological states—and to ascribe falsity to moral statements incompatible with these. Here, as we've said, these primary uses of 'true' and 'false' as predicated of moral judgments involve a kind of *fusion* of one's moral engagement with an overtly semantic form of appraisal.

So, in effect, the critic charging relativism conflates the primary use of the truth predicate, in moral contexts, with one of its two secondary uses—a conflation that typically occurs in the absence of an appreciation that there are these different, contextually appropriate, uses, and that this fact is reflective of the ways that the notion of truth is subject to contextually variable parameters of proper usage. This is our general diagnosis of the conflation involved in the relativism objection. Now let us consider how the relativist charge is typically lodged and how our general diagnosis applies to the particular charge in question.

As we have said, on our view the typical role of (first person) moral judgment is to express one's moral outlook on some issue. (We think the psychological states in question are genuine beliefs—evaluative beliefs that are not in the business of representing how the world is—but put that aside.) In particular, in judging that, for instance, some action is or would be wrong, there is a semantic norm in play that requires one to judge in accordance with one's overall moral outlook. Call this a *semantic consistency norm*. So, for instance, if Charles embraces as one of his moral commitments the view that intentionally killing an innocent person is always wrong and if, in addition, he believes that a late stage human fetus is an innocent person, then (if he considers the matter) he (semantically) ought to judge that killing a late stage human fetus is wrong (or give up his unqualified general moral commitment). If, on the other hand, Leslie thinks it is not the case that intentionally killing an innocent person is always wrong and she also believes that a human fetus is an innocent person, then she is not bound by the semantic consistency norm in question to judge that it is wrong to kill a human fetus (in cases where she considers the matter). So, in this restricted semantic sense of what an individual ought to judge, we can perfectly well say that Charles would be making a mistake were he to judge that killing a human fetus is wrong, while it would not be a mistake on Leslie's part were she to make this same judgment.

But clearly, given what we have said about various proper uses of 'true' and 'false' in relation to moral judgments, making just such observations about Charles or about Leslie is not to commit oneself to the kind of relativism about moral truth that critics (and expressivists like us) find objectionable. An objectionable truth-relativism would be a view according to which the principal usage of 'true' and 'false' in moral discourse—or perhaps even the only semantically legitimate usage—is relativistic. This is emphatically *not* our view. Rather, to slide from the claim

> (1) that judgment M is required of an individual (or a group) by the norm of semantic consistency under discussion, and so is semantically appropriate relative to that individual's (or group's) moral outlook,

to the conclusion

> (2) that judgment M (as made by that individual or group) is *true*,

is to be guilty of a kind of metalinguistic conflation that should be avoided. In effect, such reasoning conflates the primary usage of 'true' in morals—namely, the morally engaged, nonrelativistic, categorical, disquotational, nondescriptivistic usage—with one of the two secondary usages—namely, the morally detached, overtly relativistic, noncategorical, nondisquotational, descriptivistic usage. Thus, we would deny the relativism charge as posed by Bloomfield when he writes: 'Metaethical relativism infects any position that is committed to thinking that from a morally disengaged point of view, all moral outlooks that are equally consistent from an internal point of view are all equally correct; equal consistency yields equally true moral truth' (2003: 514–15). Granted, insofar as one uses 'true' in relation to moral judgments from a morally detached perspective, one will rightly say that all equally consistent moral outlooks 'yield equal truth': either (1) one will say that these outlooks all fail to yield *any* truths or falsehoods concerning morals, because there aren't any (the nonrelativistic detached usage, which is descriptivistic), or (2) one will say that each of these outlooks yields outlook-relative truths (the explicitly relativistic detached usage, which is also descriptivistic). But, as we've explained, to use the truth predicate in either of these nondisquotational, descriptivistic ways is not to commit oneself to affirming incompatible moral judgments, or to claiming that an action that is right for one individual or group to perform (in some set of circumstances) is wrong for some other individual or group to perform (in the same circumstances), or to claiming that there really are no semantically legitimate truth-ascriptions to moral statements and judgments that are categorical and nonrelativistic. On the contrary.

We are not, by the way, saying that moral relativism, as a metaethical account of the semantics of moral thought and discourse, necessarily involves this conflation of primary and secondary uses of the truth predicate in moral contexts. Rather, the claim is simply that (1) it would be a mistake to conflate semantic consistency with what one calls 'moral truth' in ethics when one is using the truth predicate in the way that is semantically primary in moral discourse, and (2) the slide from the former to the latter may be behind the temptation to think that expressivists must really be relativists in denial.

So, like Stevenson, we maintain that expressivism is, in a sense, an *answer* to moral relativism—at least it has an answer to the charge that it is a (disguised) version of moral relativism.

8. Fear of Expressivism

At this point we think the anti-expressivist should refocus the relativism objection and (to add some rhetorical flair to the objection) claim that our response to the objection is a sort of philosophical bandaid: it covers over something that is *really wrong* with expressivist views. What's really wrong with expressivism (so the critic might plead) is that on this view there is no objective or nonarbitrary backing in relation to moral discourse that, as it were, 'decides' things in ethics (ignoring cases of indeterminacy). When it comes to claims about physical reality, reality decides things. When it comes to mathematics, even if we eschew the need for some mathematical ontology, at least an objective methodology decides things (again, bracketing indeterminacy). But, in rejecting moral realism and (nonrelativist) moral constructivism, an expressivist rejects these ways of how (apart from our actual decisions, of course) things are 'decided' in ethics: there is no moral 'high ground', as Bloomfield puts it. This 'no objective backing of or grounding in an objective reality' is, of course, the robust irrealist part of the metaethical picture. But now that we have set aside the relativism issue, what *exactly* is the objection on offer?

Sometimes, we think, the objection amounts to a kind of fear—fear of expressivism.[21] And one way of articulating this fear is to say that if expressivism were true, then categorical moral evaluations would be groundless. But what does the claim of groundlessness come to? Here, we

[21] Scanlon's 1995 article is called 'Fear of Relativism'.

only have space for a few brief remarks regarding a battery of worries that deserve careful attention.

One possibility is that the critic is raising a *moral* objection: 'If expressivism were true, then what *moral* reason is there for taking moral thought and discourse seriously?' Now a moral question deserves a moral answer, and an expressivist, being invited to engage in moral disputation, can certainly oblige. The question may be construed as a question about the institution or practice of morality generally speaking, as in 'Why should any of us take the whole of moral thought and practice seriously?', or as raising a question about some of our moral practices—practices of criticism as in 'Why should any of us feel justified in taking a critical stance toward others, including our former selves?' Both questions can be answered by the expressivist, though the answers will engage one's moral outlook—as you might expect.

Less charitably perhaps, the complaint sometimes just seems to be: 'This sort of metaethical view is robustly irrealist—no OBJECTIVE moral facts that make true certain moral judgments.' To which we say (with suitable elaboration of philosophical detail and defense): 'Expressivism, Yes!'[22]

REFERENCES

Ayer, A. J., *Language, Truth and Logic*, 2nd edn. (New York: Dover, 1946).

Blackburn, Simon, 'Securing the Nots', in Walter Sinnott-Armstrong and Mark Timmons (eds.), *Moral Knowledge? New Readings in Moral Epistemology* (New York: Oxford University Press, 1996).

____ *Ruling Passions* (Oxford: Oxford University Press, 1998).

Bloomfield, Paul, 'Is there Moral High Ground?', *Southern Journal of Philosophy*, 41 (2003), 511–26.

Brandt, Richard B., 'Ethical Relativism', in Paul Edwards (ed.), *The Encyclopedia of Philosophy*, iii (New York: Macmillan, 1976), 75–8.

____ 'Relativism Refuted?', *The Monist*, 67 (1984), 279–307. Reprinted in Moser and Carson (2000).

Carson, T. L., and Moser, P. K., 'Introduction', in Moser and Carson, 2000: 1–21.

Edwards, Paul, *The Logic of Moral Discourse* (New York: Free Press, 1955).

Ewing, A. C., *Second Thoughts in Moral Philosophy* (New York: Macmillan, 1959).

Foot, Philippa, 'Moral Relativism', The Lindley Lecture, University of Kansas, 1978. Reprinted in Moser and Carson (2000): page references are to the reprinted article.

[22] Which, of course, should not be taken as a breezy affirmation of (our brand of) expressivism. We acknowledge that there is hard philosophical work to do in defending our view and we have tried to make some headway in doing it (see Timmons, 1999; Horgan and Timmons, 2000, 2006).

Gibbard, Allan, *Wise Choices, Apt Feelings* (Cambridge, Mass.: Harvard University Press, 1990).

Hare, R. M., *The Language of Morals* (Oxford: Oxford University Press, 1952).

——— 'Ontology in Ethics', in T. Honderich (ed.), *Morality and Objectivity: Essays in Memory of John Mackie* (Oxford: Routledge, 1985). Reprinted in R. M. Hare, *Essays in Ethical Theory* (Oxford: Oxford University Press, 1993): page references are to the reprint.

Horgan, Terry, and Timmons, Mark, 'Troubles for Michael Smith's Metaethical Rationalism', *Philosophical Papers*, 25 (1996), 203–31.

——— and ——— 'Nondescriptivist Cognitivism: Outline of a New Metaethic', *Philosophical Papers*, 29 (2000), 121–53.

——— and ——— 'Conceptual Relativity and Metaphysical Realism', *Philosophical Issues*, 12 (2002), 74–96.

——— and ——— 'Cognitivist Expressivism', in Horgan and Timmons (eds.), *Metaethics after Moore* (Oxford: Oxford University Press, 2006).

Lyons, David, 'Ethical Relativism and the Problem of Incoherence', *Ethics*, 86 (1976) 107–21. Reprinted in Moser and Carson (2000).

Moser, P. K., and Carson, Thomas L. (eds.), *Moral Relativism: A Reader* (New York and Oxford: Oxford University Press, 2000).

Nowell-Smith, Patrick, *Ethics* (New York: Philosophical Library, 1954).

Rawls, John, 'Kantian Constructivism in Ethics', *Journal of Philosophy*, 77 (1980), 515–72.

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

——— 'Fear of Relativism', in R. Hursthouse, G. Lawrence, and W. Quinn (eds.), *Virtues and Reasons* (Oxford: Oxford University Press, 1995). Reprinted in Moser and Carson (2000).

Shafer-Landau, Russ, *Moral Realism: A Defense* (New York and Oxford: Oxford University Press, 2003).

Smith, Michael, *The Moral Problem* (Oxford: Blackwell, 1994).

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

——— *Facts and Values* (New Haven: Yale University Press, 1963).

——— 'Relativism and Nonrelativism in the Theory of Value', in Stevenson (1963). Cited as Stevenson 1963*a*.

——— 'Retrospective Comments', in Stevenson (1963). Cited as Stevenson, 1963*b*.

Timmons, Mark, *Morality without Foundations: A Defense of Ethical Contextualism* (New York and Oxford: Oxford University Press, 1999).

——— *Moral Theory: An Introduction* (Lanham, Md.: Rowman & Littlefield, 2002).

——— 'The Limits of Moral Constructivism', *Ratio* 16 (2003), 391–423. Also published in P. Stratton-Lake (ed.), *On What we Owe to Each Other* (Oxford: Blackwell Publishers, 2004), 90–122.

——— 'Objectivity in Moral Discourse', *Elsevier Encyclopedia of Linguistics and Language*, 2nd edn. (2006).

Williams, Bernard, 'Consistency and Realism', *Proceedings of the Aristotelian Society supplementary vol. 40* (1966). Reprinted in B. Williams, *Problems of the Self* (Cambridge: Cambridge University Press, 1972).

____ 'The Truth in Relativism', *Proceedings of the Aristotelian Society* (1974–5). Reprinted in B. Williams, *Moral Luck* (Cambridge: Cambridge University Press, 1981).

Wong, David, 'Relativism', in Peter Singer (ed.), *A Companion to Ethics* (Oxford: Blackwell, 1991).

4

Anthropocentric Constraints on Human Value

Justin D'Arms and Daniel Jacobson

According to Cicero, 'all emotions spring from the roots of error: they should not be pruned or clipped here and there, but yanked out' (Cicero, 2002: 60). The Stoic enthusiasm for the extirpation of emotion is radical in two respects, both of which can be expressed with the claim that emotional responses are never appropriate. First, the Stoics held that emotions are *incompatible with virtue*, since the virtuous man will retain his equanimity whatever his fate. Grief is always vicious, both bad and bad for you, even when directed at events commonly considered tragic, such as the loss of one's child. Second, they buttressed this view with an account of the nature of emotion and its relation to value. Emotions are evaluative judgments that are *systematically false*, because they attribute significant value or disvalue to external things and events beyond an agent's control—such as wealth, honor, pain, and death—which are merely 'indifferents': neither good nor bad.[1]

Although neo-stoic views continue to attract some defenders, contemporary philosophers typically regard the Stoic antipathy toward emotion

We are grateful to the audience at the UW-Madison Metaethics Workshop, and we wish to thank the National Endowment for the Humanities and the Princeton University Center for Human Values for their generous support. We would also like to thank Sarah Buss, Nir Eyal, Stephen Gardiner, Pamela Hieronymi, Mark LeBar, Philip Pettit, and Peter Railton for their helpful comments on an earlier version of the paper. Any views, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect those of the National Endowment for the Humanities.

[1] See Nussbaum (1994). We will not be concerned with the famously problematic notion of 'preferred indifferents' used by the Stoics to avoid paradoxes about rational action, since they claimed that even preferred indifferents have no real value.

as extravagant. The most striking problem with this aspect of Stoicism is that it seems so psychologically unrealistic. It is tempting to think that the plausibility of ideals for how human beings should live and what they should value must depend partly on facts about human nature. And surely our emotional sensibilities are among the central features of a distinctively human life—the sort of feature that demands at least some degree of accommodation, rather than extirpation, in ideals that could be well suited to human beings. Nevertheless, not every aspect of actual human interest or concern should be enshrined in the ideals to which we aspire. Some things people care about are bad, or bad for them.

Such thoughts leave proponents of a realistic moral psychology with the task of trying to say something more substantive about the ways in which facts about human beings make a difference to ethics and the theory of value. This is a formidable task, on which many naturalistic approaches to ethics have foundered. Rather than approaching it directly, this paper hopes to make progress by focusing specifically on challenges to the emotions like those of the Stoics. (Since we will not be doing classical scholarship, we will henceforth refer to these challenges as stoic with a lower-case 's'.) We hope to use insights garnered from a critique of the stoic program of emotional extirpation in order to motivate some more general observations about the respects in which certain values might depend upon, or be constrained by, human psychology.

We will focus primarily on the claim that emotional responses are inappropriate because they are systematically false, in that they accord importance to matters of indifference. Most philosophers are inclined to reject this suggestion in various modest and local ways, for instance by suggesting that particular emotions (such as pity, indignation, or admiration) are often responses to genuine value or disvalue. By contrast, we will attempt to defend a hypothesis that may seem as extravagant as the stoic claim, though antithetical to it. We contend that all of what we call the sentiments—a class we will presently circumscribe—are sometimes *fitting*. Very roughly, that is to say: we claim that every sentiment has some instances which descry genuine and distinctive forms of value (such as the funny and the shameful). This is true not only of the more attractive cases but also of several emotions philosophers often reject as vicious, including disgust, envy, and jealousy. These more dubious emotions are also responsive to real, albeit human, values. This ambitious claim is supported in part by the observation that the concerns embodied in sentiments play what we will call a *wide psychological role*. By this we mean that these sentiments are sensitivities to things that matter to people in various ways, not limited to their emotional responses. Nevertheless, these human values descried by the sentiments cannot be understood independently of them; our view is therefore a form of sentimentalism.[2] Our arguments rest partly on some speculative empirical claims, which may ultimately be disproved. But even if our most ambitious hypothesis is wrong, its defense develops some novel suggestions about the connection between a particular sort of value and the distinctively human concerns manifest in the emotions.

We begin by examining stoic objections to two sentiments, regret and envy, in order to define some crucial terms necessary for homing in on the relevant argument for extirpation. Next, we will explicate several respects in which certain values are distinctively human, and offer some general reasons supporting our hypothesis that all sentiments are sometimes fitting. Finally, we return to the case of envy—perhaps the sentiment about which stoic claims are most tempting to moderns—and argue in more detail that it too is sometimes a fitting response. We will not be defending envy by 'prettying it up', as its few philosophical champions are wont to do, and then defending only its most benign instances.[3] We admit that envy is just as morally dubious as its antagonists suggest, while claiming that it is sometimes a sensitivity to value nonetheless.

1. Envy and Regret, Zealotry and Evangelism

Let us begin by characterizing the class of states on which we will focus, and then turn to distinguishing different strands in the stoic critique of emotion. The sentiments are syndromes of thought, feeling, and motivation, which constitute a core subset of the broad and diverse group of states that commonly get called emotions. First, they have certain characteristic thoughts, which can be given only a rough-and-ready gloss, even in principle. (While traditional cognitivist theories of emotion attempt to differentiate emotion-types by their 'constitutive thoughts', we think this enterprise fundamentally mistaken, at least when it comes to the paradigmatic class of emotions we're calling the sentiments—but it would take us too far afield to pursue that argument here.[4]) Second, they typically involve feelings caused

[2] For more detail on sentimentalist theories of value, see D'Arms and Jacobson (2006). There we argue that the core thesis of sentimentalism is that (some form of) value is not merely descried by the emotions but partly constituted by our emotional propensities.

[3] For instance, some authors draw a distinction between so-called friendly or admiring envy and an invidious or malicious strain (Roberts, 2003; Neu, 1980). Our defense of envy eschews any such distinction, which we find at best misleading and at worst confused.

[4] We argue against the cognitive theory of emotion, including its most sophisticated 'quasi-judgmentalist' version, in D'Arms and Jacobson (2003). We also argue there that

by physiological changes, often accompanied by a specific expression, such as laughter, tears, or blushing. Third, and perhaps most important, they each have a specific, though often complex, pattern of motivation.

Sentiments are distinct from evaluative judgments, we hold, and in this respect we depart from the stoic theory of emotion, which is an especially stark form of cognitivism. Yet we accept that they involve taking things to matter in some distinctive way, which is amenable to interpretation. They can therefore be appraised in terms of their fittingness. This form of rational appraisal, which we will explicate in the following section, focuses on whether a sentiment's object really matters in the way one takes it to matter when in the grip of that sentiment. For instance, to think that your amusement at a joke is fitting is to judge the joke funny—not merely to be amused by it. Sometimes we are amused by jokes that are not really funny, perhaps because we are giddy from being awake too long; and sometimes we fail to be amused by a genuinely funny joke, because we're depressed or we've heard it before.

The sentiments are paradigmatic types of emotion such as amusement, anger, contempt, disgust, envy, fear, guilt, jealousy, joy, pride, regret, shame, and sorrow, which we believe will prove to be robust psychological kinds. This list is provisional, and we are not crucially committed to including all its members, but we will note that it closely resembles the lists of pan-cultural emotions adduced by psychologists with disparate theoretical approaches, such as Paul Ekman (a defender of affect programs) and Richard Lazarus (an appraisal theorist).[5] Many of these syndromes come together as suites of response because they are, or are products of, relatively discrete psychological mechanisms that evolved for their adaptive value in dealing with 'recurrent adaptive situations' (Tooby and Cosmides, 1990) or 'universal human predicaments' (Johnson-Laird and Oatley, 1992).

The claim that all sentiments are sometimes fitting entails that these specific states correspond to distinctively human values. This must not be confused with the far broader claim that everything that might be called an emotion is sometimes fitting. We will use the term 'emotion' more broadly, so as to include various object-directed and affect-laden states, including those fine-grained attitudes that specify their objects more precisely (like *schadenfreude*: pleasure taken in the misfortune of others). Emotions can be type-identified however one likes, though not every such grouping will be

emotions can be appraised on grounds of fittingness even though they are not well understood on the cognitivist model. Nothing in the present argument depends upon our rejection of the cognitivist theory, and readers who are congenial to it can understand fittingness as the truth of an emotion's constitutive thought.

[5] See Ekman (1994) and Lazarus (1994).

equally fruitful. Thus, one can take all the instances of too-much-pride-at-things-that-don't-actually-speak-all-that-well-of-you, for example, and treat them as a type of pride: 'false pride', perhaps. So understood, false pride is, by stipulation, never fitting—it has been picked out so as to ensure that result. False pride is an example of what we will call a *sharpening* of the sentiment pride. Sharpenings are constructed by specifying a subclass of instances of a sentiment in terms of some characteristic they happen to share.

There are indefinitely many possible sharpenings, some of which already have names, such as homesickness and fear of flying. Cognitive sharpenings have a common thought or judgment, motivational sharpenings a common motive, and causal sharpenings a particular sort of elicitor. Normative sharpenings group emotions by appeal to some shared normative classification. Most important for present purposes are cases like false pride, which are grouped together as fitting or unfitting. Our claims here do not apply to sharpenings or other emotional states, only to the sentiments. It is controversial how many such pan-cultural syndromes exist, and we take a more capacious view than some; but our argument does not crucially depend on whether envy and regret, or any other specific example, count as sentiments.

Consider some common questions about the appropriateness of envy and regret. For different reasons, these two sentiments are among those most likely to be deemed always inappropriate.[6] Thus Rüdiger Bittner (1992) argues, following Spinoza, that regret is always 'unreasonable', and that therefore one should never feel it.[7] His argument presumes that it is possible to make the evaluative judgment characteristic of regret—very roughly, that one has done something bad—dispassionately. Hence, one does not count as regretting a decision just in virtue of thinking it mistaken and wishing one had done otherwise. That is an evaluative judgment, whereas regret is an emotion. We accept this distinction; indeed, we insist upon it. Yet we reject his conclusion that regret is always unreasonable.

Bittner also must contend that the dispassionate attitude of recognizing one's mistakes without regretting them is psychologically possible. While surely this is sometimes true, it is far-fetched as a more general claim about how people are capable of feeling about their mistakes. He writes:

> To look things straight in the face, unburdened by [regret], is just a very good idea, and a good idea in the same sense in which it is a good idea, say, to have the snow

[6] As we have stated previously and will explain presently, the claim that a sentiment is or is not appropriate is crucially ambiguous. See esp. D'Arms and Jacobson (2000*a*).

[7] We will take Bittner's interpretation of Spinoza on faith; doubters can replace 'Spinoza' in what follows with 'Bittner's Spinoza'.

> tires mounted before the first snowfall. True, there is this difference: some people do mount their snow tires in time, but few people are likely to take in their failings without regret. Still this shows only that some forms of unreason are more common and harder to eradicate than others. (Bittner 1992: 272)[8]

Bittner thus concludes that regret is always unreasonable, even if few people are capable of avoiding it. His argument is simple and superficially compelling: regret is painful, and that is a reason not to feel it.

In order to show that regret is always unreasonable, however, it does not suffice to adduce a reason never to feel it. Rather, one must show that there are not more weighty countermanding reasons in its favor. If the balance of reasons favored feeling regret on some occasion, then it *would* be reasonable (in the relevant sense), despite the countermanding reason given by its painfulness. In order to land his sweeping claim, Bittner must argue that one can get the benefits of regret without paying the cost in suffering. And indeed he makes just this argument, by claiming that, insofar as it is possible to avoid repeating one's mistakes, one can do so without regretting them. Bittner thus endorses Spinoza's conception of the reasonable person as one who, when he errs, 'understands what he did, he knows that it was bad, it just does not pain him' (1992: 267). Spinoza is not telling us always to avoid the misery of making mistakes. Since we are imperfect agents acting in uncertain conditions, we could not follow the useless advice to always choose correctly. It is not error but regret that is claimed to be always unreasonable; and it is unreasonable because it pointlessly adds pain to the badness of error. Bittner concludes, 'that is what regret is, double misery, the second for the sake of the first. So regret is not reasonable' (1992: 265).

An obvious problem with this argument springs to mind, along with a less obvious source of potential confusion. The problem is that, even if all the benefits of regret can *in principle* be had without the added suffering of regret, this achievement seems rather more difficult in practice. Bittner's argument against regret combines two dubious psychological claims: he is overly optimistic in one respect and overly pessimistic in another. First, he is too optimistic about the possibility of eliminating regret. 'Feeling is not in principle insensitive to insight' (1992: 264), he writes, and we agree. But the real question is not a matter of principle. It is whether regret can be

[8] Bittner uses the word 'grief ' rather than 'regret' in this passage, but he clearly means to be talking about the specific sort of grief characteristic of regret—indeed, his thesis demands it. Moreover, since his argument rests on the painfulness of the state, it applies both to instances of regret that are more like (self-directed) anger or guilt and to those more like grief.

eliminated in practice, by humans, and at what price. The human capacity for criticizing and changing our emotional dispositions and norms, though real and important, is not unlimited or without cost. So Bittner's quick argument will not do. What he needs is evidence that it is psychologically possible to eliminate regret from our mental lives and, moreover, that the benefits of doing so outweigh the costs. Only then could one conclude that regret is always unreasonable.

Second, Bittner seems excessively pessimistic about the likelihood that the pain of regret effectively motivates people to learn from their mistakes. 'Regret does not in fact make doing better in the future more probable' (1992: 267), he claims without elaboration. The haste with which Bittner makes this assertion suggests that this is another argument about what is possible in principle, which might be secured on a priori grounds. However, the pertinent question concerns whether human tendencies to regret in fact prove useful in helping us to avoid repeating our errors. And the answer to this question cannot be given simply by observing that, either in principle or even in some actual cases, people can recognize and learn from their failures without the pain of regret. Bittner here needs an a posteriori, economic argument, comparing the costs of regret to its benefits, rather than an aesthetic argument focusing on the sublimity of facing one's errors without misery or illusion.

We also warned of a potential confusion about Bittner's argument. It is important to note that his argument for the unreasonableness of regret does not speak to whether or not regret is fitting: that is, to whether people have grounds for regret. Indeed, he grants that some actions and decisions are regrettable. [9] 'Spinoza is not uselessly advising never to have grounds for regret', Bittner notes; on the contrary, when such grounds exist, 'the reasonable person without regret has done something bad' (1992: 265)—that is, something regrettable. The painfulness of regret is a legitimate reason not to feel it—though there may be stronger reasons in its favor—but it would be the *wrong kind of reason* to bring to bear on the question of whether some decision or action was regrettable.[10]

Bittner, to his credit, does not make this mistake. His Spinoza is not a *zealot* about regret: someone who claims that it is never fitting. He

[9] This is one of the places where ordinary language can mislead: 'Regrettable' often just means bad, without any implication of a mistake by the agent (or anyone). The sense of 'regrettable' meaning 'befitting of regret' should be familiar, but we do not claim it to be the only meaning of that term. In this paper we will just stipulate that 'regrettable', 'enviable', etc. are terms serving to express the response-invoking concepts closely tied to regret, envy, and the like. We explicate this further in the next section.

[10] On this point, see D'Arms and Jacobson (2000*b*) and Rabinowicz and Rønnow-Rasmussen (2004).

is instead an *evangelist* against it: someone who always opposes feeling regret—whether because it's inadvisable or wrong. Bittner's evangelism rests on the plausibility of his optimism about the eliminability of regret and his pessimism about its utility. Since we find both claims dubious, we are unconvinced by the argument; but we are not primarily concerned here with arguing against evangelism about the emotions. Our main concern is with zealotry. And, as Bittner recognizes, the zealot's claims are independent of the evangelist's. Even if regret, or any other emotion, were always wrong or inadvisable to feel, it would be far from obvious that this provides any support for the claim that it is always unfitting. The wrongness or inadvisability of feeling regret does not speak to regret's fittingness: to the question of whether one has made a bad decision.

Compare envy, an emotion that is both painful and ugly. These facts provide two good reasons not to feel it—albeit reasons that do not bear on whether envy is fitting. Moreover, zealotry is far more tempting for envy than regret. Whereas it is quite difficult to deny that regret is ever fitting, many philosophers are tempted to make this claim about envy. But what exactly is the zealot claiming about envy? To answer that question, one needs to determine envy's generic locus of concern. Here we will simply state a conclusion for which we have argued elsewhere (D'Arms and Jacobson, 2000*a*; D'Arms, 2002). An episode of envy presents some difference in what might be termed 'position or possession' between the agent and some rival as being bad for the agent. But this gloss must not be taken too literally. As we will explain, just what counts as a possession, for the purposes of characterizing envy's concern, is driven not by some independent notion of the concepts *possession* or *rival*, but by the best interpretation of patterns in what people envy and why, and in the motivations people display when in the grip of that emotion.[11] And similarly for pride, amusement, fear, disgust, and so forth.

We can now voice the zealot's challenge to envy, which we take to be his most intuitively compelling case. The challenge is that what really matters to one's flourishing is what one has, not what others have that one lacks. Differences in possession are never bad *per se*; hence envy's concern is systematically mistaken. Thus envy is always unfitting, and nothing is truly enviable. Or so the zealot claims.

[11] What elicits envy is malleable, but not entirely plastic, across times, cultures, and individuals. Note that such nonmaterial, even abstract, things as accomplishment, reputation, and status can be the object of envy; it's not just about the accumulation of stuff. Indeed, in a less materialistic culture, the gloss of envy would be better put in terms of position than possession; but since positions are (at least metaphorically) possessed, this poses no great worry.

2. Human Nature and Human Value

We have claimed that all sentiments are sometimes fitting, and that they both descry and circumscribe a distinctive realm of human values. These suggestions issue from a general theory concerning the relationship between emotion and value, which we call *rational sentimentalism*. While we cannot explore the details of that theory here, we attempt in this section to lay out some of its central elements in a programmatic way, so as to sketch what we hope is a compelling view of the connection between sentiments and a certain group of values. In addition to clarifying our central notions (such as fittingness and human value), we aim to motivate the importance of the under-theorized class of values on which our discussion focuses. We will try to demonstrate that the category we call human value is worthy of the name, despite being grounded in and constrained by contingent facts of human psychology.

Judgments of fit

The best way to understand the notion of fittingness is to begin by investigating the role of fittingness judgments: ask what it is to *think* an emotion fitting rather than for it to *be* fitting.[12] We will offer an account of these judgments as constituting a distinctive form of normative regulation for the emotions. Though we defend these practices of emotional regulation, this defense will not address itself explicitly to the metaphysical status of facts about (the right kind of) reasons to feel. Rather, we suggest that evaluative practices that presuppose such reasons play a crucial and perhaps ineliminable role in our actual moral psychology—or in any moral psychology available to humans. When we later turn to offering substantive claims about when emotions are fitting, we will be engaging in such regulation. But our defense of these claims rests on general and abstract considerations, which rely in part on our theory of the function of this kind of evaluative discourse. We think the zealot is mistaken partly because the norms of fit he embraces are ill-suited to the regulation of emotion in creatures like us.

Appraisal in terms of fit is a specific and familiar form of rational assessment of emotions. It is the assessment one makes in thinking a joke merits amusement or an action calls for outrage. Roughly, our proposal is

[12] In adopting this approach, however, we do not deny that these judgments have truth values.

that *to think an emotion a fitting response to some object is to think there is (pro tanto) reason, of a distinctive sort, for feeling the emotion toward it.*[13] Such judgments differ from other sorts of appraisals of an emotion, in virtue of the kind of reasons they concern. Reasons of fit are those reasons that speak directly to what one takes the emotion to be concerned with, as opposed to reasons that speak to the advisability or propriety of having that emotion.[14] So reasons of fit for fear are roughly those that speak to whether or not something is a *threat*. (The consideration that fear of some threats is counterproductive, or that a brave person would not be afraid, are reasons telling against fear which might ultimately win the day—but they are not reasons of fit.)

In our view, fitting emotions correspond to different ways of being valuable; they do not merely descry different things, or different aspects of things, that are all of some generic value. Hence, the sentimental values we defend are plural. They can conflict with one another, and with the demands of morality or self-interest. The values descried by fitting emotions are avowedly human values, normative for emotional responses to which humans are prone but other possible valuers need not be. To think an emotion F (e.g. shame, fear, amusement, etc.) a fitting response to some object X amounts to thinking that X is Φ (shameful, fearsome, funny, etc.).[15] Here and throughout this paper, we utilize one familiar sense of terms such as 'shameful', 'fearsome', and 'funny'—the sense in which they call for the relevant responses.[16]

For example, to think a trait shameful is to think that there is reason to be ashamed of it because it is bad in the distinctive way that shame presents

[13] We use the term 'object' broadly to refer to whatever the emotion's characteristic appraisal is directed at—whatever one is pleased at or bothered by. This might be some material particular (literally, an object) as when one is afraid of that tiger; or it might be something more general as when one is afraid of death; or it might be some state of affairs such as fear that the bridge will crumble while one is crossing it. Reasons of fit are pro tanto because, while they are not defeated by countermanding considerations against feeling an emotion, they can be outweighed by such considerations.

[14] An emotion can be held to be vicious in part because it is unfitting. Hence, the same consideration—the same discrepancy between object and response, in this case—can be a moral reason and a reason of fittingness; but these are different judgments.

[15] Our notion of fittingness is closely related to John McDowell's (1998) talk of 'merited' responses and Allan Gibbard's (1990) notion of what it 'makes sense' to feel, although both authors seem to differ with us on some particulars. See D'Arms and Jacobson (2006).

[16] Ordinary language uses such terms in various other ways as well, and terms such as 'enviable' and 'regrettable' are perhaps most often used in other senses than ours. Our account of the function of Φ-concepts offers an explanation of why this is so, but we will not address ordinary language further here.

it as being. But just what way is that? Somewhat contentiously, we will say that shame presents something as a *social disability*. This constitutes a gloss of shame's generic evaluative presentation: it is an attempt to interpret the emotion's characteristic concern in terms that help to clarify it, despite being inevitably vague and potentially misleading. (To say that shame presents something as shameful would be more accurate, but rather less edifying.) Such glosses are not definitional or stipulative.

The proper method for glossing an emotion involves abstracting away from those features that seem peculiar to specific instances. These interpretations focus on commonalities in the characteristic causes of an emotion, the motives it provides, the thoughts and wishes accompanying it, as well as facts about what would palliate the emotion and what seem to be its 'paradigm scenarios' (de Sousa, 1987). The principle of charity can help to adjudicate between rival interpretations, by favoring those glosses that make better sense of the occasions on which the emotion is felt. Since we deny that sentiments have constitutive thoughts, however, glosses must inevitably be rough and ready, and open to interpretive dispute. This helps explain the ubiquitous and largely inconclusive internecine disputes among cognitivists about the precise contours of supposed defining propositions; and it suggests that such disputes are interminable—not just apparently but in fact. No gloss in emotion-independent vocabulary will perfectly capture a sentiment's locus of concern. But this is not to say that one gloss is just as good as another, nor to denigrate the attempt to provide such glosses. On the contrary, the search for (inevitably tendentious) terms with which to express the characteristic concern of an emotion is crucial for persuasive and reflective purposes.

People routinely disagree in their views about when emotions are fitting, and it is important to distinguish between interpretive and evaluative disagreement. First, some disagreements arise from different interpretive norms. If you and I take shame to be concerned with different things, then we can expect to disagree about what befits shame. Thus, some philosophers suggest that shame presents an aspect of oneself not merely as bad but also as essential to one's character. Someone who accepts this gloss might argue that shame at being underdressed for an important social occasion is unfitting because it is clearly not an essential trait. We dispute this gloss of shame, and deny that any connection to one's identity is required. Hence, this trait might be shameful, depending on the particulars of the case. But we would not deny that our antagonist is making a judgment of fittingness—unlike Bittner's Spinoza. We dispute this judgment of fittingness not as an equivocation or category mistake, but because we reject the underlying account of shame's concern.

Secondly, there are common evaluative disagreements over what traits are shameful.[17] Someone who accepts our gloss of shame as a concern for social disability can disagree with us over what counts as such. We can disagree about whether something, such as inappropriate dress at a formal occasion, really is a social disability. This claim can be denied even by someone who grants that, as a matter of fact, society takes such things to be shameful. Indeed, we expect that everyone will want to hold that some things societies have deemed shameful are not so. Of course someone who finds himself in such a society will be subject to contempt and the other forms of coercion with which social norms are enforced. Though he regards these reactions of others as unfitting, he can recognize that the reactions impose real harms on him. But these extrinsic harms are fundamentally different from the disvalue of having some shameful attribute.

A principal virtue of this account of judgments of fit is that it clarifies what is at issue in disagreements over when emotions are fitting, even among people with disparate evaluative standards. What is at issue is whether or not there is a reason of the relevant kind for feeling this way. This will be something people can sensibly disagree with us about even if we find their views repugnant. Nothing we have yet said attempts to settle these evaluative questions, but our theory does impose some structure on their resolution by requiring them to proceed within a framework of relevant considerations determined by the character of the emotion itself.

Emotional regulation

As many have observed, emotions affect people in ways that they cannot govern at will. What then is the point of making judgments about when we have reason to feel? Were the emotions utterly impervious to reflection, like itches, there would be no point in it—as there is no point in asking whether you have reason to feel itchy. But in fact emotions are amenable to some degree of regulation by means of norms, and especially by norms of fit. When I conclude that it would be unfitting to feel guilty at giving some benefit to my friend rather than a stranger, that can both change how I feel about the situation and set a precedent for what feelings I can endorse as fitting toward others' similar behavior. Yet the capacity to regulate emotions in accordance with norms is imperfect, to say the least. Sometimes one's responses are *recalcitrant*: they continue despite one's considered judgment

[17] In discussion of what is shameful or fearsome or regrettable, people often assume that they share an understanding of the emotion's concern and, hence, that their disagreements are evaluative rather than interpretive. But this assumption is sometimes mistaken.

that they are unfitting.[18] Nevertheless, since normative reflection imposes some governance on our responses, we need to think about when there are reasons to feel.

For several reasons it is important for humans to regulate their emotional responses. Emotions involve powerful motivational tendencies, so regulating them is an indirect way of regulating behavior. Furthermore, because emotions are characteristically pleasant or unpleasant feelings, they reinforce or punish the behavior that provokes them. When you find yourself feeling guilty over what you've done, for instance, you will tend to be less prone to act that way again. Such effects are often salutary, but they need not be. Unfitting guilt provoked by someone's unreasonable anger at you might deter you from acting properly the next time. While the sufficiently reflective and strong-willed can overcome such aversions, others cannot, and many decisions must be made too quickly for reflection. What seems like a good idea, unreflectively, is strongly influenced by previous rewards and sanctions—often paid in the coin of positive and negative emotional response. Since sentiments insinuate themselves into evaluative thought and action in various ways, people need to regulate them in order to function as agents who plan courses of action. The tendency to be angry or contemptuous of certain sorts of action, for instance, can lead us to judge such acts wrong or contemptible, to refrain from them, and to shun those who act—or merely feel—differently.

Emotional tendencies can also unseat evaluative convictions reached on the basis of theoretical deliberation. For example, someone convinced that certain rules of etiquette are vestiges of bourgeois attitudes, which ought to be overcome rather than respected, might be undone by the disposition to be ashamed of being the only person at the parties he attends who acts accordingly. The mutually reinforcing nature of emotions and norms for their fittingness gives point to reflection on our emotional tendencies, and to the attempt to regulate them with reasons. Finally, as Gibbard (1990) stresses, there are especially powerful reasons to seek coordination of social emotions (such as anger, shame, and contempt), stemming from the need to avoid and resolve conflict. Indeed, the desire to coordinate with others, even over feelings that have no direct connection to action, runs deep. People seek shared standards for feelings such as amusement, disgust, and sorrow, and find it unsettling to discover that they have idiosyncratic sensibilities.

These considerations provide good reason to regulate emotional responses. But why should one attempt to do so by means of judgments of fit, which seek reasons that speak to the distinctive concerns of the

[18] See Greenspan (1988), Gibbard (1990), and D'Arms and Jacobson (2003).

emotions themselves? Why not focus instead on Spinoza's notion of the reasonableness of regret and the like, attempting to regulate emotions on the basis of the consequences of feeling them? This strategy would be unsatisfactory for several reasons. First, people do not care simply about whether emotions do them good, or are useful to society at large; we also care about having fitting feelings and about the values they descry. We sometimes think prudentially about when it would be counterproductive to get angry, for instance, but we almost always care about the objects of anger's concern: insults, slights, and outrageous behavior—at least when directed at us or those with whom we sympathize. Moreover, while all attempts to regulate emotional responses are at best imperfectly effective, considerations of fit seem generally much more effective than are prudential or moral assessments of feeling these ways, when it comes to influencing how we actually feel. The thought that a vicious dog can 'smell' your fear gives you good reason not to fear him, but this is much less likely to mollify your fear than is the observation that the dog is too old and decrepit to do more than bark. Finally, even when there is most reason *not* to feel a fitting emotion, norms of fittingness can serve to mark and acknowledge the significance of the human value forgone for the sake of better reasons. At any rate, so we will suggest.

We must acknowledge some limits on the aspiration to settle questions of emotional fittingness, not to mention the broader question of what (there is most reason) to feel. Indeterminacy pervades these issues. Sometimes an emotional response may be neither fitting nor unfitting, in that our norms permit but do not require it. Then differences between those who are amused by something and those who are not, for instance, are what Hume called 'blameless disagreements' of sensibility, which one should not attempt to arbitrate. The tendency to adopt this neutral stance seems to vary dramatically with the particular sentiment at issue, as well as with personality—some people being notably more prescriptive than others. But almost all of us are willing to fight over some such judgments, for instance by insisting that someone's sexual orientation is not a matter befitting shame, regardless of social norms to the contrary. In general, the more closely an emotion's characteristic concern is connected with questions of conduct and social interaction, the less tempting it is to think that disagreements in response are blameless. Thus people are far more likely to eschew judgments of fittingness with respect to amusement or disgust than shame or anger.

Reasons to feel

Judgments of emotional fit involve claims about reasons to feel, we propose; but who do we claim have such reasons? Almost all humans, when they are in the right context—a notion we will explicate presently. And we think that almost all humans are sometimes in the right context to feel the emotions corresponding to each of these values. Hence, there is another sense in which these are human values, beyond their relation to specifically human responses: the reasons they provide are reasons for humans, but not for all possible evaluators. In order to be subject to the reasons some Φ-concept invokes, one must have a susceptibility to the relevant F. Only those capable of amusement—or, as we will colloquially put it, those who have a funny bone—have reason to be amused by a witty remark. But everyone with a funny bone has reason to be amused by Wilde's last words (subject to the contextual provisos), whether or not they actually would be amused—that is, regardless of whether they have a good sense of humor.[19] Surely this claim is sufficiently bold and imperious. It seems excessive to insist further that all rational agents, including humorless aliens (were they to exist), have such reasons. Our point is that almost all humans have the emotional capacities necessary for being sensitive to what we are calling human values.[20] Some autistics, sociopaths, and other outliers may lack these capacities; if so, then they do not have the reasons given to the rest of us by such values.

In addition to possessing the relevant emotional capacities, a creature must be capable of rational self-regulation of its emotions, at least to some extent, in order to be subject to the reasons given by human values. Thus, while Molly the dog has the capacity for fear—which she manifests toward both the kennel and the bath—we do not claim that she has reason to fear the kennel but not the bath. It is true that the kennel is bad for her, since there she is deprived of the comforts of home, whereas the bath is just mildly unpleasant. But, although dogs are susceptible to fear, they are beyond the pale of the fearsome as an evaluative concept. Roughly, then, when X is Φ, almost all humans have reason to feel F at it. This is rough, though, because it ignores the crucial qualifications about context that we have hitherto postponed considering.

Common sense has it that a shark attack is a fearsome prospect, for example. Yet most people never have reason to fear sharks, simply because they don't swim in shark-infested waters. Similarly, while it is shameful to steal from one's friends, you have no reason to be ashamed of that unless you do it. In general, then, for something's being Φ to give any particular person reason to feel F toward it, he must be properly situated—that is, he

[19] Namely, 'Either that wallpaper goes or I do.'

[20] In this respect our defense of human values follows Peter Railton's (2003) account of aesthetic value, which has influenced our thinking, though in some other respects the accounts diverge.

must be in the right context with respect to the value in question. There are various difficult issues about context, but because it is impossible to fully specify contextual parameters in advance, we must leave such judgments elliptically context-relative.

When it is granted that an emotion would be fitting, we claim that anyone in the right context has reason to feel F, irrespective of his values and emotional propensities. People have such reasons, that is, whatever they happen to feel on some occasion, however they are prone to respond, and whether or not they themselves judge the feeling fitting. Suppose we can agree that it is shameful for a professor to deliver sloppy and misinformed lectures. Then there is reason for him to be ashamed of his behavior even though he is not ashamed, is not prone to be ashamed, and denies that what he is doing is shameful. This bullet we are prepared to bite.[21]

One further complication must be noted, since it will be significant in what follows. With respect to the prudential emotions, whose characteristic concern is best interpreted in terms of what is good or bad *for the agent*, questions of fit are more deeply relational. It is contentious within the philosophy of emotion how many emotions are prudential in this respect. We think that some are but others are not, and that this is often a matter of degree. Fear, grief, and envy are at least partly prudential, for instance, whereas amusement, shame, and disgust do not concern the agent's interests. Of course, it is in one's interests to be *prone* to nonprudential emotions such as amusement, but that is a different point. Our lives go better when they are leavened with a sense of humor; and without the tendency to be pained at one's inabilities by being ashamed of them, we would lose an important source of motivation for self-improvement. Even disgust helps us avoid intercourse with things—like spoilt milk—that are bad for us. None of this, however, shows that amusement, shame, or disgust are *about* one's interests, merely that it serves one's interests to be sensitive to these values.

[21] The analogous bullet may seem less appetizing in the case of amusement. As long as you have a funny bone, we are claiming, you have reason to be amused by what is genuinely funny. Does this mean that a Chinese peasant, who speaks no English and is culturally a world apart, has reason to be amused at the quips of Wilde or Berra? It does not. One only has reason to be amused at a joke in the right context, and this surely requires at least that one have heard the joke and understood it. Nevertheless, we do have to accept another implication that might seem unattractive. On our view of judgments of fit, someone who has a bad sense of humor, but is in the right context to appreciate Wilde, fails to see the humor in witticisms that do give him reason to be amused. We think it is not implausible that people are committed to such claims more often than they might suppose. For now we will simply note that we are not forced to conclude that this character with a bad sense of humor has reason to *do* anything (such as buying tickets for the revival of *The Importance of Being Earnest*).

In contrast, whether something is fearsome for you—and, hence, whether you have reason to fear it—depends partly on your interests. This is true even when you are in (what would be) the proper context for fear. The same object might pose a threat to one person and not to another. But even here the relational element focuses not on the agent's tendency to feel F, nor even on the sensibility manifest in his judgments about what is Φ, but on what is good or bad for him. Hence, the conditions under which it is fitting for a person to fear, to grieve, and to be envious are influenced by differences in our interests. Whereas the shamefulness of the negligent professor and the humor of Wilde's quips, by contrast, do not depend on anything about the interests of the person whose feelings are being appraised.

Anthropocentric constraints

We shall now turn to defending our initial claims that all sentiments are sometimes fitting and, moreover, that almost all of us sometimes have reason to feel them. These are normative claims and, as such, they do not simply follow from our account of human value concepts and reasons to feel. A zealot could in principle accept that account, granting our theory of what it is to think something shameful and the like, and agreeing that human values give almost all of us reasons to feel. Being a zealot, he would then insist that nothing is Φ: there are no human values, only human value concepts that are never instantiated.[22] We think it a virtue of our theory that it makes sense of what the zealot is claiming and helps to isolate our disagreement with him.[23] But although our account of human value concepts gives a clear meaning to the zealot's claims, it is not neutral or without normative consequence. Once one understands these concepts as fundamentally in the business of regulating human emotional responses, as we suggest, it becomes harder to see the point of adopting standards that are doomed to fail at this task.

What then do we claim is fearsome, enviable, fitting of anger, shameful, and so forth; what substantive norms of fit are we inviting you to accept? We wish to steer a course between two familiar positions, each of which we find untenable. The first is a crude sentimentalism according to which

[22] Another style of zealot might try granting that some things are (say) funny—even in our reason-conferring sense—and yet deny that this suffices to make it worth calling the funny a value. We suspect that such skepticism issues from the conviction that all values must, in the first place, provide reasons to *act*, not merely to feel. While we reject this conception of value, counterargument would take us too far afield; we are content to argue here that Φ-values are reason-giving for feeling.

[23] This is one reason for preferring our account of Φ-concepts to various possible dispositionalist and speaker-relativist accounts of such concepts. See D'Arms (2005).

any 'normal' emotional response is fitting. One would have to spell out normalcy, of course, but neither of the two obvious options is compelling. If normal responses are those that issue from a properly functioning emotional mechanism, then there will be disparate and conflicting normal responses; most things will be both Φ and not-Φ. But if normal means statistically normal, then the view will be committed to enshrining some abhorrent tendencies, which we see no reason to accept simply in virtue of their popularity. The opposite error is a bloodless rationalism according to which all ideals, including evaluative norms about the Φ, are unconstrained by actual human concerns. As a general thesis about Φ-concepts, this sort of rationalism seems bizarre. It is hard to imagine a defense of standards for what is funny or disgusting, for instance, that was not guided in part by human propensities to these responses. This position will appear more tempting in cases such as the fearsome and the enviable, where a rationalist will be able to articulate standards that seem to speak to the concerns characteristic of the emotion—and, perhaps, to reject them.

Against these two rivals we offer rational sentimentalism, which sees human values as animated by our emotional sensibilities, yet answerable to reason. On this view, although normal human emotional tendencies can be criticized and rejected, some actual human concerns nevertheless impose constraints on the tenability of norms of fittingness. In particular, human concerns that are not merely 'deep' but 'wide' supply such constraints, and make it the case that the sentiments are sometimes fitting for almost all of us. Deep concerns are those that are firmly entrenched in their possessors, such that it would be either impossible or extremely costly to excise them. We suspect that regret is a deep concern for humans, which explains why we find Bittner's suggestion that humans should do without regret overly optimistic.

Wide concerns play a broad psychological role in the mental economy of their possessor. When the object of a concern prompts a variety of evaluative attitudes, not just a single emotion or desire; when desire for it (or aversion to it) arises in many different situations; when it is implicated in the ability to get or avoid many other things people care about; when its pursuit or avoidance grounds disparate actions and plans; when, in short, it is firmly enmeshed in our web of psychological responses, this is evidence of the width of a concern. Hence, width is often part of the explanation of depth: the further some concern reaches into different sources of motivation, the more difficult it is likely to be to eliminate. We hope that this suffices to capture the general idea of a concern's having a wide psychological role, although of course much more remains to be said.

We now venture the empirical conjecture that the best glosses of sentiments will be in terms of concerns that have a wide psychological role in almost all human beings. These are concerns for such things as threats (for fear), losses (sorrow), social disabilities (shame), slights and outrages (anger), contamination (disgust), and so forth—if we've got interpretive matters right. Our conjecture, then, is that these concerns mesh with a wide range of human interests, and that this fact constrains what norms of fittingness it is tenable to hold for these sentiments. In particular, it grounds our claim that they are all sometimes fitting.

This claim might seem to make the sentiments somehow self-ratifying: their concerns matter because they matter to us humans; and because they matter to humans, they count as human values. But our argument is not that anger is fitting because humans are ineliminably prone to anger, and so forth for the other sentiments. It is crucial to our defense of these emotions that the things they concern are not merely of interest to the emotions themselves. If the concern with differences in possession between yourself and a rival were merely a product of envy, then it might be a deep concern—if envy turns out to be difficult to eradicate—but it would not play a wider psychological role. Were it possible to eliminate the tendency to be envious, then we would no longer be concerned with such things. That is, so it would be if envy did not play a wide role in human psychology. In the final section of this paper, we will offer some reasons for thinking that this is not true about envy; rather, the perhaps ugly truth is that concern for the things to which envy is sensitive matters to people in a host of ways, many of which are quite independent of the disposition to be envious. But of course we could be wrong about envy. Even some deep-seated concerns and desires are much less likely to mesh with other desires and attitudes and, hence, are relatively narrow in their psychological role. This seems likely for the widespread and tenacious human taste for salty, fatty, and sweet foods—an appetite that has a plausible evolutionary explanation but is now merely vestigial (under local conditions).

Consider anger, by contrast. The stoic and Christian foes of anger like to point out that it involves being prone to acts of retaliation that can be both costly and wrong, and we do not dispute this point. Yet anger is not just a passion for vengeance. It also manifests concern for social regulation, which focuses on personal slights and social transgressions. Moreover, the concern people take in respectful treatment, and in ensuring that others are complying with rules of conduct, is not itself dependent upon anger; rather, the kinds of transgression with which anger is concerned are important on other grounds. We think that similar arguments can make plausible the claim that all sentiments have a wide psychological role. This is an empirical claim, of course, and hence subject to refutation; but this implies that our arguments do not make the emotions self-ratifying, since their concerns might be shallow, or they might be deep but not wide.

Even if we are correct empirically, however, our claim that psychological facts constrain the tenability of norms of fittingness is still contentious. Rationalists can point to a sublime Socratic ideal of a person so self-sufficient in his virtue that he does care about honor, wealth, or even life; or to an impartial observer whose only concern is to maximize net happiness. If nothing matters but the state of one's soul, and no harm can befall the virtuous person, then there is truly nothing to fear. If honor, status, and possession are indifferents, then envy is never fitting and there is nothing shameful (except perhaps for the lack of virtue).

Notice, however, that this means it does not matter whether your company is an appealing prospect to others, or is something they seek to avoid. The fact that you have characteristics that render you unable to function well in society is claimed to be no blemish and to provide you no reason for shame. But why should the fact that the stoic has been able to describe a logically possible human being who can embrace these consequences be thought to show that they are suitable standards of fittingness for humans? If you find, as we do, that your concern for not having social disabilities is supported by and supportive of a wide range of interests that structure your social life, then it should take a very powerful argument indeed to unseat that concern. The observation that we would not care about such things, were we a different kind of creature, does not suffice to show that they are not of human value.

The zealot is mistaken because his claim flouts human nature. There are two ways to make this point. One is to do what we have already done: to point out all the things that would have no value if the zealot were right, and then count on you, our human audience, to recognize this as a *reductio ad absurdum* of the zealot's position. The other way is to remind ourselves of what judgments about fittingness are, and why we bother to make them. The purpose of these judgments is the regulation of one's emotional responses. These are responses that we will be subject to whatever norms we try to accept, and that furthermore will make a substantial difference to our thinking and our actions. They will guide much of our unreflective behavior and our intuitive sense of what outcomes are worth pursuing and avoiding. There are very good reasons, then, for adopting norms that have significant psychological traction with our emotional propensities. But if the zealot's standards are as psychologically unrealistic as we suspect, then to adopt them would be, in effect, to abdicate this important form of self-governance.

3. In Defense of Envy

We understand envy as presenting some difference in possession, broadly construed, between the envying agent and a rival as being bad for the agent. The zealot claims that this emotion is systematically unfitting, because such differences do not matter in themselves. While some privations are bad for you, surely the fact that someone else has the good you lack does not make matters worse. Can envy be defended against this admittedly plausible challenge?

It is important to be clear about what the zealot must establish, lest his argument gain illegitimate support from some irrelevant points. Perhaps it seems obvious that the rival's possession of the good cannot be bad in itself: better for someone to benefit from a possession than no one. We grant that it is better, from an impartial point of view, that someone possesses the good than that no one does. It does not follow, however, that this is a better state of affairs from the perspective relevant to envy. Compare resentment, understood as a morally sensitive emotion. That is, when resentment focuses on what others have, it involves a moral complaint—such as that the possession is unjust—which can be made impartially. (This is not to say that resentment is always fitting, of course; people often resent others unjustifiably.) Envy, by contrast, is better interpreted as taking the rival's possession to be bad *for the envier* in a particular way.[24] So if envy can be fitting, that is because some such differences are indeed bad for the envier in the way envy suggests. This is what the zealot must deny.

Our claim to the contrary, that envy is sometimes fitting, rests in part on the psychological proposition that envy's concerns are both deep and wide. While we readily acknowledge the difficulty of establishing either claim definitively, we will put forward some considerations in their favor. Envy's concerns run deep, we think, in that eradicating the propensity to envy in most people would be extremely difficult and, even if possible, very costly. (Various ascetic and utopian programs have attempted to eradicate envy, of course, and we take their failure—at least, their failure to win many adherents—to be evidence for the thesis that envy has

[24] There is a sense in which this interpretation of envy is less 'charitable', but this is not interpretive charity. Our way of understanding envy makes it an admittedly unattractive emotion. But any interpretation that tried to treat envy as concerned with something less objectionable—with only the unjust possessions of others, say—would be confounded by the data. It would fail to explain what brings envy about and what makes it go away, and by the fact that people are more prone to envy local rivals than distant millionaires.

psychological depth.[25]) But our argument rests more heavily on the claim that what envy cares about plays a wide psychological role, in that concern for fundamentally comparative matters can be found in several forms of motivation other than bouts of envy. Many of the things people are interested in count as evidence of the width of envy's concern. One is the ubiquity of sport and other forms of competition across cultures. Another is the desire to win, which is different from, though obviously related to, the interest in competition. Indeed, people are moved to do well by comparative standards in many respects, even when they do not focus on the performance of others. While it is easy to imagine a world with the benefits of achievement without the costs of envy, as it is easy for Bittner's Spinoza to imagine a world in which we learn from our mistakes without ever feeling regret, it is much more difficult to square these pretty pictures with a realistic moral psychology. Moreover, even if this were possible, a clear-eyed view of life without competition and comparative excellence would leave most of us, even those who are dubious about envy, with a sense of loss for what would be missing.

The comparative matters at the focus of envy's concern are commonly called *positional goods*. These goods are essentially relational: for one person to do better with respect to such a good, it is necessary that another do worse.[26] Obvious examples include one's position in social hierarchies of various kinds. To support the claim that envy is sometimes fitting for almost all human beings, we will argue for the ubiquity and importance of positional goods. But first we need to explain why establishing the value of positional goods helps demonstrate the fittingness of envy.

The paradigm cases of envy concern goods that contribute to determining an agent's social position. With respect to these goods, the achievements of others can clearly damage his standing. For instance, when a rival employee gets some special commendation, this gives her an edge over the agent in their competition for promotion. If the promotion (a positional good) is granted to matter, then it is bad for the envier not only that he did not get the commendation, but that she, this particular rival, got it. It's bad because it helps her get the promotion, which might otherwise have gone to him. Note that it would *not* have been equally bad for him were the commendation to have gone to an employee from a different division, who

[25] It is worth reiterating that the presence of nonmaterialistic cultures is not in itself evidence of their freedom from envy. In order to demonstrate that, it would have to be shown that social position and other forms of prestige were not broadly envied.

[26] The notion of positional goods is due to Fred Hirsch (1976), though it is a matter of dispute how best to understand it. As we are using the notion, mere scarcity is insufficient for positionality, inasmuch as scarcity is contingent.

is not in the running for this promotion. Situations of this sort are quite common, we think, because in most social hierarchies, only those who are in someone's vicinity (at least figuratively) directly affect his position. These are his rivals—not necessarily in the vernacular sense that implies he must see himself as being in competition with them, but rivals in that their successes come at a cost to his position, and vice versa. They are the people whose accomplishments he can fittingly envy—at least, if the positional goods they are jockeying for are granted to matter. We will now argue that positional goods are ubiquitous, and that they do matter.

While the paradigms of envy are not always obviously positional, they often serve in less direct ways as means of comparison. Certain conspicuous but fungible goods, such as fancy cars and expensive jewelry, are plausibly seen as eliciting envy primarily because of their role in determining one's standing in competition for something that is not fungible, namely status. It's not for nothing that such goods are called status symbols. These paradigmatic examples reveal some commonplaces about envy, but they also tend to caricature it as focused exclusively on winning competitions, accumulating possessions, and monitoring one's status. This focus makes positional goods out to be more shallow and less ubiquitous than we think they are. Consider competition, to start. Not only those who strive to be on top compete. Comparative success matters to the rest of us as well, especially because people tend to move in smaller social groups within which comparison reasserts itself. Many who do not aspire to greatness care about not being at (or near) the bottom of their cohort. Hence, people tend to drop down a league to avoid that fate, or to disengage entirely from some realm of competition and seek other sources of self-esteem, which often—we do not say inevitably—become avenues of interpersonal comparison.

Indeed, concern about one's position can have nothing to do with status, competition, or luxury items. Imagine a father who invests his energy playing with and mentoring the neighborhood kids, and comes to be proud of being their coach and teacher. This role may become one of the most meaningful aspects of his life, to the extent that he would feel displaced were another parent to usurp it. But not everyone can be Number One Dad, no matter how many Father's Day mugs suggest otherwise. Were he to contemplate losing his position, we would not be surprised to find that he took this prospect to involve a real loss, not only because he would no longer have those activities around which to structure his life. We suspect that this will be especially likely insofar as he prides himself on being the local mentor, no matter how alien it seems to him to think of himself as being in competition with the other parents on the block for that distinction. And, as we consider his life from outside, we take his position to count toward making his life a good one—in addition to the goods internal to the activities that engage him.

The prevalence of concern for positional goods is easy to underestimate, because a person need not be envious, or even consciously focused on relative position, in order for positional goods to play an important role in his self-esteem and welfare. Nor does our claim that people are often motivated by comparative position exclude other kinds of value and motivation. One of the great benefits of academia is that it allows for the pursuit of goods, particularly knowledge, that lend themselves to internal appreciation. Yet the academy is hardly free from competition. Certainly some scholars are more keenly attuned to their professional standing than others. Even those who eschew competition, however, are likely to find it gratifying to be invited to conferences, to place their papers well, be given teaching awards, and the like. While they may sincerely disavow any ambition to pre-eminence, it remains important to them to be recognized as an authority in their field. These forms of recognition are all positional, though, and we suspect that their ability to serve as grounds for self-esteem hinges more than one might like to suppose on the ways in which they signal something about the recipient's standing in the profession.[27]

It might be objected, at this point, that what moves many of us is not a competition for relative standing but simply the desire to achieve: to produce interesting, valuable work for its own sake.[28] We readily acknowledge other sources of motivation to achievement. People are moved by genuine interests specific to various fields of endeavor: a desire to understand and assess some complex and interesting theory, to play a difficult concerto with passion as well as precision, or to develop a new vaccine. Moreover, such desires can be inspired by the example set by another person, without this showing that what we *really* care about is outdoing a rival. We do suspect that this attitude is more commonly taken toward avocational interests, and that it is more easily held when one's position is relatively secure.

[27] By invoking the importance of comparative position to feelings of self-esteem, we do not mean to suggest that such feelings fully determine people's welfare, merely that they contribute to it importantly. One thing that makes a life go well is that it feels satisfying to the one living it. Some readers will be tempted to suggest that what is really of value here is not occupying a certain position but rather the pleasure one takes from this; we doubt this claim but need not dispute it here. This dialectical move is almost always available to the hedonist, and it is not our aim to refute hedonism in the theory of value. We would be content for the hedonist to concede that the goods with which envy is concerned are inherent goods: things in which we take satisfaction independently of their contribution to further goals.

[28] We are indebted to Sarah Buss and Peter Railton for pressing objections along these lines.

The fact that people are often motivated by specific goals, however, should not be permitted to obscure the respects in which desire for achievement also manifests their stake in positional goods. Part of the desire to achieve is surely an aspiration to excellence, and excellence in various endeavors—from scholarship to the arts, industry, even athletics—contributes to human flourishing. Yet which accomplishments count as excellent, or sufficiently good to be worthy of pride, is largely a function of the performance of others (especially those who are nearby).[29] This is especially obvious with respect to athletic excellence. What counts as an excellent sprinter depends on how fast people are sprinting during the period in which one competes. But reflection shows that such comparisons play an important role not merely in identifying but in determining excellence in many domains. Hence, whether one has excelled, in some of the ways that contribute to one's flourishing, is partly determined by the degree to which one's achievements stand out—albeit among various comparison classes.

We doubt that the aim of achieving non-comparative goals captures all of what most people are moved by in pursuit of excellence. Our skepticism is due partly to the observation that the standard of what a person will count as a valuable achievement tends to shift when too many people can reach it. This suggests that the content of these standards is partly dependent upon what it takes for an accomplishment to stand out, at least within a local comparison class. It would be a mistake to underestimate the degree to which an achievement's contribution to one's self-esteem hangs on such comparisons, even if one does not attend to them, and it would be a further mistake to underestimate the importance of self-esteem as a human motivation.

These mutually supportive considerations suggest that envy's concern plays a wide psychological role. The propensity to envy is a price of caring about relative standing: it is the painful counterpart to the aspirations people harbor and the pride they take in accomplishment. If we are right, then one's position in various social hierarchies matters to people quite generally, not merely as a potential source of envy. Furthermore, people do not merely desire but value positional goods upon reflection, for instance by thinking that excellence contributes to their flourishing. We also find it reasonable to defend the claim that excellence is valuable by citing respects in which it produces further goods, and thereby gets integrated into lives that prove rewarding in several different respects. Once it is granted that positional goods matter for human flourishing, then it follows that envy is

[29] It is perhaps worth noting, on this point, that the root of the word 'excellence' is 'excel': 'to do better than; surpass' (*The American Heritage Dictionary of the English Language*, 4th edn. (Buston: Houghton Mifflin, 2000).

sometimes fitting, because the success of rivals can be bad for an agent in the way envy suggests: it marks a comparative loss. Moreover, if the concern for positional goods is nearly as widespread as we suggest, then we are entitled to make the stronger claim that the enviable is a human value—or, rather, disvalue. That is to say, almost all human beings sometimes have reasons of fit to be envious.

In order for the zealot to make his case against envy, then, he must deny the importance of positional goods across the board. He must deny that comparative excellence is worth aiming at, and that it contributes in itself to a good human life. By pointing out the range of commonly accepted values with which this view conflicts, we hope to have shown that this is a difficult position to adopt. But we acknowledge some familiar grounds for being attracted to it. Surely it is possible for a person to minimize, and for some people to forego altogether, the pursuit of competitive success. Different cultures seem to vary in their balance of competitive and cooperative endeavors—though status distinctions of some sort are a pan-cultural phenomenon. Furthermore, we grant that there is something attractive in stoic ideals of serenity, and in the homelier injunction to get out of the rat race. Such lives are actually lived by some, and perhaps lives focusing on cooperation (not just cooperation to compete) and achievement (measured in absolute rather than relative terms) should be held up as the ideal to which we should aspire.

Nothing we have said conflicts with these claims, however. Our position is a form of pluralism, which recognizes various human goods and grants that no one sort of life will be able to realize them all. Hence, our limited defense of envy is consistent with the admission that there are genuine values realized in lives forsaking competition. We insist merely that such lives do give up something of human value: the goods distinctive of the pursuit of excellence. Similarly, a life focused largely around the pursuit of competitive accomplishment will forsake other values, but this acknowledgment does not undermine the value of what it does achieve.

Recall that envy is only one example of our general thesis that all sentiments are sometimes fitting, and perhaps the most difficult. Of all those values we have considered, the enviable seems to conflict most systematically with the demands of virtue. Even when envy is fittingly responsive to some genuine human disvalue, perhaps morality forbids acting from envy or even feeling it. Many actions motivated by envy would surely befit guilt if actually performed, and could only be done by someone with shameful traits. Indeed, perhaps the best sort of person never feels envy—though we have our doubts about that. Nevertheless, it would be a mistake to conclude that envy is never fitting. The unmatched achievements of our rivals can make us worse off, just as envy implies.

On the pluralist view we have been sketching here, the human values arising from sentiments generate reasons to feel that not do not merely compete with each other, but often conflict with the demands of morality. Nevertheless, if we are right, they sometimes provide reasons to feel for almost all of us.

REFERENCES

Bittner, Rüdiger, 'Is it Reasonable to Regret Things one Did?', *Journal of Philosophy*, 89 (1992), 262–73.

Cicero, *Cicero on the Emotions: Tusculan Disputations 3 and 4*, tr. M. Graver (Chicago: University of Chicago Press, 2002).

D'Arms, Justin, 'Envy', *The Stanford Encyclopedia of Philosophy* (Winter 2002 edn.), ed. Edward N. Zalta, <http://plato.stanford.edu/archives/win2002/entries/envy/>.

—— 'Two Arguments for Sentimentalism', *Philosophical Issues*, 15 (2005), 1–21.

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

—— and —— 'Sentiment and Value', *Ethics*, 110 (2000*b*), 722–48.

—— and —— 'The Significance of Recalcitrant Emotions (or, Anti-Quasijudgmentalism)', in Hatzimoysis (2003).

—— and —— 'Sensibility Theory and Projectivism', in D. Copp (ed.), *The Oxford Handbook of Ethical Theory* (Oxford: Oxford University Press, 2006).

Delancey, Craig, *Passionate Engines: What Emotions Reveal About Mind and Artificial Intelligence* (New York: Oxford University Press, 2002).

de Sousa, Ronald, *The Rationality of Emotion* (Cambridge, Mass.: MIT Press, 1987).

Ekman, Paul, 'All Emotions are Basic', in Ekman and Davidson (1994).

—— and Davidson, Richard (eds.), *The Nature of Emotion: Fundamental Questions* (New York: Oxford University Press, 1994).

Gibbard, Allan, *Wise Choices, Apt Feelings: A Theory of Normative Judgment* (Cambridge, Mass.: Harvard University Press, 1990).

Greenspan, Patricia, *Emotions and Reasons: An Inquiry into Emotional Justification* (London: Routledge, 1988).

Hatzimoysis, Anthony (ed.), *Philosophy and the Emotions* (Cambridge: Cambridge University Press, 2003).

Hirsch, Fred, *The Social Limits to Growth* (London: Routledge & Kegan Paul, 1976).

Johnson-Laird, P. N., and Oatley, K., 'Basic Emotions, Rationality, and Folk Theory', *Cognition and Emotion*, 6 (1992), 201–23.

Lazarus, Richard, 'Appraisals: The Long and the Short of it', in Ekman and Davidson (1994).

McDowell, John, *Mind, Value, and Reality* (Cambridge, Mass.: Harvard University Press, 1998).

Neu, Jerome, 'Jealous Thoughts', in A. Rorty (ed.), *Explaining Emotions* (Berkeley, Calif.: University of California Press, 1980).

Nussbaum, Martha, *The Therapy of Desire: Theory and Practice in Hellenistic Ethics* (Princeton: Princeton University Press, 1994).

Rabinowicz, Wlodek, and Rønnow-Rasmussen, Toni, 'The Strike of the Demon: On Fitting Pro-Attitudes and Value', *Ethics*, 114 (2004), 391–423.

Railton, Peter, 'Aesthetic Value, Moral Value, and the Ambitions of Naturalism', in Peter Railton, *Facts, Values and Norms* (Cambridge: Cambridge University Press, 2003).

Roberts, Robert, *Emotions: An Essay in Aid of Moral Psychology* (Cambridge: Cambridge University Press, 2003).

Taylor, Gabrielle, *Pride, Shame, and Guilt: Emotions of Self Assessment* (Oxford: Clarendon Press, 1985).

Tooby, J., and Cosmides, L., 'The Past Explains the Present: Emotional Adaptations and the Structure of Ancestral Environment', *Ethology and Sociobiology*, 11 (1990), 275–424.

5
The Meaning of ‘Ought’

Ralph Wedgwood

What does the word ‘ought’ mean? Strictly speaking, this is an *empirical* question, about the meaning of a word in English. Such empirical semantic questions should ideally be answered on the basis of extensive empirical evidence about the use of the word by native speakers of English.

As a philosopher, I am primarily interested, not in empirical questions about the meanings of words, but in the nature of the *concepts* that those words can be used to express—especially when those concepts are central to certain branches of philosophy, as the concepts expressed by ‘ought’ are central to ethics and to the theory of rational choice and rational belief. Still, it is often easiest to approach the task of giving an account of the nature of certain concepts by studying the meanings of the words that can express those concepts. This is why I shall try here to outline an account of the meaning of ‘ought’.

I shall try to argue that this account of ‘ought’ can deal adequately with some of the empirical linguistic data; but I shall not be able to undertake a sufficiently thorough investigation to be in a position to claim that my account deals adequately with *all* the linguistic data that need to be accounted for, nor that it deals better with the data than any alternative account. In particular, although I shall argue that the word ‘ought’ can express a large number of systematically related concepts (so that whenever the word ‘ought’ is used, the linguistic context must determine which of these concepts this occurrence of ‘ought’ expresses), I shall not be in a position to argue that my account of ‘ought’ captures *all* the concepts that the word can express. Still, I hope to give some reasons for thinking that

I am indebted to audiences at the University of Wisconsin, Madison, and at the University of Glasgow, and to my Oxford colleagues John Broome, Krister Bykvist, and Timothy Williamson, for penetrating comments on earlier drafts. The final revisions to this paper were carried out while I held a Research Leave award from the UK Arts and Humanities Research Board, to whom I should also like to express my gratitude.

my account captures at least *some* of the concepts that the word can express, and that these concepts are among those that are central to ethics and to the theories of rational choice and rational belief. In this way, I hope that my account should be able to play a useful clarificatory role within those branches of philosophy.

I should emphasize that I am concerned here purely with 'ought' (and its near synonym 'should'), not with all normative or deontic concepts as such. Many philosophical discussions of the meaning of 'ought' seem to assume that it is an obvious analytic truth that whenever one "ought" to do something, one has a "duty" or "obligation" to do it. This assumption seems eminently questionable to me. I ought to buy a new pair of shoes, but I surely do not have any duty or obligation to buy a new pair of shoes. Duties and obligations are in some sense "owed" to someone or something that is the object or beneficiary of the duty or obligation, while it is far from clear that anything like that need be true of everything that one "ought" to do. So for at least these reasons, 'ought', 'is obliged', and 'has a duty' must be distinguished. But I shall say nothing further about 'duty' and 'obligation' here. I shall focus exclusively on the term 'ought' instead.

1. Understanding and Logic

A good account of the meaning of a term should do two things: first, it should explain what it is to *understand* the term, or to count as a competent user of the term; secondly, it should explain the term's *logical properties*—which sorts of inferences involving the term are valid, and why. In the case of the term 'ought', explaining the logical properties of the term involves explaining the basic principles of *deontic logic*.

Different philosophers of language have taken radically different approaches to both of these tasks. In addressing the first task, most philosophers assume that it is at least part of understanding a term that one has the ability to use declarative sentences involving that term to express certain mental states. However, philosophers differ over what sort of mental state is normally expressed by the use of declarative sentences involving 'ought': cognitivists think that these mental states are just straightforward beliefs, of basically the same kind as the beliefs that are normally expressed by most other declarative sentences; non-cognitivists think that they are mental states of some crucially different kind, such as emotions, or desires or intentions of the sort that are typically expressed by commands or prescriptions.

Philosophers have also taken various different approaches to the second task, including what I shall call the "factualist" approach and the "non-factualist" approach.[1] According to the factualist approach, the fundamental explanation of the logical properties of the term essentially involves the idea that the content of any declarative sentence involving the term is a *proposition* that is either *true* or *false*. According to the non-factualist approach, even if one eventually "earns the right" to speak of propositions that are true or false, the *fundamental* explanation of the term's logical properties need say nothing about sentences involving these terms having as their contents propositions that are either true or false.

In this paper, I shall just assume that the cognitivist, factualist approach is correct. That is, I shall assume that the mental states that are normally expressed by the use of declarative sentences involving 'ought' are perfectly straightforward *beliefs*; and I shall explain the logical properties of 'ought' in terms of its contribution to the *truth conditions* of sentences in which it appears; in more technical terms, I shall explain the logical behaviour of 'ought' in terms of the word's *semantic value*.

More specifically, I shall assume that the semantic value of 'ought' is some *property* or *relation*, which features in the propositions that are the contents of sentences involving 'ought'. So I shall assume an ontology of propositions, properties, and relations—where propositions, properties, and relations are *universals*, which may have a complex structure, being composed, by means of operations analogous to predication, conjunction, negation, and so on, out of objects such as individuals, properties and relations.[2] (In effect, propositions are 0-place universals, monadic properties are 1-place universals, and the other relations are *n*-place universals for some $n > 1$.) A further feature of my conception of propositions can be articulated by reference to *possible worlds*: every proposition divides the possible worlds into those worlds where the proposition is true and those where it is false. A *fact* can be identified with a proposition that is true at the *actual* world. (There are ontological controversies about how these universals and possible worlds are related to each other: on some accounts, the universals can be constructed out of these possible worlds, while on other accounts, these possible worlds are in effect just big propositions. I shall avoid committing myself to any position on these controversial ontological questions here.)

[1] I avoid the term 'expressivism' here because an "expressivist" semantics for a class of statements—that is, a semantics that gives its fundamental explanation of the meaning of these statements in terms of the type of mental state that they express—is only one possible form that a non-factualist semantics of these statements could take.

[2] In effect, I shall be assuming something like the ontological framework outlined in the first four chapters of Bealer (1982).

Many philosophers have objected to this cognitivist, factualist approach, and especially to the application of this approach to broadly normative terms like 'ought'. Unfortunately, I shall not be able to answer most of these objections here; nor shall I be able to explain why I believe the cognitivist, factualist approach to be superior to its non-cognitivist and non-factualist rivals. I shall simply assume cognitivism and factualism for the sake of argument, in order to investigate what sorts of semantics are possible for the term 'ought' on the assumption that cognitivism and factualism are correct.

Nonetheless, I shall at least implicitly address one objection to the factualist approach. It might seem that if the word 'ought' has a property or relation as its semantic value—or in less precise terms, as its reference—it will be hard, if not impossible, to explain *why* the word 'ought' has the precise semantic value that it has. In this paper, I shall try to show that this is not so: we can give an illuminating, non-trivial explanation of why the word 'ought' has the precise semantic value that it has.

Specifically, I shall attempt to show that the semantic value of the 'ought' can be explained on the basis of the word's *essential conceptual role*. This "conceptual role" is a certain way of using the term in reasoning. It is "essential" in the sense that it is an essential part of understanding the term, or of being a fully competent user of the term, that one has some ability to use the term in this way. In this way, our account of what it is to understand the term can be integrated with our account of the term's logical properties: to understand the term, one must have some mastery of its essential conceptual role, and it is this conceptual role that explains the term's semantic value, which in turn explains the term's logical properties.[3]

[3] For my first attempt at this sort of conceptual role semantics, see Wedgwood (2001). Another philosopher who has developed a form of "conceptual role semantics" for normative vocabulary—according to which, as for my account, the essential conceptual role of normative vocabulary is its role in *practical reasoning*—is Robert Brandom (1994: 229–71; 2000: 79–94). Nonetheless, there are profound differences between Brandom's approach and mine. (1) I do not aim to give a reductive account of semantic or intentional notions in general: I simply *presuppose* that we are dealing with a term that expresses some concept or other; I aim only to explain what it is about the term that makes it the case that it has *this* particular meaning, and *this* particular property as its semantic value. (2) I do not take conceptual role semantics to be a *rival* to truth-conditional semantics: on the contrary, I assume a rich ontological framework of propositions, properties, and relations, and I take it to be an essential feature of the meaning of a term that it has some semantic value that is included within this ontology. Indeed, in my view, the "external" norm of correctness—which in the case of belief is a norm of *truth*—is more fundamental than any "internal" norm of rational inference or reasoning (after all, what is the *point* of rational inference or reasoning, if not to arrive at the truth in one's beliefs?). (3) I reject Brandom's radical holism; in my view it is not the *total* conceptual role of 'ought' that fixes its meaning, but only a special privileged part of its conceptual role.

2. The Logical Form of 'Ought'

One controversial question emerges immediately, concerning the *logical form* of 'ought'. Many philosophers understand 'ought' as a *propositional operator*—that is, as a term whose semantic value is a function from an embedded proposition (which is indicated in the sentence in which 'ought' occurs) to a further proposition. But other philosophers—most notably Peter Geach (1991)—hold that it is a mistake to assume that 'ought' is always a propositional operator; according to these philosophers, at least sometimes, 'ought' must be understood as a *relational predicate* applying to triples consisting of an agent, a possible course of action, and a time.[4]

In this paper, I shall treat 'ought' as a propositional operator wherever it occurs. There are at least some sentences where it certainly seems overwhelmingly plausible that 'ought' functions as a propositional operator. For example, consider:

(1) Drinking water ought to be clean and safe.

No particular agent is explicitly mentioned in this sentence: so how can this occurrence of 'ought' stand for a relation between an agent, a possible course of action and a time?

It might be suggested that in a particular context of utterance, (1) will contain an implicit reference to an agent, a time, and a possible course of action—namely, the course of action of *bringing it about that drinking water is clean and safe*. But it would be extraordinary if (1) could contain an implicit reference to a particular agent, in a given context of utterance, unless the speaker actually had that agent in mind in making that utterance; and a speaker in uttering (1) need not have any particular agent x in mind such that by uttering (1) she means to say that x ought to bring it about that drinking water is clean and safe. In that case, it might be suggested that the speaker means to express the proposition that there is at least *some* agent who ought to bring it about that drinking water is clean and safe. But this proposition has a radically different logical form: it is an existentially quantified proposition, not an atomic proposition. It is surely preferable if the logical form of the proposition that our semantics assigns to an utterance of a sentence bears some systematic relationship to the compositional structure of the sentence. But our semantics will preclude the possibility of any such systematic relationship if (1) sometimes expresses an atomic proposition (when the speaker has a particular agent in mind)

[4] Compare also Harman (1973).

and sometimes an existentially quantified proposition (when the speaker has no particular agent in mind).

We can avoid all these problems if we treat ‘ought’ in (1) as a propositional operator. Grammatically, ‘ought’ in English is an auxiliary verb, like the modal auxiliaries ‘can’ and ‘must’. When an occurrence of ‘ought’ modifies the main verb of a sentence, it can be taken as a propositional operator applying to the proposition that would be expressed by the unmodified form of that sentence. Thus, in (1), ‘ought’ is a propositional operator applying to the proposition that would be expressed by the sentence ‘Drinking water is clean and safe’.

If we treat ‘ought’ as sometimes functioning as a propositional operator, we would clearly achieve a more unified account if we suppose that it always functions as such an operator. We would also be able to unify our account of the auxiliary verb ‘ought’ with that of the modal auxiliaries ‘can’ and ‘must’, which practically all philosophers and semanticists would interpret as propositional operators.[5]

Moreover, there is a further argument, due to Bernard Williams (1981: 119–20), for the conclusion that ‘ought’ always functions as a propositional operator. The kind of ‘ought’ that philosophers like Geach regarded as standing for a relation between an agent and a possible course of action is what Williams called the “practical or deliberative *ought*”. The way in which this kind of ‘ought’ differs from other kinds can be illustrated by this example:

(2) Fred ought to have enough food for his family for Christmas.

We can distinguish at least two different readings of this sentence. The first reading would be appropriate if the reason for uttering this sentence is that Fred has promised to do the Christmas food shopping for his family, but is an unreliable person who is all too likely to forget to go to the shops before they close. The second reading would be appropriate if the reason for uttering the sentence is that Fred is too desperately poor to buy enough food for his family for Christmas, and the speaker is commenting on what a deplorable state of affairs this is. These two different readings could differ in truth value: on the first reading, the sentence is false unless Fred has a reasonably reliable *ability* to ensure that he has enough food for his family for Christmas, while on the second reading, the sentence could be true even if Fred has no such ability.

[5] In many languages, the closest equivalent to ‘ought’ is an impersonal verb followed by a noun clause, which is a construction that it is particularly tempting to interpret as representing a proposition embedded inside a propositional operator: *il faut* in French, *dei* and *chrē* in ancient Greek, *prepei* in modern Greek, *rhaid* in Welsh, *opportet* in Latin, and so on.

The first sort of 'ought' is often used to express either advice or a conclusion of deliberation or practical reasoning about what to do. This is why Williams called it "the practical or deliberative *ought*". This label might be misleading if it suggests that this sort of 'ought' can *only* be used to express conclusions of deliberation (in first-person contexts), or advice (in second-person contexts). There is no reason to think that this sort of 'ought' cannot occur in third-person or past-tensed contexts (as in 'Napoleon ought not to have invaded Russia') where there is no question of the speaker's giving advice or deliberating about what to do; and we should not assume that 'ought'-statements that are more naturally described as *theoretical* rather than practical (such as 'You ought to proportion your belief to the evidence') must involve a different kind of 'ought'. The point is just that this sort of 'ought' is particularly appropriate for expressing advice or deliberation.

The second sort of 'ought' is what Sidgwick called "the political *ought*".[6] This label is also potentially misleading, since many occurrences of this sort of 'ought' have nothing to do with politics (it might be better to call it "the *ought* of general desirability"); but I shall stick with Sidgwick's term here.

It is the first sort of 'ought'—the practical or deliberative 'ought'—that Geach construed as standing for a relation between an agent and a possible course of action, rather than as a propositional operator. But suppose that a group of people are involved in a joint deliberation, as a result of which a speaker concludes:

(3) Someone ought to go and inform the manager.

Even if one keeps constant the interpretation of 'ought' as having its practical or deliberative sense here, this sentence is clearly ambiguous. The ambiguity is most naturally interpreted as involving a scope ambiguity: on one reading, (3) means 'It ought to be that: someone goes and informs the manager'; on the other reading, it means 'Someone is such that: *he* ought to go and inform the manager'. On the first reading, the only agent who could possibly be the "subject" of the 'ought' is presumably the group involved in the joint deliberation, viewed as a collective agent. But this collective agent is not explicitly mentioned in the sentence, and so, for similar reasons to those that applied in the case of (1), 'ought' in this first reading of (3) also seems to be a propositional operator; and as Williams says (1981: 116), "it

[6] See Sidgwick (1907: book 1, ch. 3, n. 10, p. 34). Sidgwick illustrates this "political *ought*" by means of the following example (p. 33): 'when I judge that the laws and constitution of my country "ought to be" other than they are, I do not of course imply that my own or any other individual's single volition can directly bring about the change'.

is hard to see what requires it, or even allows it, to turn into something else" in the second reading. So there seems to be a reason for treating even the practical or deliberative 'ought' as a propositional operator.

If that is right, then the crucial difference between the two readings of (2) is not a difference in logical form. Rather, it seems that they must involve different kinds of 'ought'-operator—namely, the "practical" and the "political" 'ought'-operators respectively. One of the main differences between the practical and the political 'ought' seems to be that the practical 'ought' is at least implicitly *indexed* to an agent and a time. For example, in the reading of (2) on which it involves the practical 'ought', the 'ought'-operator is indexed to Fred and to some period of time (presumably, some period of time before the food shops close for Christmas); for this reading of (2) to be true, the proposition to which this 'ought'-operator is attached ('Fred has enough food for his family for Christmas') must be capable of being realized by *Fred*'s exercising some of the abilities that he has *at that time*.[7] The political 'ought', on the other hand, is not indexed to any particular agent and time in this way; this is why the reading of (2) on which it involves the political 'ought' can be true even if Fred lacks the ability to realize this proposition at that (or indeed any other) time.

As we have seen, the main difference between the two readings of (3) is not in the kind of 'ought' involved (both readings involve the practical or deliberative 'ought'), but in the relative scope of the quantifier and the 'ought'-operator. However, once we recognize that the practical 'ought' is always indexed to some agent, we see that in these two readings of (3), 'ought' must be indexed to different agents: on the first reading, it is implicitly indexed to "us" (the group engaged in the joint deliberation), whereas on the second reading, it is indexed to the agent-variable bound by the quantifier 'Someone ...'.

For most of this paper, I shall focus on the practical or deliberative 'ought'. (In the last section, I shall explore how my account can be generalized to deal with other kinds of 'ought' as well.) I shall represent the practical 'ought'-operator that is indexed to the agent *A* and time *t* by the symbol '$O_{<A,t>}$'.[8]

[7] Some philosophers believe that we must distinguish between "the time of the act" and "the time of the 'ought' ". I think this is wrong. In my view, there is no "time of the 'ought' "; at most, the fact that makes the 'ought'-statement true may be a fact about some particular time, such as the fact that one made a certain promise at a certain time. However, the proposition embedded inside the time-indexed 'ought'-operator may itself concern a different time from that to which the operator is indexed. For example, an adviser might say to you, 'Your nephew ought to inherit your property after you die'; in this case, 'ought' is indexed to you and the time at which you have the ability to draw up your will, not to the time after you die when your nephew will inherit.

[8] To avoid certain complications, let us suppose that this symbol has no content unless '*A*' refers to someone who is an agent at the time referred to by '*t*'. I should note

In the spirit of classical logic and unrestricted compositionality, I shall suppose that if there is a propositional operator '$O_{<A,t>}$', then this operator can be attached to *any* proposition *p*, to yield a further proposition '$O_{<A,t>}(p)$' that will have a definite truth value, either true or false. But we should note that it will in many cases be hard to find a sentence of standard English (or any other natural language that I know) that has the complex proposition '$O_{<A,t>}(p)$' as its content.

In English, one common way to convey that an occurrence of 'ought' has its "practical or deliberative" sense, and is indexed to a particular agent *A*, is to make *A* the grammatical subject of 'ought'. (Making an agent the grammatical subject of 'ought' does not always indicate that this occurrence of 'ought' is indexed to that agent: one mafioso might advise another 'Alfredo ought to be killed before he talks to anyone'; if this is the practical 'ought', it is indexed not to Alfredo—the grammatical subject of the verb 'ought'—but rather to the advisee.) But in English, the proposition to which the 'ought'-operator is attached is indicated by an *infinitive*—where the grammatical subject of the infinitive must be the same as the subject of the auxiliary verb 'ought'. So there is simply no way in grammatical English to affix the phrase 'You ought . . .' to an expression that indicates a proposition that does not somehow involve the person referred to as 'you'. For this reason, when the practical 'ought'-operator '$O_{<A,t>}$' is conveyed in English by the phrase 'At *t*, *A* ought . . .', there is a *grammatical* barrier to attaching this 'ought'-operator to any propositions that do not in some way involve *A*. Nonetheless, according to my assumptions, there is no *logical* barrier to attaching the operator '$O_{<A,t>}$' to propositions that have nothing to do with *A*. (If *p* is a proposition that does not in any way involve *A*, then we *cannot* convey '$O_{<A,t>}(p)$' by saying 'At *t*, *A* ought to bring it about that *p*'; the proposition *p* and the proposition '*A* brings it about that *p*' are obviously distinct propositions, which must not be confused with each other.[9])

Another way of conveying the operator '$O_{<A,t>}$' (more common in other languages than in English) is to use an impersonal construction like 'It ought to be the case that . . .', and leave it *implicit* in the context that this occurrence of 'ought' is indexed to a particular agent *A* and time *t*. Even if one uses a personal construction, so that the relevant agent is the

that I am being rather free and easy with the use of quotation marks, which sometimes form expressions that refer to linguistic types, sometimes to propositions, and often function as Quinean corner-quotes. I hope that no serious confusions will result.

[9] Even 'At *t*, *A* ought to be such that *p*' does not really convey '$O_{<A,t>}(p)$', but rather '$O_{<A,t>}$(*A* is such that *p*)'. '*A* is such that *p*' is not strictly speaking the same proposition as *p* itself: the former entails that *A* exists, while the latter may not.

grammatical subject of the auxiliary verb 'ought', it is still merely implicit in the context that this occurrence of 'ought' has its practical or deliberative sense (as opposed to its "political" sense, or some other sense). Because the practical 'ought' is especially connected with deliberation and advice, the easiest way to indicate that it is the practical 'ought' that is in play is if the context somehow makes it clear that the statement is made from the standpoint of the relevant agent's deliberations about what to do at the relevant time (or of someone advising the agent about what to do at that time). It will be very hard to convey that a statement is made from this standpoint if the proposition embedded inside in the 'ought'-operator is causally independent of everything that the agent might do or think at that time; as Aristotle famously observed,[10] no one deliberates about things that they cannot affect in any way. So, if nothing that the agent could do or think at that time will make any difference to whether or not p is the case, then it will be almost irresistible to hear the sentence 'It ought to be the case that p' as involving a different sort of 'ought'. For example, if someone says, 'You ought to have been born ten years earlier than you were', or 'You ought to have been born at exactly the time that you were born', it will be almost impossible to hear this as involving the practical 'ought' (as opposed to some other kind of 'ought'). Still, I am assuming that, in principle, *any* proposition p can be embedded inside the practical 'ought'-operator indexed to an agent A and time t, '$O_{<A,t>}$', to yield another more complex proposition '$O_{<A,t>}(p)$'.

We might try enriching natural language by introducing an explicitly indexed 'ought'-operator: 'It ought, from the standpoint of A and t, to be the case that . . .'. But we have no clear intuitions about sentences like 'It ought, from the standpoint of me and now, to be the case that there are nine planets in the solar system', even though, as noted above, I shall assume here that this proposition has a truth value. In the absence of any clear intuitions about these propositions, the question of what their truth conditions are must be decided by theoretical considerations, rather than by any direct appeal to intuition.

To sum up: I shall treat 'ought' as a propositional operator whenever it occurs. The "practical or deliberative 'ought' " (unlike what Sidgwick called the "political 'ought' ") is implicitly indexed to a particular agent and time. It will be hard to hear 'ought' as having this practical or deliberative sense, and as indexed to a particular agent A and time t, if the proposition that is embedded within the 'ought'-operator is causally independent of all of A's thoughts and actions at t. But this does not make it impossible

[10] See Aristotle, *Nicomachean Ethics* 3.3, 1112^{a}18–30.

for such propositions to be embedded inside this operator. Indeed, I shall suppose that the proposition '$O_{<A,t>}(p)$' has a definite truth value whatever the embedded proposition p may be. It might be hard to express this proposition using 'ought' in ordinary English; but this proposition will be true or false nonetheless.

3. Conceptual Role Semantics for the Practical 'Ought'

According to my version of conceptual role semantics, the semantic value of the practical or deliberative sense of the term 'ought' is determined by the role that the term essentially plays, when it has this sense, in *practical reasoning* or *deliberation*. Specifically, when it is used in this sense, the term's essential conceptual role is given by the following rule:

> Acceptance of the first-person statement '$O_{<me,t>}(p)$'—where 't' refers to some time in the present or near future—commits one to making p part of one's plan about what to do at t.

As I noted earlier, I am assuming a *cognitivist* interpretation of 'ought' sentences here; so I shall assume that to "accept" the sentence is just to *believe* the proposition that the sentence expresses. To say that a belief "commits" one to making a certain proposition part of one's plan is to say that, if one holds this belief, and the belief is itself rational, then that would make it *irrational* for one not to make that proposition part of one's plan.

A "plan about what to do at t", as I am understanding it, is just a proposition—roughly, a proposition that represents a way in which one might behave at t, and a way things might be if one did behave in that way. To "adopt" the proposition p as one's plan about what to do at t is to have a set of intentions about what to do at t such that, if the conjunction of the contents of those intentions is the proposition q, one believes the proposition 'If it were the case that q, it would be the case that p'. Then we can define "making the proposition p a part of one's plan" simply as: adopting as one's plan a proposition that logically entails p.

We could also introduce a similar operator 'P'—the practical or deliberative 'may', which some philosophers indicate by the term 'permissible'—whose essential conceptual role is given by the following rule:

> Acceptance of the first-person statement '$P_{<me,t>}(p)$'—where 't' refers to some time in the present or near future—permits one to treat p as allowed by one's plan about what to do at t.

To treat a proposition p as "allowed" by one's plan is, in effect, to be disposed not to adopt as one's plan any proposition that is *inconsistent* with *p*. To say that a belief "permits" one to treat a certain proposition as allowed by one's plan is to say that if one holds this belief, and the belief is rational, then that would make it *not irrational* for one to treat that proposition as allowed by one's plan.

If this rule gives the essential conceptual role of the practical or deliberative 'ought', then understanding this sense of 'ought' will involve having some mastery of this rule; and to have some mastery of this rule, one must presumably have at least some disposition to follow the rule. To follow this rule, one must respond to any rational belief in a proposition that can be expressed by a sentence of the form '$O_{<me,t>}(p)$' by making the embedded proposition *p* part of one's plan about what to do at *t*. Thus, anyone who understands the practical or deliberative 'ought' must have some disposition to respond to their own rational beliefs about what they ought to do by planning accordingly. In this way, the claim that the essential conceptual role of the practical 'ought' is given by this rule can explain why a certain form of "normative judgment internalism" is true: rational beliefs involving this sort of 'ought' must have some disposition to be accompanied by a corresponding plan about what to do (at least so long as the agent to whom this occurrence of 'ought' is indexed is the thinker herself, represented in the first person, and the time to which it is indexed is represented as in the present or near future).[11]

In following this rule, it is crucial that one should exhibit some sensitivity to whether or not one belief in this proposition is rational. This is a fundamental difference between rules about how one mental state *commits* one to having another mental state, and rules about how one mental state counts as a *ground* or *basis* for having another mental state. In some cases, simply *having* a mental state is enough to make that mental state a ground or basis for a further mental state, regardless of whether or not that first mental state is rational; in these cases, the first mental state does not in my sense "commit" one to that further mental state. This point helps to explain the particular way in which, according to my account, the essential conceptual role of this term '$O_{<A,t>}$' can explain the term's semantic value.

Within the "factualist" semantic framework that I am assuming here, the semantic value of the operator '$O_{<A,t>}$' will be a certain property of propositions—presumably, a relational property that propositions have in virtue of some relation in which they stand to the agent *A* and the time *t*. But how can the essential conceptual role of this operator, as given by the rule specified above, determine the operator's semantic value?

[11] For an argument for the claim that this is the best way for a realist about the normative to explain such "normative judgment internalism", see Wedgwood (2004).

The rule specified above can determine this operator's semantic value because this semantic value must be the *weakest* property of propositions that guarantees that all instances of that rule are *valid*—as I shall put it, it is that semantic value that "makes" the instances of the rule valid. But what does it mean to say that an instance of this rule is valid?

An instance of a rule can be regarded as having "inputs" and an "output", where these inputs and outputs are types of mental state. Where the rule is a rule about how one type of mental state *commits* one to another mental state, it would not be plausible to say that for an instance of such a rule to be valid, whenever one is *in* the input state, the output state must be a correct or appropriate state to be in. (That might be plausible for a rule that is merely about how one mental state counts as a ground or basis for another.) What is required rather is, roughly, that the *correctness* of the inputs guarantees the correctness of the output.[12] In the case of certain rules of *inference*, the inputs and output can be regarded as beliefs; and a belief is correct if and only if the proposition believed is true. So an instance of such a rule of inference is valid if and only if the truth of the contents of its inputs guarantees the truth of the content of its output. In this way, the notion of the "validity" of an instance of a rule is closely related to the notion of the logical validity of an inference. But other mental states besides beliefs can also be called correct or incorrect. So the notion of the validity of instances of a rule has wider application, besides its application to rules of inference.

More precisely, if the content of the rule is that the input mental states *commit* one to having the output mental state as well, then the semantic value of the operator in question must make it the case that the correctness of the input mental states guarantees that the output mental state is *uniquely* correct—that is, that it is the *only* correct mental state of that kind to have towards the proposition in question. If the rule is a rule about how the input mental states *permit* one to have the output mental state as well, then although the correctness of the input mental states must guarantee the correctness of the output state, it need not guarantee that that output state is uniquely correct. (This distinction between correct mental states and uniquely correct mental states is particularly important with respect to plans and intentions about what to do: if one is in a "Buridan's ass" situation, then it is correct to form an intention to go to the left, and also correct to form an intention to go to the right, but neither intention is *uniquely* correct.)

[12] This is my response to the principal objection that was made against my approach by Schroeter and Schroeter (2003): they overlook the fact that I have a way to distinguish such rules of "commitment" from other rules of reasoning.

On this approach, then, the semantic value of the practical 'ought'-operator '$O_{<A,t>}$' will be that property of a proposition *p* that makes it the case that the *only* correct way for *A* to relate the proposition *p* to her plan about what to do at *t* is to make *p* part of that plan. As I have explained, to make *p* part of one's plan is to adopt as one's plan a proposition that logically entails *p*. The obvious alternative way for *A* to relate *p* to her plan is to adopt as her plan a proposition that logically entails the *negation* of *p*. If the *only* correct way for *A* to relate *p* to her plan about what to do at *t* is to make *p* part of that plan, then it must be correct for *A* to adopt as her plan a proposition that entails *p*, and not correct for *A* to adopt as her plan a proposition that entails the negation of *p*. Thus, my account leads to the following account of the semantic value of '$O_{<A,t>}$': for any proposition *p*, '$O_{<A,t>}(p)$' is true just in case there are correct plans (for *A* to have about what to do at *t*) that logically entail *p*, and no such correct plans that logically entail the negation of *p*.

In this way, this approach to the semantics of 'ought' rests on the idea that there is a notion of "correctness" that can be applied to plans. It is admittedly not very common in ordinary English to describe plans as "correct" or "incorrect". But we do often speak of someone's making the "right choice" or the "wrong decision", or describe someone's decision as a "mistake". In these contexts, the terms 'right', 'wrong' and 'mistake' seem to be being used in the same sense as when we talk of a *belief*'s being right or wrong or a mistake; and choices and decisions are mental events in which we adopt or revise our plans about what to do. So we can say that a plan is correct if and only if it is a plan that it is in this sense right (not wrong or a mistake) to adopt. If there is indeed a genuine notion of "correct plans", then there should be no more objection to using this notion in the metalanguage in which we are giving our semantic theory than there is to using the notion of a "correct belief" or a "true proposition" in our metalanguage.

It seems plausible that this notion of a "correct plan" is itself a broadly normative notion. Indeed, we might try to explain what it is for an attitude to be "correct" along the lines suggested by Wiggins's (1989: 147) idea that "truth is the primary dimension of assessment for beliefs", together with Dummett's (1993: 42–52) idea that the root of our concept of truth is our grasp of what it is for a belief or an assertion to be correct. Following this suggestion, we might say that for a mental state to be "correct" is just for it to satisfy the "primary" norm (or "dimension of assessment") that applies to mental states of that type. Unfortunately, however, I cannot undertake to give a full account of the relevant notion of "correctness" here.[13]

[13] For more on this sense of 'correctness', see Wedgwood (2002).

If the notion of a "correct plan" is indeed a normative notion, then my account of the semantic value of 'ought' does not give any identification of this semantic value in non-normative terms; on the contrary, its identification of this semantic value uses the broadly normative notion of a "correct plan". In this sense, my account of the meaning of this sort of 'ought' is not a "naturalistic" account. (My account is, at least prima facie, *compatible* with the claim that the property that these uses of 'ought' refer to is in fact a natural property—that is, a property that can be picked out in wholly non-normative terms. But my account does not *imply* that this property is a natural property. If it is a natural property, that is not something that one could simply read off the semantics for the practical 'ought' that I have given here.)

It is because my account is not "naturalistic" in this strong sense that it can escape the dilemma that Terry Horgan and Mark Timmons (2000) have deployed against all forms of "naturalistic moral realism". According to Horgan and Timmons, every naturalistic account of the reference of a moral term will be vitiated by one or the other of the following two fatal flaws. The *first* flaw is that the account will simply fail to assign any determinate reference to the moral term at all. If the account is to avoid this first flaw, and to assign a determinate reference to the moral term, it will have to pick on a certain relation R in which we stand to a unique property, and claim that it is in virtue of our standing in relation R to that property that our moral term refers to the property. But now, according to Horgan and Timmons, the account will fall into the *second* flaw, since they claim for every such relation R, it is possible for there to be a community of speakers that do *not* stand in that relation to that particular property—even though intuitively it seems that the members of that other community also use terms that express moral concepts and have the very same reference as our moral terms.

This argument is plausible only if it is assumed that this relation R is a purely natural relation, and not itself a *normative* relation. But in my account, the relation in virtue of which these uses of 'ought' refer to the relevant property is itself a normative relation. In my account, this relation is the following: first, these uses of 'ought' express a concept whose essential conceptual role consists in the way in which certain beliefs involving this concept *commit* one to incorporating a certain proposition into one's plans; and secondly, this concept refers to the property that makes this sort of practical reasoning *valid*—that is, the property of a proposition p that makes it *correct* for one to incorporate the proposition p into one's plans about what to do at t, and *incorrect* to incorporate the negation of p into one's plans about what to do at t.

It seems plausible to me that a community that had *no* term that ever expressed a concept whose essential conceptual role was this role in practical reasoning and planning would *not* have any terms for the practical or deliberative 'ought'. However, so long as certain uses of a term in their language do express such a concept, then according to my account, those uses of that term *must* have the same reference as the corresponding uses of our term 'ought'. In the case of belief, it seems to be the very same property of a proposition *p*—namely, *truth*—that makes it correct for members of one community to believe the proposition *p* as makes it correct for members of any other community to believe *p*. But the same point, it seems to me, holds for the case of plans as well. It is the very same relation between a proposition *p*, an agent *A*, and a time *t* that makes it uniquely correct for members of one community to incorporate the proposition *p* into their plans as makes it uniquely correct for members of any other community to do so. So long as a community uses a term to express a concept that has this essential conceptual role in practical reasoning and planning, my account will demand that if one of those uses of the term is indexed to an agent *A* and time *t*, then it refers to the property of standing in that relation to *A* and *t*. In this way, then, my account escapes both horns of Horgan and Timmons's dilemma.

Since my account of the meaning of 'ought' itself makes use of normative terms, some philosophers may complain that my account of the meaning of 'ought' is viciously circular. But this complaint is mistaken. No one demands that an account of what it is for a word to mean *cow*, for example, must make no mention of any relation in which that word stands to actual cows. No one demands that an account of what it is for a word to mean *not* must refrain from using any words (like 'not') that have that very meaning. All that it is reasonable to demand is that the account should not presuppose the idea of *a word's having that meaning* (or *expressing that concept*). It should instead give an informative account of *what it is* for a word to have that meaning (or of what it is for the concept that is expressed by the word to be that concept). One way to dramatize this demand is by imagining the situation of a "radical interpreter".[14] An adequate account of the meaning of the practical or deliberative 'ought' would give an illuminating explanation of how, at least in principle, such a radical interpreter could identify a term in an unknown language as having this meaning. According to my account, to identify a term as having this meaning, an interpreter would have to acquire some reason to think that the term expresses a concept that has the essential conceptual role that I have sketched above. In principle, one could acquire reason to think this in just the same way as one could acquire

[14] The idea of such "radical interpretation" is due to Davidson (2001: essay 9). We could appeal to this idea without accepting Davidson's full-blown "interpretivism". As Lewis (1974*a*) suggests, the reference to interpretation could just be taken as a way of dramatizing what is objectively constitutive of a word's having the meaning in question.

reason to think that a term in an unknown language expresses a concept that has the essential conceptual role that is given by the introduction and elimination rules for one of the logical constants like 'or' and 'if'. For this reason, my account is not viciously circular.

It would also not be fair to complain that my account is trivial or uninformative. First, as we have already seen, my account can give an explanation of why a certain sort of "normative judgment internalism" is true. Secondly, in the next section, I shall give another example of how my account of the meaning of the practical 'ought' has substantive consequences. Specifically, I shall explain how, given plausible claims about the nature of planning and practical reasoning, my account of the semantic value of the practical 'ought' can explain which principles of *deontic logic* are correct for this sort of 'ought'. (I should warn my readers that the next section will be fairly technical; readers who are not interested in deontic logic are invited to skip this section.)

4. The Logic of the Practical 'Ought'

The general idea of how this account of the semantics of the practical 'ought' can provide an explanation for the principles of deontic logic is fairly straightforward. According to this account, the meaning of this kind of 'ought' is given by its essential conceptual role in practical reasoning; and the term's semantic value is that property of a proposition that makes it correct for the relevant agent to adopt plans that entail that proposition, and incorrect for her to adopt plans that entail the negation of that proposition. So, if there are consistency constraints on correct planning and practical reasoning, then there will be corresponding consistency constraints on statements involving the 'ought'-operator. These consistency constraints are in effect precisely what deontic logic consists in—namely, principles, flowing from the very meaning of the term 'ought' itself, about which sets of 'ought'-statements are consistent and which are not. So, on the approach that I am recommending, the source of deontic logic lies in these consistency constraints on planning and practical reasoning.

It certainly seems plausible that there are consistency constraints on planning. Many of these consistency constraints stem from the idea that to be correct our plans must be *realizable*. In some sense, it is part of what plans are *for* that they should guide us to act in such a way as to realize those plans. Thus, a plan that simply cannot be realized fails to achieve the result that plans exist to achieve. Hence, I shall suppose, no such plan can be "correct". (Strictly speaking, the realizability constraint on planning takes two forms. First, there is a realizability constraint that is relative to the agent's *beliefs*—that is, the agent should not adopt a plan if he *believes* that it cannot be realized; this constraint is what I shall call a "constraint on *rational* planning". Secondly, there is a realizability constraint that is relative to the *facts* of the agent's situation—that is, the agent should not adopt a plan that cannot *in fact* be realized; it is constraints of this second kind that I shall call "constraints on *correct* planning".)

In fact, however, my specification of the semantic value of the practical 'ought' already reflects some of these consistency constraints on correct plans. For any two propositions p and q, if p is logically equivalent to q, then there are correct plans that logically entail p and no correct plans that logically entail the negation of p if and only if there are correct plans that logically entail q and no correct plans that logically entail the negation of q. So if p and q are logically equivalent, then so too are '$O_{<A,t>}(p)$' and '$O_{<A,t>}(q)$'. In this sense, the operator '$O_{<A,t>}$' behaves like a *classical* modal operator: it permits the substitution of logical equivalents.[15]

Moreover, suppose that there are correct plans that logically entail 'p & q', and no correct plans that logically entail the negation of 'p & q' (so, given my account, '$O_{<A,t>}(p \;\&\; q)$' is true). Then there are correct plans that logically entail p and no correct plans that logically entail the negation of p (since any plan that entailed the negation of p would also entail the negation of 'p & q'); and similarly, there are correct plans that logically entail q and no correct plans that logically entail the negation of q. So, the operator '$O_{<A,t>}$' also behaves like a *monotonic* modal operator: that is, it *distributes over conjunction*; '$O_{<A,t>}(p \;\&\; q)$' entails '$O_{<A,t>}(p)$' and '$O_{<A,t>}(q)$'.[16]

To defend the other logical principles that apply to the practical 'ought'-operator, however, we need to appeal more explicitly to the idea that any correct plan for an agent A to have about what to do at a time t must be *fully realizable* by A at t. I propose that this idea should be understood in the following way.

First, let us define what it is for a proposition to be realizable by A at t. To say that a proposition p is "realizable" by A at t is to say that A has

[15] For a useful account of the various sorts of non-normal modal operators, see Schurz (1997: 160–1).

[16] The claim that 'ought' distributes over conjunction has been disputed; e.g. Jackson (1985) has proposed analysing '$O(p)$' in counterfactual terms, as meaning, roughly, 'If it were the case that p, things would be better than they would be if it were not the case that p'. This analysis allows for counterexamples to distributivity. Suppose that (i) the nearest possible world in which p is true is one in which q is not, and (ii) such worlds are very bad, although worlds in which both p and q are true are very good. Then given Jackson's analysis, '$O(p \;\&\; q)$' is true, but '$O(p)$' is false. But it seems to me that 'ought' is not well analysed in such counterfactual terms. We often say that something "ought" to be the case when it is very much only a *part* of everything that ought to be the case.

some set of abilities such that there are possible worlds in which all the actual truths that are causally independent of whatever A might do or think at t hold, and A exercises those abilities at t, and in all those worlds, p is true. (Thus, all the actual truths that are causally independent of whatever A might do or think at t will, in a degenerate sense, be realizable by A at t. Roughly, for a truth p to be "causally independent of whatever A might do or think at t" is for it *not* to be the case that there is some thought or course of action such that there are nearby possible worlds in which A has that thought or performs that action at t, and in all such worlds, p is not true.)

Secondly, it is a crucial feature of plans that we can adopt a *partial* plan, and then fill in the details of the plan (by adding further conjuncts to the proposition that we have adopted as our plan) as time goes by. Let us say that a *maximally detailed plan* for an agent A and a time t is one such that, for every proposition p that is realizable by A at t, the plan logically entails either p or its negation. Then we can articulate the constraint on correct plans as follows: a plan is correct only if it is possible to extend the plan into a maximally detailed correct plan that is itself a realizable proposition.

Now suppose that (i) there are correct plans (for A to have about what to do at t) that entail p, and no such correct plans that entail the negation of p, and in addition (ii) there are correct plans that entail q and no such correct plans that entail the negation of q. Since every correct plan is fully realizable, the propositions p and q must be realizable. So the correct plans that entail p must be capable of being extended into a maximally detailed plan that entails either q or the negation of q. But there are no correct plans that entail the negation of q. So the only correct maximally detailed extensions of these plans entail q. So there are correct plans that entail both p and q; hence there are correct plans that entail 'p & q'. But there cannot be any correct plans that entail the negation of 'p & q' (if there were such correct plans, there would have to be correct maximally detailed extensions of those plans that entailed either the negation of p or the negation of q; but by hypothesis there are no such correct plans). Hence, given my account of its meaning, the practical 'ought'-operator '$O_{<A,t>}$' also behaves like a *regular* modal operator: that is, it *agglomerates over conjunction*; '$O_{<A,t>}(p)$' and '$O_{<A,t>}(q)$' taken together entail '$O_{<A,t>}(p \,\&\, q)$'.

Moreover, if correct plans must be realizable, then any proposition that is logically entailed by a correct plan must also be realizable. Hence, given my account of its meaning, if '$O_{<A,t>}(p)$' is true, then p must itself be realizable. Clearly it is logically impossible for any logically false proposition to be realizable. Hence, the practical 'ought'-operator '$O_{<A,t>}$' also conforms to the so-called D principle of modal logic: if p is logically false, then '$O_{<A,t>}(p)$' is also logically false.

So far, I have argued in favour of all the principles of von Wright's original (1951) deontic logic. But in fact, the account that I have given so far also supports the final principle that is needed to turn von Wright's system into standard deontic logic. This principle is the *rule of necessitation*, according to which if p is a logical truth, then so is '$O_{<A,t>}(p)$'. Now, the logical principles that I have already defended are enough to show that if there is any truth of the form '$O_{<A,t>}(q)$', then for every logical truth p, p follows from q, whatever q may be, and so '$O_{<A,t>}(p)$' is true as well. But need there be any truth of the form '$O_{<A,t>}(q)$'? (Perhaps for some A and t, there are *no* correct plans for A to have about what to do at t?) If so, then this argument will not show that '$O_{<A,t>}(p)$' is a logical truth whenever the embedded proposition p is also a logical truth.

The most intuitive way to argue for the rule of necessitation is probably to focus, in the first instance, not on the 'ought'-operator, but on the 'may' operator '$P_{<A,t>}$'. The semantics that I suggested above for this operator naturally leads to the conclusion that the semantic value of this operator '$P_{<A,t>}$' is that property of a proposition p that makes it the case that it is correct (though not necessarily uniquely correct) for A to treat the proposition p as "allowed" by her plans about what to do at t. As I suggested earlier, to treat p as "allowed" by one's plans is to be disposed not to adopt any plans—even maximally detailed plans—that are inconsistent with p. So the natural conclusion to draw is that the semantic value of this operator '$P_{<A,t>}$' is that property of propositions that makes it the case that there is at least one maximally detailed correct plan (for A to have about what to do at t) that is consistent with p (which is not to say that there cannot also be *other* correct plans that are inconsistent with p).

Obviously, however, a logical falsehood is not consistent with anything; so in particular if p is a logical falsehood then p is not consistent with any correct plans (let alone maximally detailed correct plans) for A to have about what to do at t. So if p is a logical falsehood, then '$P_{<A,t>}(p)$' cannot be true. Since we relied on nothing but logic and the semantics of the operator '$P_{<A,t>}$' to establish that '$P_{<A,t>}(p)$' cannot be true, '$P_{<A,t>}(p)$' must be a logical falsehood too.

It is also plausible that the two operators, 'ought' and 'may', '$O_{<A,t>}$' and '$P_{<A,t>}$', are *duals* of each other.[17] 'It may permissibly be the case that ...' is definable as 'It is not the case that it ought not to be the case that ...',

[17] This claim might be thought to conflict with the claim of some philosophers that (i) deontic logic can be used to understand the logical structure of *legal* codes, and (ii) there are "gappy" legal codes, according to which certain courses of action are neither permitted nor forbidden. However, this is not a problem for my view, it seems to me, since the concept of what is "legally required" is not a kind of 'ought'; the logic of legal codes is not a sort of deontic logic.

and vice versa (that is, '$P_{<A,t>}$' is definable as '$\neg O_{<A,t>} \neg$', and '$O_{<A,t>}$' as '$\neg P_{<A,t>} \neg$').[18] But then if p is a logical truth, '$\neg p$' is a logical falsehood, and so '$P_{<A,t>}(\neg p)$' must also be a logical falsehood, and '$\neg P_{<A,t>}(\neg p)$' must be a logical truth. So if p is a logical truth, '$O_{<A,t>}(p)$' must also be a logical truth. That is, the rule of necessitation is sound.

A simpler but perhaps less intuitive argument for the rule of necessitation starts from the point that, as I am understanding the term, one's "plans" for what to do at t do not just consist of one's *intentions* about what to do at t. As I put it earlier, to "adopt" the proposition p as one's plan about what to do at t is to have a certain set of intentions such that, if the conjunction of the contents of those intentions is the proposition q, one *believes* the proposition 'If it were the case that q, it would be the case that p'. In this way, the proposition that one adopts as one's plan incorporates not just one's intentions but also one's *beliefs about the causally independent facts*. It is especially important for one's plan to incorporate one's beliefs about the causally independent facts that will determine what the causal consequences of one's actions will be. Of course, many of the other causally independent facts will be less practically relevant than these; and in a sense, it is quite redundant for one to incorporate these practically irrelevant facts into one's plan. But however practically irrelevant these facts may be, it is not *incorrect* to incorporate such facts into one's plan (indeed, if a correct plan is "maximally detailed" in the sense that I defined above, it would have to entail all such causally independent facts). Logical truths are always among the truths that are causally independent of what one does. So it will always be correct to incorporate such logical truths into one's plans (and of course it will never be correct to incorporate the negations of such logical truths into one's plans). Thus, the rule of necessitation is guaranteed to be sound: if p is a logical truth, so too is '$O_{<A,t>}(p)$'.

More generally, it is correct to incorporate any causally independent truths into one's plans. So if p is such a causally independent truth, then '$O_{<A,t>}(p)$' is true.[19] (Unless the causally independent truth p is itself a logical truth, then '$O_{<A,t>}(p)$' will be a truth but not a logical truth; this is

[18] If '$O_{<A,t>}$' and '$P_{<A,t>}$' are duals of each other, and for '$O_{<A,t>}(q)$' to be true, the embedded proposition q must be realizable, then the natural conclusion to draw is that for '$P_{<A,t>}(p)$' to be true, p must be, as we might put it, at least *practically possible*: A must have some set of abilities such that there is a possible world in which all the actual truths that are causally independent of everything that A thinks or does at t hold, A exercises those abilities at t, and p is true.

[19] This may also give us a reason for accepting the S4 principle for the deontic operator: '$O(p) \rightarrow OO(p)$'. There are several other principles that have been suggested as part of the logic of 'ought' that would also have to be considered in a fuller treatment of this topic—e.g. '$O(O(p) \rightarrow p)$', '$O(p \rightarrow OP(p))$', and '$P(p) \rightarrow OP(p)$'.

because unless p is a logical truth, then logic alone cannot tell us whether or not p is a causally independent truth.)

It must be conceded that, unlike the other principles of deontic logic that I have argued for, the rule of necessitation is not intuitively obvious. As I already mentioned (in § 2) in defending my view of the logical form of 'ought', it is hard to hear the term 'ought' as having its practical or deliberative sense and as indexed to an agent A and time t, unless the proposition embedded inside the operator is one whose truth value is causally dependent on A's thoughts or actions at t. So it is hard to hear the term 'ought' as having its practical or deliberative sense in sentences like 'It ought to be the case that the number 3 is not both prime and not prime', and it is all but impossible to hear this occurrence of 'ought' as indexed to a particular agent and time. As I emphasized earlier (at the end of § 2), we cannot rely on a direct appeal to intuition to evaluate sentences of this kind: we must appeal to theoretical considerations instead; and as I have argued, these theoretical considerations come down in favour of the rule of necessitation.[20]

If these logical principles involving the practical 'ought'-operator '$O_{<A,t>}$' are indeed correct, then there is a natural possible-worlds semantics for this operator. First, for any possible world w, there is a set of propositions that are true in w, and causally independent of all the agent A's thoughts or actions at t in w. Let us call the worlds at which all these propositions are true the worlds that are "available" to A at t in w. Then there is some selection function that picks out a subset of these "available" worlds; let us say that it picks out the "favoured" available worlds. It is a constraint on this selection function that the set of "favoured" available worlds must be a realizable proposition (in the sense defined earlier). Then we can say that for any proposition p, '$O_{<A,t>}(p)$' is true in w if and only if p is true at all these favoured available worlds. This possible-worlds semantics leads to standard deontic logic under the assumption that the set of favoured available worlds is never empty.

For a thorough list, see Åqvist (1984). Unfortunately, I will not be able to consider whether these principles are genuinely logical truths here.

[20] Various objections have been raised against the rule of necessitation in deontic logic. For example, it might seem that it makes it "too easy" to answer a radical "error-theorist" who believes that 'ought' is meaningful but all sentences in which 'ought' has largest scope are false. But the quest for a semantics for 'ought' that is neutral on absolutely all metaethical controversies seems misguided. Certainly, this radical sort of error theory is incompatible with the account that I have given of the meaning of 'ought'. But that only shows that a full defence of my account would have to involve an argument for regarding this radical error theory as false. It also need not follow that if my account is correct, then this radical error theorist is irrational, or that he doesn't understand the term 'ought'. It often happens that a philosopher understands a term perfectly well but embraces a false theory of what the term means.

In effect, this possible-worlds semantics corresponds fairly closely to the account that was proposed by Fred Feldman (1986).[21] The main difference is that, instead of speaking of the "favoured" available worlds, Feldman speaks of the "best" available worlds. But nothing that our discussion has covered so far justifies the claim that the "favoured" worlds are in any sense the "best" worlds.[22] Hence I have used a more non-committal term in characterizing the relevant selection function simply as a "favouring" function. (We should also note that the relevant selection function is itself indexed to the relevant agent *A* and time *t*; so this semantics is compatible with rejecting a consequentialist moral theory in favour of a more agent-relative, deontological theory. For example, it may be that a world in which *A* fails to prevent two murders at *t* is "favoured", while a world in which there are fewer murders overall but *A* himself commits a murder at *t* is not "favoured" in the relevant way.)

I have argued that the logic for the practical 'ought' is nothing other than standard deontic logic. Many objections have been raised against standard deontic logic over the years. First, there is the paradox of Ross (1941: 62): in standard deontic logic, '$O_{<A,t>}(p)$' entails '$O_{<A,t>}(p \vee q)$'; so 'You ought to post this letter' entails 'You ought to: either post this letter or burn it'. But it seems to me that if we bear in mind that this entailment holds only if 'or' has its truth-functional sense, then it is clear that the statement 'You ought to post this letter or burn it' is actually true. There is an obvious Gricean explanation for why it seems an odd thing to say: it is much less informative than something else that one might say—namely, 'You ought

[21] Compare also Belzer (1998), Humberstone (1983), and Loewer and Belzer (1983).

[22] In fact, I believe that there are further considerations that justify the claim that there is a *ranking* of worlds, such that the "favoured" worlds can be identified with the worlds that come highest in this ranking; and I also believe that the English words 'better' and 'best' express a sufficiently large number of notions that we can convey the idea that one world w_1 comes higher up in this ranking than another world w_2 by saying that w_1 is "better" than w_2. The considerations that justify the claim that there is such a ranking of worlds have to do with the logical relations between the *conditional* 'ought'-statements 'Given that p_1, it ought to be that q' and 'Given that p_1 & p_2, it ought to be that q'. So long as these statements involve the same type of 'ought', then it is plausible that the same logical relations hold between them as between the *counterfactuals* 'If it were the case that p_1, it would be the case that q' and 'If it were the case that p_1 & p_2, it would be the case that q'. Then as Lewis (1973: 58–9) has shown, these logical relations imply that any adequate possible-worlds semantics, either for the conditional 'ought' or for counterfactuals, will be equivalent to one that involves a ranking of worlds. (In the case of counterfactuals, the ranking of worlds is in terms of *closeness to what is actual*; in the case of the conditional 'ought', the ranking is in terms of *closeness to what is ideal.*) However, a long and intricate argument is needed to defend the claim that these conditional 'ought'-statements are logically related in this way; so I shall not try to defend the claim here.

to post this letter'. Asserting the weaker claim would tend to be a useful contribution to a conversation only if one was not in a position to assert the stronger claim—that is, only if it is not true either that you ought to post the letter, or that you ought to burn it, but only that you ought to do one or other of these things. Thus it is easy to explain why 'You ought to either post the letter or burn it' may seem false even if it is actually true.[23]

A second alleged paradox of deontic logic focuses on the more general point that, in standard deontic logic, if p entails q then '$O_{<A,t>}(p)$' entails '$O_{<A,t>}(q)$'. So for example in the Good Samaritan paradox of Prior (1958: 144), 'You ought to help the traveller who was beaten and robbed' entails 'There ought to be a traveller who was beaten and robbed'. However, once we remember that we are dealing with an 'ought'-operator that is indexed to an agent and a time, it becomes clear that the conclusion 'It ought to be that the traveller was beaten and robbed' only follows if the occurrence of 'ought' in the conclusion has the *same* sense, and is indexed to the *same* agent and time, as in the premise. Presumably the premise is only true when indexed to a time t such that the fact that the traveller has been beaten and robbed is causally quite independent of everything that the relevant agent thinks or does at t. But as I argued earlier, there is no natural way in English of expressing the proposition that results from attaching a practical 'ought'-operator that is indexed to a particular agent A and time t to an embedded proposition whose truth value is causally independent of all A's thoughts and actions at t. We simply have no intuitions about the sentence 'From the standpoint of you now (when there is absolutely nothing that you can do that will change the fact that the traveller was beaten and robbed), it ought to be the case that the traveller was beaten and robbed'. When this sentence strikes us as false, that is because we are not hearing it as involving a practical 'ought' that is genuinely indexed to that agent and that time. Instead, we may be hearing it as equivalent to 'From the standpoint of you and some time at which there *was* something that you could do that would determine whether or not the traveller was beaten and robbed, it ought to be that the traveller was beaten and robbed'

[23] Here is an objection to my response to Ross's paradox. My response entails that if there is anything that you ought to do, then whatever you do, you will do something that you ought to do. (If you burn the letter, you will have done something that you ought to do—namely, post the letter or burn it; similarly, if you throw the letter away, and so on.) But surely it cannot be that easy to do something that one ought to do? Reply: There are many problems with this objection (it plays very fast and loose with quantification over "things that one might do", e.g.). But even if my response to Ross's paradox does entail this result, the result is not obviously counterintuitive at all. On reflection, it seems clear that it *is* easy to do *something* that one ought to do: what is hard is to do *everything* that one ought to do . . .

(that is, roughly, 'You ought to have seen to it that the traveller was beaten and robbed'). But my account of the logic of the agent- and time-indexed practical 'ought' certainly does not imply that *this* follows from the original premise. Thus, when the conclusion of this inference strikes us as false, that is because we are sliding between the original practical 'ought', which was indexed to a particular agent and time, and *another* 'ought', which differs either in not being a practical or deliberative 'ought', or else in not being indexed to the same agent and time. For these reasons, then, it seems to me that these objections to standard deontic logic are not compelling.[24]

5. The Context-Sensitivity of 'Ought'

So far, I have only given an account of one kind of 'ought'—the practical or deliberative 'ought'. But there is extensive linguistic evidence that there are in fact several different kinds of 'ought': the term 'ought' expresses different concepts in different contexts of use.

I have already cited the distinction between the practical 'ought' and what Sidgwick called the "political 'ought' ". The most striking difference between these two kinds of 'ought', as I have suggested, seems to be this: the practical 'ought' is clearly indexed to a particular agent and time, and it is a constraint on what "ought" to be the case, in this sense, that it should be realizable by what the agent thinks or does at that time; the political 'ought', on the other hand, is not indexed to any particular agent and time in this way. I might say, 'The British constitution ought to be radically reformed', without having any particular agent *x* in mind (either individual or collective) such that I mean to say that *x* ought to bring it about that the British constitution is radically reformed. In that case, as I argued earlier,

[24] For this reason, I find it somewhat surprising that many recent deontic logicians (e.g. Hansson, 1997; Belzer, 1998) have been persuaded by these familiar "paradoxes". I suspect that part of the reason is that these deontic logicians seem not to have seen the evidence in favour of the hypothesis that 'ought' is systematically context-sensitive, and is implicitly indexed in different contexts of use to various different parameters; hence they have been rather uncritical in relying on their linguistic intuitions, without investigating whether these intuitions in fact involve different 'ought'-operators—that is, occurrences of 'ought' that are indexed to different parameters. Admittedly, many other "paradoxes" have been raised against standard deontic logic. But according to my account, most of these (including Castañeda's (1981) "Paradox of the Second Best Plan" and Åqvist's (1967) "Paradox of the Knower") can be solved in the same way as the Good Samaritan Paradox. The main exception is Chisholm's (1963) "Paradox of the Contrary-to-Duty Imperative". The most promising solution to this paradox is the familiar solution in terms of the *conditional* 'ought'; see Feldman (1990).

my statement does not contain any implicit reference to any particular agent. My acceptance of this statement hardly commits me to *planning* on the radical reform of the British constitution; at most it commits me to *favouring the goal* of such radical reform.

'Ought' exhibits other sorts of contextual variation as well. For example, on some occasions, 'ought' seems to be relative to a particular goal or purpose. Thus, someone might say, pointing to someone who is fiddling with a safe, 'He ought to use a Phillips screwdriver to open that safe' or even just 'He ought to use a Phillips screwdriver'.[25] Intuitively, this statement is true just in case using a Phillips screwdriver is necessary for opening the safe in the best or most effective way—even if, in many other salient senses of the term, the person ought not to be opening the safe at all. On other occasions, on the other hand, 'ought' is not relative to a particular goal or purpose in this way. Thus, in saying that the person in question ought not to be opening the safe at all, one is not simply saying that the person's refraining from opening the safe is necessary for achieving some particular goal or purpose in the best or most effective way.

Another crucial dimension of context-sensitivity is seen in the fact that, on some occasions, 'ought' seems to be relative to the information that is actually available to the relevant agent, whereas on other occasions it is not. Sometimes, it might be true for us to say, 'Given how little we know about what will happen, we ought to play safe'; here what we "ought" to do depends only on the information available to the agent. But on other occasions, 'ought' is not relative to the information available to the agent in this way: thus, it might sometimes be true to say 'It turned out that I ought not to have done that, although I couldn't have known it at the time'.

There are yet other examples of context-sensitivity in 'ought'. For example, there is the epistemic 'ought', as in 'Tonight's performance ought to be a lot of fun', which seems to mean, roughly, just that it is highly *probable* that tonight's performance will be a lot of fun.

I shall argue that this contextual variation in the concept that the term 'ought' expresses is not mere random ambiguity (like the way in which 'bank' in current English is ambiguous between *river bank* and *money bank*). Rather, the term 'ought' is systematically context-sensitive. There are certain specific contextual parameters that are fixed by the context of a statement involving the term 'ought'; and these contextual parameters determine which of these many 'ought'-concepts the term 'ought' expresses in the context.

[25] I take this example from Williams (2002). Compare Prichard's (1949: 91) discussion of the 'ought' that is "hypothetical" on the agent's intentions.

My account of the meaning of the practical or deliberative 'ought' was based on the idea that the essential conceptual role of this type of 'ought' is its role in *practical reasoning*. There is a natural way of generalizing this approach so that it can cover other kinds of 'ought' as well. In the widest sense, *deliberation* involves considering a certain domain of propositions, against the background of a certain body of information, and then making a certain assessment of those propositions.

In the kind of deliberation that I focused on in my account of the practical 'ought', the domain of propositions consists of those propositions that are compatible with all the truths that are causally independent of everything that one thinks or does at the relevant time; and these causally independent truths form the background against which one assesses the propositions that are compatible with them. The distinctive sort of assessment that forms the output of this kind of deliberation is *incorporating* some of these propositions into one's *plans* about what to do at the relevant time.

The essential conceptual role of other kinds of 'ought' is their role in other kinds of deliberation. For the political 'ought', the relevant domain of propositions is a wider domain (not just those propositions that are compatible with everything that is causally independent of what a particular agent thinks or does at a particular time): roughly, it is the domain of propositions that are compatible with those features of the actual world that could not easily be otherwise—the features that hold in all the possible worlds that are "nearby" the actual world, such as the laws of nature. The kind of assessment that forms the output of this kind of reasoning is not incorporating any of these propositions into one's actual plans, but only forming a *preference* for some of these propositions over the alternatives that are incompatible with them. To form a preference for a proposition over the relevant alternatives is in effect to form a *conditional plan*—in effect, the plan of acting in such a way that the proposition in question is true, rather than in such a way that the relevant alternative is true, if one does either.

The essential conceptual role of the purpose-relative 'ought' (as in 'He ought to use a Phillips screwdriver to open that safe') is its role in a kind of *purpose-relative practical reasoning*. In this kind of practical reasoning, one reasons merely about how to achieve a certain purpose, ignoring the question of whether or not to pursue that purpose in the first place. The relevant domain of propositions is, as with the practical 'ought', the propositions that are compatible with everything that is causally independent of what the relevant agent thinks or does at the relevant time. The difference lies with the kind of assessment that forms the output of this reasoning. This is not incorporating the relevant propositions simply into one's plans about what to do at the relevant time, but incorporating those propositions into one's *contingency plans* about how (if at all) to achieve the purpose in question. In effect, it is to plan on acting in such a way that the proposition in question is true in the event that one plans on achieving the purpose in question.

In general, the semantic value of each of these kinds of 'ought' will be the property of a proposition that makes it uniquely correct to assess the proposition in the relevant way, out of the domain of propositions that are compatible with the relevant background information. So, for example, according to this account, the statement 'He ought to use a Phillips screwdriver to open that safe' will be true just in case the proposition that the person in question uses a Phillips screwdriver follows from some correct contingency plans for how (if at all) to open the safe, and the negation of that proposition does not follow from any such correct contingency plan.

It seems to me that for every one of these kinds of 'ought' (practical, political, and purpose-relative), there is both a version that is relative to the information that is available to the relevant agent and a version that is not information-relative in this way. For example, sometimes we might say, 'Given that he didn't know what sort of safe it was, he ought to have tried opening it with an ordinary screwdriver first', whereas on other occasions we might say, 'He couldn't have known it at the time, but he ought to have used a Phillips screwdriver for that safe'; and similarly with the other kinds of 'ought'. I shall call these "information-relative" and "objective" uses of 'ought' respectively.

The difference between the "information-relative" and "objective" uses of 'ought' does not consist in the kind of deliberation in which they have their essential conceptual role, but rather in the precise role that these uses of 'ought' play in those kinds of deliberation. I shall illustrate the difference with respect to the practical 'ought'; a similar difference will apply to the other kinds of deliberation as well. A belief involving the objective practical 'ought', of the form '$O_{<me,t>}(p)$', *unconditionally* commits the believer to incorporating the proposition p into his plans; the only way in which the believer can escape this commitment is by giving up this belief. The information-relative 'ought', on the other hand, is relativized, at least implicitly, to a particular body of information; and the essential conceptual role of the information-relative practical 'ought' is that the canonical rational ground or basis for beliefs involving this sort of 'ought', of the form 'In relation to information I, $O_{<me,t>}(p)$', is the fact that being in information state I commits the believer to incorporating p into his plans about what to do at t.

The epistemic 'ought' (as in 'Tonight's performance ought to be a lot of fun') seems to be a sort of information-relative 'ought', implicitly relative to a certain body of information that counts in the context as *evidence*. The relevant sort of deliberation here is not practical reasoning, but deliberation about what to believe; this sort of deliberation starts out from the information that counts as evidence in the context, and concludes with the thinker's forming at least a tentative belief in one of the propositions that are compatible with that evidence. So the essential conceptual role of the epistemic 'ought' consists in the fact that the canonical rational ground or basis for beliefs involving this sort of 'ought', of the form 'In relation to evidence *E*, it ought to be that *p*', is the fact that evidence *E* commits one to forming at least a tentative belief in *p*.

If this account of the essential conceptual role of the information-relative 'ought' is correct, it may be plausible to say that its semantic value will just be that relation between a body of information and a proposition that makes it the case that that information really does commit the relevant agent to making the relevant sort of assessment of that proposition. Thus, the semantic value of the epistemic 'ought' will be that relation between a body of information and a proposition that makes it the case that that information commits one to forming at least a tentative *belief* in that proposition. In other words, given how I am understanding the notion of 'commitment', this is the relation that makes it the case that if having that information is itself a rational state, then it is irrational not to form at least a tentative belief in that proposition. Presumably, this relation has something to do with the proposition's *probability* on that evidence. Thus, an epistemic 'ought'-statement, of the form 'In relation to evidence *E*, it ought to be that *p*', is true if and only if *p* is sufficiently probable given evidence *E*.

In the previous section, I argued that the logic of the practical 'ought' reflects the consistency constraints that apply to correct planning. In a broadly similar way, the logic of each of these other kinds of 'ought' reflects the consistency constraints that apply to the relevant kind of assessment. (In the case of the information-relative 'ought', these will be consistency constraints on *rational* assessments of the relevant kind; in the case of the objective 'ought', they will be consistency constraints on *correct* assessments of the relevant kind.)

According to the suggestions that I have made here, the relevant kind of assessment (at least in the case of the kinds of 'ought' that I have considered here) involves incorporating the proposition in question into some sort of (conditional) plan—or, in the case of the epistemic 'ought', into one's system of beliefs. It seems plausible to me that essentially the same consistency constraints apply to conditional plans and to belief systems as to unconditional plans. First, for a conditional plan to be correct, the conditional plan must be logically consistent; and likewise, for a system of beliefs to be correct, the contents of the system must be logically consistent. Secondly, for a conditional plan or a system of beliefs to be correct, it must be possible to extend it into a maximally detailed plan or system of beliefs which is also itself correct. Finally, it will always be correct to incorporate a logical truth to any plan or system of beliefs, and never correct to incorporate the negation of a logical truth. Just like the practical 'ought', then, these other kinds of 'ought' are subject to all the consistency constraints of standard deontic logic. It may also be plausible that there are similar consistency constraints on *rational* conditional plans and on *rational* beliefs. If so, then it is plausible that the informative-relative 'ought' is also subject to the consistency constraints of deontic logic.

We can capture these logical features of these sorts of 'ought' by means of a generalized version of the possible-worlds semantics that I sketched in the previous section for the practical 'ought'. As we have seen, the context must determine two parameters for each occurrence of 'ought'. First, the context must determine the conceptual role of the concept that this occurrence of 'ought' expresses. Determining this will involve settling the following two issues: (i) whether it is the sort of conceptual role that is characteristic of the objective 'ought', or the sort that is characteristic of the information-relative 'ought'; and (ii) what kind of deliberation figures in this conceptual role—that is, what kind of assessment of propositions is the output of this sort of deliberation (for example, this output might be incorporating the proposition into one's plans, or into one's contingency plans, or into one's system of beliefs). Secondly, the context must determine "the relevant domain of propositions" and the relevant background information; this information can be represented by means of a set of propositions *S*, which is "held fixed" in the context, so that only those propositions that are consistent with *S* count as "the relevant domain of propositions" in the context. (For the objective 'ought', the set of propositions that is "held fixed" will typically be some set of truths that need not be known or believed by any of the participants to the relevant conversation, such as all the truths that are causally independent of what the relevant agent thinks or does at the relevant time. For the information-relative 'ought', the set of propositions that is "held fixed" will typically be some set of propositions that are known or believed by the participants in the conversation.)

So, in a context in which a set of propositions *S* is being "held fixed", and 'ought' expresses a concept with essential conceptual role *C*, 'It ought to be the case that *p*' is true at a world *w* if, and only if, *p* is true in all possible worlds that (i) are compatible with all members of *S* and (ii) belong to the "favoured" subset of those worlds (from the standpoint of *w*) according to the relevant selection function that is associated with *C*.[26] As before,

[26] As Kratzer (2001) put it, 'ought'-statements involve two contextually determined parameters: (i) the "modal base" (which delimits the relevant class of worlds), and (ii) the

this account of the semantics leads to standard deontic logic so long as the "favoured" subset of the worlds that are compatible with *S* is never empty.

My account of the objective practical 'ought' can be seen as an instance of this general pattern. According to my account, a statement involving the practical 'ought', of the form '$O_{<A,t>}(p)$', is true if and only if *p* is true in all worlds that (i) are compatible with all the propositions that are true in *w*, and causally independent of what *A* thinks or does at *t*, and (ii) belong to the "favoured" subset of those worlds, when assessed in the appropriate way with respect to how *A* acts at *t* in those worlds.

Similar accounts can be given of the other kinds of 'ought', including the epistemic 'ought' (as in 'Tonight's performance ought to be a lot of fun'). A statement involving the epistemic 'ought', of the form 'It ought to be the case that *p*', is true at a world *w* if, and only if, *p* is true in all worlds that (i) are compatible with what *counts as evidence in the context*, and (ii) belong to the "favoured" subset of those worlds when evaluated with respect to *probability on the evidence* from the standpoint of *w*.

Often, the set of propositions *S* that is "held fixed", and so determines which domain of propositions (or possible worlds) is relevant to an 'ought'-statement, is just determined implicitly by the context. In some cases, however, it may be indicated more explicitly. To take an example involving the practical 'ought', one may say: 'If you are going to keep on taking heroin intravenously, you at least ought to use clean needles'. Here the proposition 'You are going to keep on taking heroin intravenously' is explicitly added to the set of propositions *S* that is "held fixed", producing a set that is different from the set that would ordinarily be "held fixed" for this sort of 'ought'. In effect, this is a *conditional* 'ought' of the sort that was analysed by David Lewis (1974*b*) among others. Thus, this statement is true just in case all members of the "favoured subset" of the worlds in which the addressee keeps on taking heroin intravenously (and in which all the other truths that are causally independent of what the addressee does or thinks at the relevant time also continue to hold) are also worlds in which he uses clean needles. So far as I can see, there is a conditional 'ought' of this sort corresponding to every one of the various kinds of 'ought' that I have discussed above.

The general picture of 'ought' as a kind of modal operator, quantifying over certain possible worlds, is familiar. What my conceptual role semantics approach adds to this familiar picture is an understanding of what selection function yields the relevant "favoured" worlds (as I have called them). There are in fact many such selection functions, each corresponding to a

"ordering source" (which supplies a ranking of the worlds and thereby a "favoured" subset of the relevant worlds).

different concept that can be expressed by ‘ought’. What these selection functions have in common is this: for each of these selection functions, there is a certain way of “assessing” propositions that forms the output of some kind of “deliberation”, such that the propositions that are true at all the worlds that are “favoured” according to this selection function are all propositions that it is uniquely correct (in the case of an objective ‘ought’) or uniquely rational in relation to the relevant information (in the case of an information-relative ‘ought’) to assess in that way.

According to the account that I have outlined, the logical principles that apply to each ‘ought’-concept stem from the consistency constraints on the kind of deliberation within which that ‘ought’-concept has its essential conceptual role. The reason why the principles of standard deontic logic are correct for each of the ‘ought’-concepts that I have discussed so far is that each of these concepts has its essential conceptual role within a kind of deliberation the output of which consists in incorporating some proposition into some sort of system of plans or beliefs. It seems essential to any system of plans or beliefs that to be correct, or even to be rational, its contents must all be consistent with each other. Since plans and beliefs are subject to fairly robust consistency constraints, so too are these ‘ought’-concepts.

There may be yet other concepts, similar to the ‘ought’-concepts that I have discussed so far, that have their essential conceptual role in some other kind of deliberation to which these consistency constraints do not apply. For example, perhaps there is a kind of deliberation or reasoning the output of which is a *desire*; and perhaps there need be nothing incorrect about simultaneously desiring both *p* and the negation of *p*. There might then be a concept whose essential conceptual role is to play a regulative role in this sort of reasoning; and there would be no reason to expect standard deontic logic to hold for such concepts. It may even be that the word ‘ought’ in English can express such a concept. (Philosophers who insist on the possibility of “moral dilemmas”—situations in which it ought to be that *p* is the case, and also ought to be that *p* is not the case—would insist that ‘ought’ can express a concept of this kind.) Unfortunately, I cannot undertake the empirical investigations that would be necessary in order to determine whether the English ‘ought’ can express any such concept.

To conclude: it appears that my conceptual role semantics for the term ‘ought’—according to which the basic meaning-constituting conceptual role of ‘ought’ is its role in deliberation—can explain many of the phenomena that such an account is called upon to explain. It can explain the precise ways in which ‘ought’ is systematically context-sensitive; it can provide an explanation of why the principles of standard deontic logic are correct for each of the many concepts that can be expressed by 'ought'; and it enables us to answer the objections that have been raised against that standard deontic logic.

REFERENCES

Åqvist, Lennart (1967) 'Good Samaritans, Contrary-to-Duty Imperatives, and Epistemic Obligations', *Noûs*, 1: 361–79.

—— (1984) 'Deontic Logic', in Dov Gabbay (ed.), *Handbook of Philosophical Logic* (Dordrecht: Reidel), 605–714.

Bealer, George (1982) *Quality and Concept* (Oxford: Clarendon Press).

Belzer, Marvin (1998) 'Deontic Logic', in Edward Craig (ed.), *Routledge Encyclopedia of Philosophy* (London: Routledge): http://www.rep.routledge.com/article/Y043

Brandom, Robert (1994) *Making it Explicit* (Cambridge, Mass.: Harvard University Press).

—— (2000) *Articulating Reasons* (Cambridge, Mass.: Harvard University Press).

Castañeda, Hector-Neri (1981) 'The Paradoxes of Deontic Logic: The Simplest Solution to All of Them in One Fell Swoop', in Hilpinen (1981: 378–85).

Chisholm, Roderick (1963) 'Contrary-to-Duty Imperatives and Deontic Logic', *Analysis*, 24: 33–6.

Davidson, David (2001) *Inquiries into Truth and Interpretation*, rev. edn. (Oxford: Clarendon Press).

Dummett, Michael (1993) *The Seas of Language* (Oxford: Clarendon Press).

Feldman, Fred (1986) *Doing the Best we Can* (Dordrecht: Reidel).

—— (1990) 'A Simpler Solution to the Paradoxes of Deontic Logic', *Philosophical Perspectives*, 4: 309–41.

Geach, P. T. (1991) 'Whatever Happened to Deontic Logic?', in Geach (ed.), *Logic and Ethics* (Dordrecht: Kluwer).

Hansson, Sven Ove (1997) 'Situationist Deontic Logic', *Journal of Philosophical Logic*, 26 (4, Aug.): 423–48.

Harman, Gilbert (1973) Review of Roger Wertheimer, *The Significance of Sense, Philosophical Review*, 82: 235–9.

Hilpinen, Risto, ed. (1981) *New Studies in Deontic Logic: Norms, Actions, and the Foundations of Ethics* (Dordrecht: Reidel).

Horgan, Terry, and Timmons, Mark (2000) 'Copping out on Moral Twin Earth', *Synthese*, 124: 139–52.

Horty, John F. (2001) *Agency and Deontic Logic* (Oxford: Oxford University Press).

Humberstone, I. L. (1983) 'The Background of Circumstances', *Pacific Philosophical Quarterly*, 64: 19–34.

Jackson, Frank (1985) 'On the Semantics and Logic of Obligation', *Mind*, 94: 177–95.

Kratzer, Angelika (2002) 'The Notional Category of Modality', in Paul Portner and Barbara Partee (eds.), *Formal Semantics: The Essential Readings* (Oxford: Blackwell), 289–323.

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

—— (1974*a*) 'Radical Interpretation', *Synthese*, 23: 331–44.

—— (1974*b*) 'Semantic Analyses for Dyadic Deontic Logic', in Sören Stenlund (ed.), *Logical Theory and Semantic Analysis* (Dordrecht: Reidel), 1–14.

Loewer, Barry, and Belzer, Marvin (1983) 'Dyadic Deontic Detachment', *Synthese*, 54: 295–319.

Prichard, H. A. (1949) *Moral Obligation* (Oxford: Clarendon Press).

Prior, Arthur (1958) 'Escapism', in A. I. Melden (ed.), *Essays in Moral Philosophy* (Seattle: University of Washington Press), 135–46.

Ross, Alf (1941) 'Imperatives and Logic', *Theoria*, 7: 53–71.

Schroeter, Laura, and Schroeter, François (2003) 'A Slim Semantics for Thin Moral Terms?', *Australasian Journal of Philosophy*, 81: 191–207.

Schurz, Gerhard (1997) *The Is–Ought Problem: An Investigation in Philosophical Logic* (Dordrecht: Kluwer).

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

Wedgwood, Ralph (2001) 'Conceptual Role Semantics for Moral Terms', *Philosophical Review*, 110: 1–30.

—— (2002) 'The Aim of Belief', *Philosophical Perspectives*, 16: 267–97.

—— (2004) 'The Metaethicists' Mistake', *Philosophical Perspectives*, 18: 405–26.

Wiggins, David (1989) *Needs, Values, Truth*, 2nd edn. (Oxford: Basil Blackwell).

Williams, Bernard (1981) '*Ought* and Moral Obligation', in Williams, *Moral Luck* (Cambridge: Cambridge University Press).

—— (2002) ' "Ought", "Must", and the Needs of Morality' (unpublished).

Wright, G. H. von (1951) 'Deontic Logic', *Mind*, 60: 1–15.

6

Knowing Enough to Disagree: A New Response to the Moral Twin Earth Argument

Mark van Roojen

At the beginning of the twentieth century, G. E. Moore's open question argument convinced many philosophers that moral statements were not equivalent to statements made using non-moral or descriptive terms. For any non-moral description of an action or object it seemed that competent speakers could without confusion doubt that the action or object was appropriately characterized using moral terms such as 'good' or 'right'. The question of whether the action or object so described was good or right was always open, even to competent speakers. In the absence of any systematic theory to explain the possibility of synthetic as opposed to analytic identities, many were convinced this demonstrated that moral properties could not be identified with any natural (or supernatural) properties. Thus Moore and others concluded that moral properties such as goodness were irreducible sui generis properties, not identical to natural

I owe thanks to a number of people for discussion of this paper and related topics as I was working on the first draft. They include David Chalmers, John Gibbons, Terry Horgan, Harry Ide, Mark Kalderon, Clayton Littlejohn, Jennifer McKitrick, Joe Mendola, Sarah Sawyer and Russ Shafer-Landau. I also owe the audience at the First Annual Madison Metaethics Workshop in Oct. 2004 for comments and questions. People I specifically recall talking about the paper on that occasion include Sarah Buss, David Copp, Janice Dowell, Jamie Dreier, Terry Horgan, Mike Ridge, Russ Shafer-Landau, Michael Smith, David Sobel, Sarah Stroud, Mark Timmons, Pekka Vayrynen, Ralph Wedgwood, and Eric Weiland. If there are others whose comments I don't recall it says more about my memory than their comments. I apologize to them. I owe further thanks to two anonymous referees for Oxford University Press, Jennifer Haley, Russ Shafer-Landau (again!), and Pekka Vayrynen for helpful comments on still later drafts. Finally Mark Timmons kindly suggested I come up with a less modest title than the original.

properties (Moore, 1903: 15). Noncognitivists used the same argument to support the idea that moral judgments have an expressive function rather than a representational function. Their explanation for the failure of competent speakers to recognize the equivalence of moral predicates with other predicates was that these terms, unlike other predicates, did not serve to represent properties at all (Ogden and Richards, 1923: 125).

Contemporary philosophers recognize the possibility of synthetically (as opposed to analytically) identifying objects or properties referred to using different terms. We can discover that water is the same stuff as H_2O without being able to infer it from the meanings of the terms involved (Kripke, 1972; Putnam, 1975). Descriptive naturalists with respect to ethics capitalized on this to point out that the openness of Moore's question to competent speakers does not rule out the possibility of discovering that a moral property is a naturalistic property through empirical evidence not dependent on the expressions in question having the same meaning. The most sophisticated version of this sort of proposal has been offered by Richard Boyd. Oversimplifying just a bit, Boyd's idea is that moral terms can refer to a property in virtue of a certain sort of causal connection between the use of the term and the property, just as the term 'water' can refer to H_2O in virtue of a causal connection between H_2O and our use of the term water. Since it need not be transparent to a speaker what object or property bears the right sort of relation to her use of a term, a competent speaker can remain ignorant of the identity in question. Hence linguistic competence will not be sufficient to close open questions about the identities of the properties involved (Boyd, 1988).[1]

Terry Horgan and Mark Timmons (henceforth abbreviated as H&T) have constructed a neat argument intended to refute Boyd's theory and all similar theories. If they are correct, their argument together with Moore's original open question argument leave us to choose between noncognitivism on the one hand and non-naturalism on the other, the same options available to our predecessors seventy years ago. Given these options, and given that their argument highlights the commendatory function of moral language, the authors suggest that a sophisticated noncognitivism is the preferred choice.

Horgan and Timmons make explicit what they take to be commitments of causal theories of reference of the sort Kripke, Putnam, and Boyd use to explain the functioning of scientific kind terms, and which Boyd also applies to moral terms. These theories are anti-Fregean in the sense that term reference was not determined by a descriptive sense grasped by a thinker or

[1] Some of my wording on this page duplicates wording in my (2004) encyclopedia entry.

speaker and uniquely satisfied by the referent of the term. Rather, reference is determined by the existence of a certain sort of causal connection between the speaker's use of the term and the referent. However, Horgan and Timmons argue, the thought experiments used by proponents to motivate such causal theories generally show that competent speakers are in fact aware that the terms refer to whatever stands in the appropriate causal relations to the use of a term when the causal theory is the appropriate theory of reference for that term. If this is right, then applying the same theory to moral terms would suggest that a competent speaker should at least tacitly know that the relevant moral term refers in virtue of the right sort of causal connection. Thus speakers' intuitions about whether or not the term refers to whatever has such a connection should be probative with respect to the truth of the theory of reference in question. Horgan and Timmons then construct a clever example involving "Moral Twin Earth" to generate intuitions in conflict with the assumption that causal regulation determines reference for moral terms, and conclude that the theory is false. Indeed, they claim their argument sounds the death knell for all descriptivist versions of naturalism.

The argument has spawned a number of replies, each designed to show that the example does not refute naturalism. Many of these replies argue that the target theory survives in the face of the Moral Twin Earth example (Geirsson, 2003). My argument takes a somewhat different line. I think the target semantic theory as understood by Horgan and Timmons is in fact refuted by the Moral Twin Earth example. And I think the internalist upshot of their argument—that a commendatory function is a constitutive feature of genuine moral discourse—is also correct. However, I will argue, we can construct a successor semantic theory to the one proposed by Boyd which takes advantage of his real insights while supplementing them in various ways. This theory is not refuted by the Twin Earth example and in fact incorporates the internalist[2] upshot of the example while classing moral property terms as genuinely referring expressions.

[2] Unfortunately, philosophers have used the words 'internalism' and 'externalism' to mark a number of different philosophical distinctions in different subfields of philosophy and at least two of these are relevant to this paper. Here I mean internalist in the sense which requires a necessary connection between accepting a moral judgment and being motivated to do what it recommends. At other points I will be defending views which are 'externalist' in a sense that does not contrast with this one, but which instead contrasts with claims that the meaning of a term and the contents of thoughts using that term are entirely determined by individualistic properties of the speaker or thinker using the term. Internalism in that contrasting sense is the view that meaning and content are determined by facts internal to the thinker's head skin. I hope that context will make clear which sort of internalism or externalism I have in mind.

The Dialectic, the Target Theory, and the New Objection

The old open question argument

The Moral Twin Earth argument is embedded in a dialectic that begins with G. E. Moore's open question argument. Moore's argument purports to show that goodness could not be identical with any naturalistic property. He challenged his readers to provide candidate natural properties to identify with goodness. He claimed that for any such candidate property it was open to a person capable of having thoughts involving the property to wonder whether it was in fact identical to the property goodness.[3] The fact this was possible could then be used as a premise in two sorts of arguments purporting to demonstrate the property goodness was not in fact the candidate property. One version rests on the assumption that an analysis ought to support substitution of one term for the other in any meaningful sentence. If goodness is to be analyzed as the property in question, it would then be appropriate to substitute the term for the property in question for any occurrence of the term 'goodness', including sentences expressing one's uncertainty about whether the property was itself goodness. But then one would be asking whether the property in question was itself—using the very same term to pick out the property. For example, "I wonder whether pleasantness is goodness," would become, "I wonder whether pleasantness is pleasantness." Since according to Moore the former makes sense and the latter does not, pleasantness is not a correct analysis of goodness. And thus, Moore believed, for any natural candidate. If we now add the assumption that identity claims are analytic—that is underwritten by correct analyses of one term employing another term for the thing—we arrive at the conclusion that goodness is not identical to any natural property.

A second version of the argument relies on Leibniz's law. It takes the fact that we might sensibly ask ourselves whether pleasantness is good but not sensibly ask ourselves whether goodness is good to show that we are in doubt about the goodness of pleasantness but not the goodness of goodness. Thus pleasantness has a property goodness lacks, and by Leibniz's law the two cannot be identical.[4]

[3] Moore actually used the term 'good' to refer to the property in question, but I think it is more natural to use the term 'goodness'. Moore's text is somewhat schematic and does not provide all of the premises needed to construct a valid argument, so I'm providing a reconstruction which I think is faithful to his intentions. More important for my purposes, I think my reconstruction captures what people took from his text and were persuaded by. See Moore (1903: 15–17).

[4] For textual support for this interpretation consider, "whoever will attentively consider with himself what is actually before his mind when he asks the question 'Is

The argument was widely influential in convincing people that no natural property was identical with either goodness or rightness. But in hindsight it is sometimes hard to see why the argument had such influence. Similar worries could be raised about almost any other informative identity claim inasmuch as it might be reasonable for someone to consider it a subject for investigation. Discussions of the paradox of analysis and Frege's puzzle should already have made this clear. And the first version of the argument contains a number of assumptions which we have reason to doubt, perhaps most importantly that identity claims must be analytic as opposed to synthetic. It is possible that those impressed by the argument thought the sorts of identity claims that philosophy was after—claims that could tell us about the real nature of an object or property—would have to be a priori, even if not all identity claims are. The Kantian idea that all claims with necessary modal force must be a priori plus the necessity of identity might underwrite the assumption. Such assumptions may make the open question argument hard to resist—even with seeming counter-examples ready to hand—at least until subsequent work on the semantics of natural kind terms made clear how identities might be matters for empirical investigation. Even now the second version of the argument employing Leibniz's Law remains somewhat persuasive until one is in a position to say which of the premises is false and why.

The target theory

This is where views which Horgan and Timmons dub "New Wave Moral Realism," come in. The new wave theorists, most notably Richard Boyd, provide a semantics for moral terms that explains how identity claims though necessary could be synthetic. The new wave theories also explain how we could sensibly have the sorts of doubts Moore's open questions express, even with respect to identical properties. The theories explain why Moorean doubts do not require us to give up Leibniz's law in order to defend the identification of moral properties with natural properties. These new wave theories are modeled on anti-descriptivist and externalist theories of meaning and content determination suggested by the work of (among others) Kripke (1972), Putnam (1973, 1975), and Burge (1979).

These theories were in part constructed to explain how any identity claim could be open to rational doubts on the part of even competent speakers of

pleasure (or whatever it may be) after all good?' can easily satisfy himself that he is not merely wondering whether pleasure is pleasant." (Moore, 1903: 16). For a nice discussion see Kalderon (2004).

the language which is used to express the identities. The basic idea was to deny that the relevant terms functioned as disguised descriptions known by competent speakers, sufficient for determining the referents of the terms. Without this assumption there is no reason to think linguistic competence makes available for each term some a priori equivalent description. It should then be no surprise that competent speakers could doubt any candidate identity, or that it should be a synthetic matter when identities are established.

Rejecting a descriptive picture of reference determination for a class of expressions carries with it the need for a replacement account of reference determination for those terms. Proponents of these externalist theories obliged by suggesting that an appropriate socially transmitted chain of causal and epistemic influence might be sufficient to secure reference for many classes of terms. The idea can be filled out in a number of particular ways. One of these is proposed by Richard Boyd and applied to moral terms.

Boyd takes moral terms to have their referents determined just as the referents for natural kind terms are determined. He thinks that referents of natural kind terms are determined by a causally composed feedback loop from the referent to the use of the term. He writes:

> *Roughly*, and for nondegenerate cases, a term *t* refers to a kind (property, relation, etc.) *k* just in case there exist causal mechanisms whose tendency is to bring it about, over time, that what is predicated of the term *t* will be approximately true of *k* (excuse the blurring of the use–mention distinction). Such mechanisms will typically include the existence of procedures which are approximately accurate for recognizing members or instances of *k* (at least for easy cases) and which relevantly govern the use of *t*, the social transmission of certain relevantly approximately true beliefs regarding *k*, formulated as claims about *t* (again excuse the slight to the use–mention distinction), a pattern of deference to experts on *k* with respect to the use of *t*, etc. . . . When relations of this sort obtain, we may think of the properties of *k* as regulating the use of *t* (via such causal relations), and we may think of what is said using *t* providing us with socially coordinated *epistemic access* to *k*: *t* refers to *k* (in nondegenerate cases) just in case the socially coordinated use of *t* provides significant epistemic access to *k*, and not to other kinds (properties, etc.) (Boyd, 1988: 195).

Thus, according to Boyd, a moral term such as ‘right’ will refer to whatever property causally regulates in the above-described manner our use of the term ‘right’.

The theory is ready-made to explain the possibility of open questions. Non-experts can have a thought with a certain content in virtue of being members of a speech community, and yet not know what the experts know. And it is not necessarily obvious to even experts, let alone ordinary speakers of a language, what naturalistic features the relevant property might have.

They may be largely in the dark about the nature of the property which lies at the other end of the causal-regulatory feedback loop, although over time they can expect to learn more about it. Given that the most competent speakers of the language may not know that the property which plays this role can truly be identified via some naturalistic description or other, such speakers may have doubts about the identity of the moral property in question and the property naturalistically described. Hence the possibility of open questions regarding the identity of such properties, even for experts and even when the right-hand side of the identity statement open to question picks out the property using features which are essential to it and thus pick it out rigidly.

Moral Twin Earth

It is this theory which Moral Twin Earth is designed to refute. H&T believe that if Boyd's proposal is true semantically competent speakers should at least tacitly recognize its truth. Thus, even if such speakers do not know which natural property regulates their moral terms, they should at least tacitly know they refer to the property which appropriately regulates their use. Further, H&T believe we should be able to elicit this tacit knowledge by presenting a speaker with the right sorts of thought experiments and asking whether the people in those examples are using the words in question to refer to the same thing we refer to. This is how it works with natural kind terms like 'water'. Putnam and Kripke argued against descriptive theories and for their own theories by eliciting audience responses to various scenarios. The scenarios were devised to generate verdicts concerning the referents of various terms which vindicate the externalist theories. For example, these theorists elicited verdicts about the meaning of 'water' on Twin Earth, a planet otherwise like ours except that XYZ takes the place of H_2O on Earth. Speakers' agreement that the term refers there to XYZ and not to H_2O is crucial in vindicating causal regulation as a component in the determination of meaning and reference. Thus Horgan and Timmons suggest we should expect such tacit knowledge whenever a term refers in virtue of similar causal regulatory roles (1992*b*: 162–3).

Moral Twin Earth is constructed to test the hypothesis that moral terms work in this way. We are to imagine a planet much like Earth on which people use moral terms such as 'good' and 'right' in much the same way we do. People on this twin of Earth apply these terms to persons' actions and institutions; they take the "goodness" or "rightness" of an option to be important, and they are normally disposed to do what they believe is "right" and to choose what they take to be "good". So on the surface

Earth and Twin Earth are indistinguishable. At the same time we are to imagine that one natural property causally regulates our use of the relevant moral term here on Earth, whereas a different property causally regulates the use of the same term on Twin Earth. The properties are similar enough to account for common ways the terms operate on the two planets, but they are still distinct. A bare minimum of subtle but real differences in the psychologies of the relevant populations is allowed, so that somewhat different properties can play the same roles for the groups on each planet (Horgan and Timmons, 1992*b*: 164–5).

H&T then ask us to decide whether we would translate the moral terms on Twin Earth with our counterpart terms, or not.[5] The verdict that we should will cause problems for Boyd's theory. If such translation is correct our counterparts must mean what we mean by the terms. But if our terms and theirs mean the same thing, it cannot be that the terms on Earth and on Twin Earth designate different natural properties. Thus intuitions that the two populations are in genuine disagreement would indicate that the terms cannot function to designate whatever natural property regulates the relevant population's use of the term, since by hypothesis these are different on Twin Earth than they are here.

Unfortunately for Boyd's theory, competent speakers do have the intuitions that Horgan and Timmons seek to elicit with their example. Most people who have read their article seem to agree that the two populations address one debate about moral goodness, and that they are not talking past one another in virtue of using words with different referents. On this basis, Horgan and Timmons claim that Boyd's theory stands refuted, along with other similar cognitivist theories.

Taking stock

I myself have the intuitions Horgan and Timmons expect when they present the Moral Twin Earth example, and I am inclined to believe they tell us something correct about our moral terms. The speakers in the two communities are using moral terms with the same meaning, so that their dispute over what to do is a real dispute. If Boyd's proposal is in conflict with this, something must be wrong with it. Granting that, it is worth

[5] Readers should be reminded here of Hare's (1952) famous missionaries and cannibals argument which shares many features with the argument here. Moral Twin Earth aims at targets not on the scene when Hare formulated his argument, and hence involves setting up the thought experiment in a somewhat different way than he did. The similarities are noted by Timmons (1999), though the Twin Earth argument was not consciously patterned on Hare's.

carefully considering how the example causes trouble for Boyd's approach and what we can learn from the example.

A First Bit of Instructive Complication

We might begin by looking closely at what sort of variation there can coherently be between Earth and Twin Earth. Only certain sorts of variation with respect to causal regulation can be built into the thought experiment consistent with the similarities between the planets. If the populations of each planet really constitute linguistic communities, the use of a word by one member of a community will play a role in explaining the use of that word by other members of the community. Obviously each member of a community does not miraculously and independently coin a term orthographically and phonologically identical to those used by her community to express the same contents she will express with her new term. Terms are passed on from one member of a community to another. People learn them from their parents, friends, and neighbors and repeat the terms they learn. All of this requires causal interaction and it is these causal mechanisms that are responsible for people speaking the same language. Thus when Horgan and Timmons ask us to imagine a place different from ours with respect to the causal regulation of terms they cannot ask us to build the difference into this part of the causal chains from properties to speakers.

Rather they must ask us to imagine some variation leading to the original use of a term with a certain meaning such that that same meaning can then be passed on to subsequent users. We need examples in which two different properties stand in the same relation to original meaningful usage which don't change things so much that the kind of relations between the property and the speakers also changes. To see if this is possible we need to pay attention to the kind of relation Boyd postulates and ask whether two different properties could stand in this very same sort of relation to the speakers in question. Boyd's theory requires a causally efficacious feedback loop, from the referent of a term back to our use of the term such that the referent itself plays a role in causally explaining how we come to modify our beliefs expressed using the term so that they become truer over time.[6] It certainly seems possible that two similar but not identical

[6] At least that is how Horgan and Timmons interpret him and I think that this interpretation is probably fair. If you look at the long quotation about regulation taken from Boyd he is less than fully explicit about this. Still in saying that the use of a term is "causally regulated by" a certain property he seems to suggest that the property or its instances causally impact the use in question.

properties would be suited to playing similar causal roles with respect to a community's use of a term, and hence there seem to be possible scenarios in which each one is related as Boyd suggests the referent of 'good' must be.

Actually the issue is a bit trickier than at first it seems. Boyd puts an epistemic constraint on the nature of the causal relation such that too radical switches in the properties at the end of the causal chain might all by themselves turn a relation which meets Boyd's specification into one that does not. The causal regulation must be such as to make the beliefs of the community truer of the referent over time. Thus each of the properties must be such that what people come to believe as a result of the relevant causal relation is more nearly true than what they believed previously. Not just any property which caused us to modify our beliefs will do. Some properties may play a causal role in belief formation and yet not be otherwise such that the beliefs formed about them will be true or more true than what was previously believed. Thus the causally regulating property in the twin scenario must be sufficiently similar to the property playing the relevant role in the actual scenario that as our beliefs about rightness or goodness evolve they become more true of the hypothetical twin property as well as of the property actually playing that role.

Suppose things are as Horgan and Timmons stipulate: property A (the one which fits the role determined by consequentialist theory A) regulates use of the term 'right' on Twin Earth, and property B (the one which fits the role determined by nonconsequentialist theory B) regulates the use of 'right' on Earth.[7] But suppose also that nonconsequentialism is correct and that B is the correct version of nonconsequentialism. On the theory we are testing, the proposition expressed by 'X is right' on Twin Earth is that X has property A, and the one expressed on Earth is that X has property B. Now on both planets the population believes that right actions are the ones that ought to be done and make the most sense to do, or at least most people believe this. This is to say, those on Twin Earth believe that actions with property A ought to be done and make the most sense to do. And those on Earth believe that actions with property B ought to be done and make the most sense to do. Furthermore they

[7] I switch the target term from 'good' to 'right' because it bypasses a problem with the example as formulated by H&T. Nonconsequentialists may not differ from consequentialists over what they believe to be good, but over how the relative goodness of an outcome determines the rightness of actions. Typically nonconsequentialists deny that we should always do what leads to the better outcome, though it is possible to model such views using an agent-relative measure of goodness. We can overcome this problem by switching the example from goodness to rightness, since consequentialists and nonconsequentialists do disagree about which actions are right.

have come to believe this as a result of regulation by A and B respectively. But, on the assumption that B is the correct moral theory, only those on Earth will have made their beliefs truer by taking this commitment on board.[8]

Whether or not this result is fatal to the coherence of the example would seem to depend partly on our metric for judging when beliefs are more or less true. For no doubt some of our beliefs expressed using the term 'right' will be true of property A and it may have played an appropriate regulating role in generating those beliefs. Depending on how we weight the true beliefs as against the false ones, perhaps we will still want to say that A too has regulated our beliefs expressed using the term 'right' so that they become truer of A as we move along. Perhaps not.

It may not matter to the overall point what we decide. For we can modify the H&T example slightly. Suppose that on one planet the moral term is not causally regulated in the appropriate way by any property at all. The term which is not appropriately regulated will have to possess a different semantic value from the same term appropriately regulated by some property, at least it must if we take the analogy with the direct reference theorist's treatment of natural kind terms seriously. A moral term not regulated in the appropriate way by any property will likely have to be treated as something like an empty name, or a purportedly referring expression for which there is no referent. Empty terms of this sort are not synonymous with non-empty terms of the same sort.[9] If the correct intuitions about this case are the same ones H&T elicit with their original case, we can bypass the worries and reach the same result. Causal regulation of the relevant sort does not determine the appropriate semantic values for moral terms.

Still there is something to learn from the near failure of the example. It may be possible to specify the kind of regulation that determines reference in such a way as to rule out twinning the relation. It may be that the addition of epistemic constraints on regulation—similar to the requirement that the regulation makes the beliefs truer—would be the sort of specification that might accomplish such a task. I will come back to this idea later.

[8] If you're worried that it is question begging to posit that one or the other of these theories is true, remember that the Moral Twin Earth thought experiment is first and foremost a challenge to the semantic theory embodied in Boyd's proposal. The conclusion that moral theories cannot be true had better not be an assumption of the argument on pain of circularity.

[9] Just as our term 'Santa Claus' does not mean the same thing as a phonologically and orthographically identical term referring to a fat guy in a red suit on another planet. For more on this see Kripke's (1973: 156–8) discussion of Sherlock Holmes.

A Second Bit of Complication

A second area where closer examination is instructive encompasses the features of the thought experiment which underwrite our attribution of a common meaning to the terms. H&T propose that the crucial feature has to do with the action-guiding nature of the judgments in question (1992*b*: 170). We can assess this suggestion by looking at the way Moral Twin Earth is introduced to us. We are told that (1) the terms 'good' and 'right' are used to reason about considerations bearing on well-being, that (2) people there are normally disposed to act in ways corresponding to what is 'good' and 'right', that (3) these people take the goodness or rightness of options to be important, even over-ridingly so, in deciding what to choose, and that (4) the terms apply to actions, persons, and institutions (1992*b*: 164). Since this list summarizes our grip on Moral Twin Earth's similarity to Earth, we should expect to find the grounds of our sense that the two populations mean the same thing in this rather short list of common features. Items (2) and (3) spell out the action-guiding character of moral judgments, and it does seem reasonable to conclude they are responsible for our regarding the terms as synonymous with ours.[10] If we have any doubts concerning this, we might ask ourselves if the features listed in (1) and (4) by themselves would be enough to sustain the verdict. It seems they would not and hence it seems that the internalist features of moral practice are essential.

But I should register my doubts that Horgan and Timmons have listed *all* of the features of moral terms which account for our ascribing the same meanings to any terms used in the same way. In addition to those they list, our moral terms are used in such a way that their application supervenes on the distribution of non-moral properties of the items up for appraisal (as H&T recognize elsewhere[11]). An action-guiding appraisal of actions, persons, and institutions that did not depend on the otherwise specifiable features of these items would not be moral judgment as we know it, and we might be loath to translate terms used in this unguided way with our moral vocabulary. I suspect most readers of the relevant papers just assumed in the spirit of charity that this feature of our practices was among the

[10] The conclusion is reinforced by the similar verdict rendered when we are asked to consider Hare's missionaries and cannibals.

[11] Horgan and Timmons use almost identical language to describe the similarities between Earth and Moral Twin Earth in each of their papers employing the example (1991: 459; 1992*a*, 246–7; 1992*b*: 164.) Only the paper explicitly about supervenience (1992*a*) mentions it among the similarities. We might think of them as intending to include it in the others as well.

similarities intended by H&T. In any case we should add supervenience to their list.

A couple of other features of our moral judgments should probably also be presumed in the same spirit. The terms 'good' and 'right' are our most general terms of moral appraisal; presumably the corresponding words on Moral Twin Earth are similar. And a full description of Moral Twin Earth had better add something to distinguish the roles of 'good' and 'right' so that we have some reason to translate each with the phonologically and orthographically identical terms. That said, it does seem that the action-guiding character of a set of judgments, together with these additional features, is sufficient for using a term with the same meaning as our terms. This is a second lesson to take from the example.

A Third Bit of Complication

This leads naturally to a third insight we might glean from a closer examination of the troubles Moral Twin Earth does and does not cause for new wave theorists: The kind of internalism supported by the example is social in nature and not to be interpreted in an individualistic fashion. Moral terms are described by H&T as being only for the most part action-guiding. They suggest that people are normally but not always disposed to do what they regard as right and to promote what they believe good. In this way of setting up the case they are certainly correct, as the literature debating the merits of internalism has made salient.[12] The connection with motivation is still necessary insofar as it is needed to sustain the verdict that the 'right' on each planet means right. But what is necessary is that this is the normal situation in the population, not that any given member of that population be so motivated.

This sort of action-guiding character actually vindicates one feature of moral practice that the new wave realists wish to emphasize. The content of sentences and thoughts expressed using moral language is a function of the community of which one is a member. People in a community can express certain thoughts and say what they say despite their not being typical members of the community in the ways they make those judgments or use those words. The original version of that idea was used in part to explain how members of a community could use a word with its standard meaning

[12] Stocker (1979) and Brink (1986) present examples that seem to show that not all competent speakers of moral language need be motivated by their moral judgments. Smith (1994) and Dreier (1990) suggest ways of accommodating these cases within a defeasible internalist view.

despite dissenting from "constitutive" truths about the referents of those terms. But the Moral Twin Earth example and our reaction to it suggests that there can similarly be constitutive features of moral practice which are not universally observed by even those competent enough to use moral terms with their ordinary meanings and to have thoughts which would be expressed using those terms. Even if there are some individually necessary conditions on possessing moral concepts and using moral terms with the meanings we do, not every constitutive feature of moral practice needs to be exemplified in that way. Some features, in particular the features H&T use Twin Earth to illuminate, need only be exemplified by subpopulations of the community of which one is a member.

A New New Wave Theory

I think we can accept all of the foregoing morals highlighted by the Moral Twin Earth example while retaining the central ideas of Boyd-style theories, namely the parts of the theory that enable it to make sense of open questions.

From a certain point of view, it can be surprising that *every* proposed identity between a moral term and any candidate property picked out in non-moral terms is subject to doubt.[13] It is this general thesis that grounds the open question argument and which the new wave theorists are in a good position to explain. They explain it by making the referent of a term a function of some fact or facts of which competent speakers may be unaware. But to be a general explanation of the purported fact that for any description a speaker can regard it as an open question whether it refers to goodness or rightness, the theory has to allow such ignorance in a pretty strong way. Thus the theory must deny there is any description sufficient for uniquely picking out the referent of these terms of which competent speakers must be aware. This rules out not only various first-level substantive

[13] In a reply to a paper of Sayre-McCord's which is itself a response to Horgan and Timmons, Ernest Sosa wonders how anyone could have found the open question argument at all surprising (Sosa, 1997: 304–7). For on any theory there will be synthetic identities open to doubt by competent speakers. That the President of the United States is the George Bush can be doubted by competent speakers, but that by itself does not show that he is not. (Evidence from Florida would be more telling than any open question argument in calling that identity into question, whether or not the identity in fact holds.) Puzzlement about the power of the open question argument should diminish when we remember that Moore claimed that *all* identity statements involving moral terms are open to similar doubts by those competent with the concepts. In the absence of an alternative proposal for what property terms mean it can seem hard to find a content for them without invoking some description or other. What the various "new wave" theories contribute is precisely that sort of alternative.

characterizations of the referent, but also descriptions which embody the semantic theory proposed to explain how words come to refer to their referents. In other words if competent speakers can doubt every descriptive analysis of moral terms, they must be able to doubt whether the term 'good' refers to whatever property regulates use of the term 'good' of people in the same community as the speaker.[14]

One upshot of this approach is that the meaning of a moral term should not be equated with any such reference-fixing description. For a speaker can count as competent enough to use the term meaningfully and to have thoughts of the sort we would use the term to ascribe without believing that any such description is satisfied by the referent of the term. Yet the term is supposed to have a determinate referent, in the present case a property. The causal regulation thesis serves to provide that determinate referent for Boyd's account. It provides a fact or set of facts about the speaker's use of the term which narrows down the available candidates for the referent from those which would be available by limiting ourselves only to properties satisfying descriptions accepted by the speaker. In fact the candidate may not be one of those satisfying what the speaker would assent to, since the speaker may well have false beliefs about the referent. Boyd's theory is thus a version of externalism about content. Facts about a speaker's

[14] Let me be clear, since it might seem that the claim is too strong to be credible. The theories deny that any description sufficient for uniquely picking out the referent is such that competent speakers must be unable to doubt it. Thus, any identity claims or analyses involving the referent will fall into the dubitable class. But the theories need not and do not deny that a speaker must have some knowledge about the referent to count as competent. Perhaps there is some fact about rightness that everyone possessing the concept must understand, or perhaps there are a number of different facts or sets of facts each of which is sufficient knowledge for competence. The point is only that such knowledge will not be sufficient to ground an identity claim of the sort involved in an analysis. H&T imply that the use of thought experiments to support the new wave theories is in tension with this claim, for they think that the experiments show that the question Q7 is open, and that its openness counts against Boyd's theory: "Q7 Given that the use of 'good' by humans is causally regulated by natural property N, is entity e, which has N, good?" (H&T, 1992; 163). But if their argument really turned on the openness of this question, H&T would have gone through needless work to conceive of Moral Twin Earth. Q7 is one of the many open questions employed by the old-fashioned open question argument, and the complaint here is a version of that argument. If Q7's openness to competent speakers by itself refuted Boyd, H&T should merely have presented themselves as competent speakers who doubt the semantic theory postulated by Boyd and hence doubt that goodness is the property that causally regulates our use of the word 'good'. Lucky for all involved, the Moral Twin Earth example works independently of the original open question argument. It turns on the sufficiency of a population's practices to underwrite their using a word with the same meaning we do in circumstances where Boyd's theory predicts that the terms are not being used with the same meaning. It doesn't turn on the fact competent speakers can question Boyd's analysis.

usage and environment which can be unknown to that speaker contribute to determining the truth conditions for her thoughts and utterances employing the terms in question.

The particular external facts on which Boyd relies to determine reference have to do with which item in the world causally regulates a speaker's use of the term in such a way that people's beliefs about the referent of the term would over time become approximately true of the item in the world. Causal regulation by itself is not enough; the regulation has to also satisfy the epistemic constraint that it lead to truer beliefs over time. As we have already seen this makes it harder to construct the sort of Twin Earth scenario needed to refute the theory. I propose we see what we can do with just these sorts of epistemic facts, without requiring that the process in question be causal.

Suppose that (mimicking Boyd) we say something like this:

> *Roughly*, and for nondegenerate cases, a term *m* refers to a property *p* just in case there exist epistemically relevant procedures whose tendency is to bring it about, over time, that what is predicated of the term *t* will be approximately true of *k* (excuse the blurring of the use–mention distinction). Such procedures will typically require some members of the community to have an ability to recognize instances of *m* (at least for easy cases) which are employed in making judgments express using *m*, the social transmission of certain relevantly approximately true beliefs regarding *p*, formulated as claims about *t* (again excuse the slight to the use–mention distinction). When relations of this sort obtain, we may think of what is said using *t* as providing us with socially coordinated *epistemic access* to *p*. We have this sort of socially coordinated epistemic access when a good number of the beliefs about the referent of 'm' in the social context non-accidentally track what is going on with *p*. In other words, we have it when a good bit of what people say and believe using 'm' is approximately true, and it is said and believed because of how it is with *p*: *m* refers to *p* (in nondegenerate cases) just in case the socially coordinated use of *m* provides significant epistemic access to *p*, and not to other properties—that is when what is said using *m* expresses knowledge about *p*.

This modification of Boyd's proposal replaces causal regulation with epistemic regulation of roughly the sort that divides merely true beliefs from knowledge. We can think of this as a generalization of Boyd's idea regarding causal regulation. Knowledge requires that our beliefs be non-accidentally true. If our beliefs about some matter count as knowledge about some thing, they must not only represent matters correctly but we must have them because things are thus with the thing about which we have knowledge. For ordinary natural kinds, facts about those kinds can only explain our knowledge of those facts if we have causal contact with the kinds. When you and I believe that water is heavier than a similar volume of sawdust, our belief is not a lucky accident or a guess precisely because we have had contact with and thereby know something about water's weight. While XYZ on Twin Earth is also heavier than sawdust, there is no sense in which our believing that could constitute knowledge. For we would need empirical access to the stuff in order to know this.

Thus for many properties a causal connection to instances of its instantiation may well constitute the epistemic relation needed. But not every required epistemic relation will be such a causal relation. Some epistemic connections are a priori. It can be right to say that a person believes a necessary truth about an abstract entity *because* it is true, without their belief being caused by that entity or that truth. The 'because' here stands for some sort of relation that need not be causally implemented. I'm not sure I know what the relation is, but I think the notion is one we sometimes invoke and one which plausibly has an epistemic role to play (Nozick, 1981: 287). The above statement of the revised Boyd-style view employs it to cash out a more general version of his suggestion that the relation between a term and its referent might necessarily be epistemic.

I need to add one further refinement to this suggestion, one I will argue for in the next section. Some properties are more eligible candidates for the referents of a term than others. We can think of the most eligible candidates as the most natural properties available for the sort of endeavor speakers engage in when they are using a term to talk about a property. A term 'm' refers to the most eligible candidate property consistent with the constraint that speakers of the language who use 'm' use it to express knowledge about *p*. Since moral thought and talk has distinct purposes from other disciplines the most eligible properties will be natural moral kinds.[15]

We are now in a position to sketch how the revised account helps us handle the Twin Earth scenario. Since the Twin Earth argument is a semantic objection to moral realism, we can make various realist metaphysical and epistemic assumptions without begging questions against it. I'll suppose the following. Some actions really do make sense to do whereas others don't and some make more sense to do than others. Actions make more or less sense to do in virtue of other features that they have, and often these features are naturalistic. Furthermore, people are not entirely ignorant of this. In particular people sometimes do things because they make sense to do and when they do so it constitutes acting from knowledge.[16]

[15] Geoffrey Sayre-McCord (1997) in his paper on the Moral Twin Earth argument has suggested that we need to substitute moral kinds for natural kinds in the Boyd account. But without also modifying the story about the kind of regulation involved in the Boyd-type theory (from causal to epistemic) I don't think the suggestion by itself does the trick.

[16] I take the idea that we are talking about what it makes sense to do from Gibbard (1990: 49).

On these assumptions, people might well want/need a vocabulary to talk about what makes sense to do and what does not. Suppose such a vocabulary develops, and people use it in ways that somewhat track the facts about what makes sense to do, and that the judgments vary with some of the naturalistic features of the actions they are assessing. Suppose further that their judgments using a term 'r' come to more and more closely track what in fact makes sense to do, and that this is not accidental. Suppose still further that they tend to do what they label with 'r'. We might be justified in thinking that they use this term 'r' as they do because of facts about which actions have the property of making sense to do. If that is correct, the proposed semantic theory for moral terms would license us to conclude that 'r' refers to the property actions have when they make sense to do. (Call this property **R**, to save space.)

Two communities could stand in such a relation to **R** without agreeing on all of their judgments. And they could stand in this relation to that property even at the limit of enquiry. One or both communities might be disposed to get things somewhat wrong about what makes sense to do, even though most of their judgments are correct and even though the correctness of these judgments is to be explained by their having knowledge of **R** which they express using the term 'r'. In other words, they could be so disposed that they were causally regulated in such a way as to track a property somewhat divergent from **R** at least over a certain range of cases.

Yet, and this is the important point, if the causal factors that explain their judgments not accurately tracking **R** count from an epistemic view as mistakes, they do not for all that vitiate the claim that **R** is the referent of their term 'r'. For our modified theory counts only epistemically relevant regulation as serving to determine the referent of a term. How do we determine when people are disposed to make judgments that track epistemically relevant factors as opposed to when they are disposed to make mistakes? That is a difficult matter, but one factor in this determination is whether they are disposed to judge in ways that track relevantly natural kinds. When two populations both use overlapping action-guiding terminology but are disposed to diverge over a range of cases, the population whose judgments most closely track the relevant moral kinds is getting things more right than the other population. But for all that, the proposed semantic theory will attribute the same content to their judgments and suggest that one population is making a mistake.

This is what I suggest is going on with the Moral Twin Earth example. The two populations are both tracking well enough what makes sense to do. At most one population might be disposed to get it right in the long run. This is how I take H&T's stipulations regarding causal regulation in the example. But both populations may be using 'right' to refer to what makes sense to do, since both are using the term because they have (some) knowledge of what makes sense to do and they express that knowledge using the term 'right'.

Using Natural Kinds to Refine the Proposal

That anyway is the capsule summary of the position. In this section I'll give more detail on using natural kinds to winnow down the eligible candidates for term reference and motivate that component of the proposal. Then I'll provide some more commentary on how that helps explain the intuitions about the Moral Twin Earth cases.

We need some way to narrow down the range of referents beyond what even a causal regulation account could do on its own. For any finite series of events in which a particular property is instantiated, there are many properties that are instantiated by each of the events in the series. At least there are on the relatively liberal conception of properties where new properties can be created by conjoining and disjoining old properties. I take it that when we say that a property causally regulates people's use of a term, we mean that to be shorthand for the idea that events in which that property is instantiated play a role in causing the beliefs that people have and express using that term. The problem arises because each of these events involves the instantiation of many properties and no matter the number of instances involved there will be multiple candidates to be the designee of the term in question.

So we need to get from the plethora of different properties, the instantiations of which have causally interacted with a community of speakers, to the specific property which is the referent of the term. There are two suggestions one might immediately think of to try to narrow the range. One would be to limit the candidates to those properties which are relevant to causing people to have the beliefs that they do. We might hope that the instantiation of massiveness is causally relevant to people's beliefs about what is massive, but that the Cambridge property of being massive or in New York would not be. But there are problems with this approach that lead me to think it cannot be right. It might be that people are prone to a certain sort of error in certain circumstances, so that the absence of those circumstances is also causally relevant to our forming the beliefs that we do. Being massive and not in near zero gravity might be a causally relevant property when we are thinking about what causes our beliefs about massiveness. Yet we want the account to allow us to express beliefs about massiveness not just massiveness in higher gravitational environments.

Furthermore, being prone to error can bring in another sort of problem. We do in fact form beliefs in circumstances where we do make errors. We don't want it to turn out that our beliefs are true of some more complex property constructed out of the property we would ordinarily take ourselves to be designating along with the property the instantiation of which on a given occasion led us to make the error that we do. We want our beliefs about loudness to be beliefs about loudness, not beliefs about loudness or high distortion. But given the way the world actually works many of us mistake high distortion for volume. It could be that the beliefs that we normally express using the word 'loud' are causally best explained by citing our experiences of either loudness or high distortion. And it might even be that they are as true of that disjunctive property as they are of the non-disjunctive property of loudness.

These shortcomings may lead one to view a second suggestion as more promising. Why not allow the dispositions of the speaker's community to determine the property in question? But once again we face a problem. For sufficiently anomalous cases our community may be disposed to error about the extension of the property. Yet that would seem to be ruled out if the actual dispositions of the community to make judgments about the extension of a property made it the case that our judgments were about a property with that extension. This is the analogue for properties of the objection Kripke runs against the idea that dispositions can be used to determine which function we have in mind when it looks like our actual behavior is consistent with any number of candidate functions. At some point our dispositions run out (Kripke, 1982: 22–39). What I think this shows is that no causal or dispositional approach can all on its own uniquely determine the referent or semantic value of any such terms. It isn't that causation cannot play a role; it can narrow things down. But it needs supplementation.

So we need some way of narrowing down the range of candidate properties so that only some are eligible to be the contents of our judgments. David Lewis (1984) has suggested that we should employ the distinction between natural and unnatural properties, or better more and less natural properties.[17] In situations where our dispositions and practices under determine the referents for our terms, the more natural properties are eligible candidates for those referents whereas the less natural are not. Lewis's idea seemed to be that natural science would be the arbiter of naturalness, though I'm not entirely sure that is what he had in mind. In any case, I want to suggest a modification of this idea. Naturalness should be

[17] I thank Michael Smith for reminding me of the importance of Lewis's idea for the sort of semantics I am trying to work out here.

seen as discipline-relative. The kinds or properties which are more natural for the purposes of physics may not be the same as those which are more natural for purposes of biology. The more eligible semantic values for one's terms when engaged in the former may or may not be the same as the more eligible semantic values for one's terms when one is engaged in the latter. Probably the naturalness of a property or kind for a given discipline is relative to the subject matter and purposes of the discipline. I can't say exactly how this is supposed to work, but my sense that naturalness must be discipline-relative stems partly from thinking that when I'm being appropriately responsive to the questions posed within a discipline and the evidence we have for different hypotheses, certain methods of classification and of picking out properties relevant to the theory seem more natural to use than others.[18] And it seems to me that what seems natural to me in one sort of domain is not what seems natural to me in another. Tables seem perfectly natural for anthropological purposes, but not for the purposes of physics.

With this modification, I propose to follow Lewis in employing the naturalness of kinds to determine the eligibility of kinds for referents of terms. Even though the actual dispositions of speakers and thinkers are insufficient to rule out gruesome interpretations of their talk, the greater eligibility of the more natural properties to be the referents of terms can be used to single out the more natural property as the referent. Once we have such a notion to be employed this way, it can also be employed to overcome the related problem that people can be disposed to make mistakes. For such cases we can allow the greater eligibility of the more natural kinds to override the actual dispositions of speakers. Two speakers who are differently disposed to use some term over a range of cases can still be said to be using it with the same meaning because we expect them both to be referring to more or less natural properties, and thus we chalk up their divergence as due to error.[19]

[18] The idea that naturalness might be discipline-relative is not particularly original with me, but it does seem to deviate from Lewis's proposal. His discussion allows that different theories will employ different classificatory kinds, but has all of them ordered along one scale for naturalness. Thus, ordinary artifact terms such as 'table' select a less natural kind than electrons, though the fact that tables form a more natural kind than more disjunctive or gruesome kinds still figures in 'table' designating tables. My own intuitions are that tables are just as natural as electrons unless we happen to be asking the sorts of questions that physics aims to answer.

[19] In a nice paper, David Copp (2000) suggests that the referential intentions of speakers, plus their interests can be used to similar effect in fixing the referent of terms within a Putnam inspired naturalistic semantics for moral terms. And he uses this to suggest that terms on Earth and Twin Earth might mean the same thing. This idea may be very similar to the suggestion I am making here, though he thinks that there is an important contrast between Putnam and Boyd with respect to the applicability of this

It should be relatively easy to see how the idea extends to speakers causally regulated by different properties in their use of a term. For I take it that the idea of causal regulation involves judgments about which properties are in fact causing the speaker in question to use the word as they do. Normally different dispositions with respect to their use of the term will reflect difference in which properties are causally relevant to their uses of the term, even where those dispositions are not manifested. On the indicated approach to kind reference, differences with respect to those dispositions can sometimes be ignored in the interpretation of a speaker's utterances. And on the modification of the Boyd approach that I advocate here, we are allowed to make a similar move. What matters is not which kinds actually regulate a speaker's use of a term, but which kinds it makes sense to think that community members are talking about, given the overall uses to which they put the terms and the ways in which they make judgments about when the terms apply.

This is also the picture suggested by some of the well-known examples employed in the externalist literature about the meaning of natural kind terms. As Putnam (1973) has pointed out, we regard people's use of the term 'water' prior to 1700 as designating and meaning just what *we* designate and mean by that term. But our use of the term as we now use it depends on things that we have learned since 1700. It can't be just the fact that we do use this term in this way that makes it the case that those in the past use it with the same meaning as we do. It has to be because it is non-arbitrary that we so use it. We use 'water' to refer to H_2O and not liquids containing either H_2O or some nearly indistinguishable liquid of a different chemical makeup which they were not in a position to rule out as the referent of their terms. Our ruling that alternative out could only be a reason to interpret their meaning as our meaning if we think that our choice is governed by the evidence and rationally responsive to that evidence. If we retrospectively decide that we have made a mistake, either because we lack some evidence or because we used bad judgment, we should change our judgments about what meaning the term 'water' had in 1700. And we should change our view of what the meaning and referent was all along, not think that our discovering a mistake changed the meaning.

On this way of looking at things, once a certain basis for use of the term has been established, there are two further factors determining the referent. There are facts about the kinds to which the speakers have epistemic access and there are facts about which kinds are most natural. Neither by itself is

idea and he thus does not try to work it into a regulation-based story of the sort Boyd suggests. I'd like to downplay the role of speaker's referential intentions more than Copp would, partly because they can often be very vague and I would like a somewhat less individual conception of speaker's interests than he seems to employ.

sufficient to determine a referent, since speakers will have some access to gruesome kinds related to those we intuitively regard as the referents and since there may be some natural kinds to which they don't have sufficient epistemic access.

Once we have this much, we should allow that even interpretations that past users of the term have ruled out might in the end come to be seen as the correct interpretation of their language. This anyway is the upshot of another well-known example, Dalton's use of the term 'atom' (Burge, 1993). Dalton apparently defined atoms as the smallest particles into which matter could be divided. He suggested that all matter was made up of such atoms and also apparently believed that something like the periodic table captured differences in features of different kinds of atoms corresponding to certain kinds (Burge, 1993: 316). People are generally inclined to take Dalton to refer to atoms when he used the term 'atoms'. Yet his choice of definition seems to have explicitly ruled out any interpretation on which the referent was not the smallest indivisible particle of matter. But, if we regard him as offering a mistaken definition of his term, a term which has its reference determined in part by the true things he thought about atoms but which he did not take to be definitive, this need not be *our* verdict. It is fair to suppose that the experiments he did to determine that something like our periodic table correctly represented real features of what he called 'atoms' gave him knowledge of atoms. And these could form enough of a foundation to make it the case that his term and ours have the same meaning and the same referent. So to use a term correctly a speaker or community of speakers must have some knowledge about the referent, but that knowledge need not be what they themselves would offer as the defining features of the referent of the term.

Filling in the Details for Moral Twin Earth

Boyd's proposal as understood by H&T[20] uses the tendencies of the causal mechanisms over time to narrow down the range of candidates beyond what can be determined from the actual events that have occurred so far. And it is precisely this that leaves him vulnerable to the Moral Twin Earth objection. Given that the relevant mechanism is causal and hence only nomically necessary, we can coherently imagine alternative mechanisms which would focus the range on another candidate property. Thus it is fair for H&T to stipulate examples in which two populations have different dispositions

[20] Or at least one fair interpretation of that proposal, which Horgan and Timmons adopt in constructing their counter-argument.

to judge, so that the sets of properties instances of which regulate their judgments diverge or will diverge over time.

The modification I defend here uses two related ideas to avoid the problem. First, by emphasizing that the tracking criterion that makes our terms designate is epistemic and not always causal, we see how that idea can still yield common designees for our terms in the absence of the sorts of common causal mechanisms the original Boyd story required. Regulation of a community's use of a term by a property is only relevant to the designation of that property when that regulation yields knowledge of the property. And if some not-wholly-causal epistemic route generates the relevant sort of knowledge it can be relevant to determining the referent. This allows us to characterize the causal or dispositional divergences in Twin Earth cases as constituting errors or dispositions to err and discounting them in determining the referents of the relevant terms. Secondly, the modified approach suggests that the naturalness of a classification is epistemically relevant and that naturalness is discipline-relative. It is partly because of this that we can classify certain responses as mistakes. Sometimes a particular judgment will be best interpreted as involving mistakes about a natural kind rather than knowledge of an unnatural kind. By using discipline-relative facts about naturalness to narrow the range of candidates for the referents of moral terms we explain why relevant twinning scenarios cannot be constructed, since they require assigning an unnatural kind as the referent of one of the terms in question.

Relatedly, the proposal cites both necessary and contingent features of our world to determine the designee of the term 'right'. Since the relevant contingent features are stipulated to be the same across Earth and Twin Earth they can be used to generate the desired semantics. And insofar as the necessary features are necessary they will be available for that purpose on both planets. The contingent designation-determining facts are that (1) the term 'right' is used to reason about considerations that bear on well-being, that (2) people there are normally disposed to act in ways corresponding to what is 'right', that (3) these people take the rightness of options to be important, even over-ridingly so, in deciding what to choose, that (4) the term applies to actions, that (5) the rightness of an action is treated by competent speakers as being determined by features of the actions describable without using that moral term, and (6) 'right' is the most general term of moral appraisal which applies specifically to actions. The first four are the features of speaker's use of the term 'rightness' that Horgan and Timmons hold fixed between Earth and its moral twin, and the last two are features I claim need to be added for their argument to go through.

The relevant necessary designation-determining facts are that (1) some actions really do make more sense to go in for than others, (2) that if an action makes more sense than an alternative that fact is a reason to do it rather than the alternative, (3) that options make more or less sense because of the non-moral features that they have, and (4) that among the relevant features are things like how they effect the well-being and happiness of oneself, other people, and other creatures, and so on.

When a someone here on Earth tells me that they are doing an action because it is kind to so and so, and that its kindness makes it right, I take them to be expressing and acting on knowledge. For I take it that kindness does (often) make actions right, and that this is a reason to do actions of that general kind. The knowledge here expressed combines necessary and contingent facts and facts which are knowable only empirically with facts that seem to me to be discoverable through reflection and which hence are a priori. It is an empirical discovery (though often an obvious one) of contingent truths that certain ways of treating people make them happier and better off, and we need to know which kinds they are to treat people kindly. It is a necessary truth that making people happier and better off is something that makes sense to do and which we have reason to do (other things equal). This, I think, is knowable a priori though it may be that some empirical input is crucial in figuring it out at least for some people.

Does anything about the shift between Earth and Twin Earth shake our confidence that Twin Earth speakers have knowledge of these same facts? It is built into the example, partly in virtue of the internalist features of the Twin Earth scenario that there will be people like our speaker here on Earth, who take actions to be right because they are kind, and who act on that knowledge by doing what they take to be kind and right. The twinning operation did not in fact shift the major facts of human psychology and biology. Thus kindness on Twin Earth should be the same as it is here, and the same sorts of epistemic procedures should be relevant to finding out what it takes to be kind to another. In cases where Earthlings have knowledge about kind actions, their counterparts on Twin Earth will as well. What about the fact that in these circumstances kindness constitutes rightness? This, I think, is one of the a priori knowable and necessary truths. One does not know an a priori knowable truth just in virtue of believing it. One might believe it for the wrong reasons or might not have gone through the a priori reasoning, imagination, or reflection necessary for putting oneself in a position to know. But supposing that our Earthling friend has put herself in a position to know by reasoning appropriately, her twin will have done the same.

It is this basis which I think forms the foundation of knowledge sufficient to determine a common referent for the terms here and on Twin Earth. But it can do so only if we can view any divergence over the extension of the term (beyond the sort attributable to ordinary vagueness of terms) as involving a mistake by one or both parties.

Suppose Earthlings and their counterparts are disposed to diverge in their judgments about what is right and what is not over a significant range of cases. One way this could come about is if the empirical component of their epistemic processes diverges. They might come to disagree about which acts are kind due to disagreement about what sort of nervous system one needs to feel pain. Here we have no trouble ruling out one view or another as just a mistake, relevant neither to the referent of 'pain', nor to the actual extension or referent of 'kindness'. Suppose instead that the disagreement comes out in disagreements over which naturalistically described actions count as right such as when kindness can be too demanding. If this is not just a disagreement about a borderline case and realism about rightness is correct, there is an answer to this question. The answer here must be necessary, whatever it is.[21] The dispute here is over which way of conceptualizing rightness gives us a more natural moral kind, that is over which conception cuts moral reality at its joints. Except for borderline vagueness, at most one view will be correct, so at most one speaker will be reasoning correctly and getting things right. And since one can be justified a priori in accepting views about what makes an action right, if the person who gets it right has gone through the relevant epistemic reasoning, her beliefs may be justified and count as knowledge.

This helps explain the difference in our intuitions between the standard Twin Earth scenarios and the Moral Twin Earth scenario. Where the referent is a naturalistic kind, and it is a contingent matter whether the truth of our beliefs about a kind is accidental or not we can stipulate relevant twin scenarios. In the original Twin Earth story, much of what we believe about water is equally true of XYZ as it is about H_2O. But only with H_2O is what we believe non-accidentally true. That's because most of what we believe about water is composed of contingent empirical claims. To say that they are contingent is to say that they could have been otherwise and hence we need information from the actual world to figure out whether they are true, and that is to say they are empirical. And to say that they are empirical is to say that these contingent facts must have a causal impact on our powers of observation. It is because H_2O has had that sort of impact on us that it is no accident that we believe what we do about water.

On the other hand, with the kind of moral case we are now envisioning, no such epistemic procedures are relevant. If we are in dispute about how much kindness morality can demand, and we have the issue clearly

[21] The correct way of conceiving of the supervenience of the moral on the non-moral will assure that much.

defined, no actual experimental evidence seems relevant to settling that issue. What could further empirical information tell us? The issue in dispute is hypothetical and does not turn on contingent and empirical matters. If kindness requires a certain sort of sacrifice is it still morally required? This does not mean that our beliefs about this could not be accidental and hence not knowledge. But it does mean that what it takes for them not to be so is not a matter of causal contact with the property in question. Rather we have to have thought about the issue in the right way so that our beliefs are correct because we have thought about the relevant issues correctly or nearly so. If I'm right this will require correct reasoning and vivid and sympathetic imagination. If after imaginatively and sympathetically thinking things through we arrive at the correct view about what to do, this will be no accident, and the epistemic constraint that I think forms the heart of the Boyd-type stories will be satisfied. Some of our beliefs about rightness will count as knowledge and thus our term 'rightness' will refer to rightness.[22] So will the analogous beliefs of those on Twin Earth.

It may be possible to stipulate a scenario as a limited test case for the modified theory. Imagine that the things we express using the terms 'right' and 'rightness' are not known because it is only an accident that people believe them even though they are true. It is hard to describe the sort of case required but it is worth trying. The scenario requires that our twins use their term 'right' in a supervenient way, not because they think it makes more sense to treat like cases alike in evaluating actions, but for some other reason. And similar things must hold for their treating rightness and well-being as linked and their thinking that the rightness of an action counts in favor of doing it. They must have reached these conclusions not by thinking imaginatively and reasonably about the alternatives but through some epistemically irrelevant process. Maybe throwing darts at a special dartboard would be an example of such a process. If that is the story I no longer think their word means what ours does. If so, this is some confirmation of the proposal.

To conclude this section I'll summarize. A suitably modified version of the sort of externalist theory that Boyd uses can explain how reference to properties can be secured for moral terms even while competent speakers

[22] Sayre-McCord (1997) has suggested that changing the Boyd account to specify that the property in question be a moral as opposed to natural kind would dissolve the troubles caused by H&T's scenario. I agree with using moral kinds as the relevant sort of natural kind for the domain. But by itself this move doesn't overcome the problem. The H&T example is a problem for the reference-determining mechanism proposed by Boyd and not for the particular referent proposed. So changing the referent without changing the mechanism will not by itself provide an answer to the objection, though providing an answer will involve changing the referent from what Boyd thinks it is.

can raise questions about their reference. And it can do this while remaining immune from Moral Twin Earth counter-examples, chiefly because it places a greater emphasis on the epistemically relevant features of the proposal, and because moral epistemology makes it harder to shift the necessary epistemic facts in such a way as to provide a counter-example.

Why Believe Any of This?

Is there any reason to accept the semantic story, apart from allowing us to avoid Horgan & Timmons's clever counter-example? Obviously I think so. Metaethical arguments are always a species of argument to the best explanation. A certain range of phenomena are taken to be data about moral practice and we construct theories to explain those data. And the proffered semantics for moral terms is part of a package which, as I've employed it, includes a commitment to a certain sort of rationalism. Whether we should accept this package depends on how it fares in explaining the various data about ethics that we think needs to be explained. In addition to being a clever counter-example, the Moral Twin Earth story highlights a number of the data that a good metaethical theory needs to explain. So a theory constructed to handle that example will, if all goes well, score rather well in the contest that arguments to the best explanation set up.

One datum is that competent speakers can doubt the identity of rightness and any property picked out in other terms. The theory I offer shares with Boyd's the ability to explain open questions about true identities wherever those identities are empirically established or only empirically establishable. But it surpasses Boyd's theory insofar as it explains how open questions are possible even when the reasons to accept a true identity statement are a priori. One thing the Twin Earth example does is to highlight the need for this. If competent thinkers can doubt the identity of rightness with any property picked out using some other form of words, either it must be possible to question even a priori establishable identities, or there must be no a priori arguments for thinking rightness identical with any property. Given the game that many of us are involved in—offering relatively a priori arguments for the identity of moral properties, we need a story to tell about how such doubts are possible. And the theory on offer suggests that we can find our explanation by noting that even a priori truths may be knowable only through a process of reasoning and that we can always sensibly wonder whether our reasoning about some matter has been correct.[23]

[23] The point is obvious but a surprising number of theories flout it. Take e.g. models of belief revision on which all of the a priori truths automatically get a credence of 1.

Another datum, put very blandly, is that moral judgments have some sort of tight connection with action. Yet that tight connection seems compatible with some people not being moved by moral judgments they accept. On the one hand we would not translate a term using our term 'right' if people were not normally inclined to do what they took the term to apply to. On the other hand it seems perfectly conceivable that some people not be so moved, and even that some people expressed doubts about the rationality or sensibleness of doing those things. Not only does it seem conceivable, but actual people seem sometimes to be like this. Insofar as Earth and the stipulated Moral Twin Earth share these features, the fact that we take them to be speaking univocally suggests that this sort of tight connection may be grounded in just these similarities.

Our package endorses this hypothesis and explains why it might be correct. If enough people in a community get enough things right about a property, so that in principle there is available to all speakers in a community an epistemic pathway to finding out more about that property, then all members of that community can be credited with thoughts and talk about that property, even if some of what they say is false and even necessarily false. One way (among others) that the belief that something makes sense to do can manifest itself is by doing it when one is in a position to do so (van Roojen, 2002). My suggestion is that those who do mostly do what they think right are acting on a belief whose content is that these things make sense to do. And, I suggest, this belief counts as knowledge if it stems from noticing that some features of actions make them make more or less sense to do, when it is the features of the action that in fact make it make sense that grounds their judgment about the case, and when they reliably (though not infallibly) discriminate the sensible from the inadvisable. Thus most of the populations of both Earth and its moral twin meet the conditions for getting things right much of the time. Those who are not motivated to do what they regard as right are making a mistake, perhaps culpable perhaps not. But given that their fellow inhabitants get things right and that their use of the term 'right' depends on the practices of their fellows, they too use the term with the same meaning and it refers to the same property in their mouths as it does in the mouths of their friends.[24]

The theory then supports the sort of internalism suggested by the Twin Earth argument. When and only when most people in a community are

[24] Conceptual role semantics of the sort advocated by Ralph Wedgwood in the conference draft of his paper for this volume may also be able to explain this form of internalism requirement allowing some people not to be appropriately motivated, but only if it takes the form of broad rather than narrow conceptual role semantics. Wedgwood prefers the narrow version.

guided by the application of a certain term in choosing what to do, and when certain other requirements are met we will think of it as an evaluative term. But this does not rule out the sort of amoralist often invoked by externalists to refute internalism. We can think of these amoralists as something like the patient who thinks that he has arthritis in his thigh—someone whose membership in a community gives him competence enough to make judgments using a concept, even when the most competent members of that community would regard some of what he believes as incoherent.[25]

There is one further desideratum that many metaethicists find important. This is that the theory fit into a broadly naturalistic world view, whatever that comes to.[26] So it might be appropriate to say something about whether the resulting theory counts as naturalistic. My answer is that it depends, and that a lot of what it depends on is not special to ethics.

Naturalism is usually stated as a contingent thesis. The world might have been such as to contain non-natural things but it does not. And while this idea seems coherent enough when applied to concrete entities, how to apply it to properties is not all that straightforward. It seems like there could be three sorts of properties—those that could be had only by natural objects, those that could be had only by non-natural objects, and those which might be had by either. Leaving aside the difficult question of what it takes for an object to be natural, only the third sort of property cannot be instantiated in a world with only natural objects. So probably naturalism should be conceived in a way consistent with the existence and even the instantiation of properties which could be had by non-natural entities.

Now it looks like rightness and goodness ought to be thought of as this sort of property. On any favored partition of entities into natural and non-natural, it seems like the non-natural entities would—if there were any—be morally evaluable, and actions with effects on them might vary in rightness in virtue of those effects. If we think naturalism rules out gods and ghosts, it does not seem to follow that killing a god or a ghost would not be wrong, if only we could do it. Nor does it follow that some gods or ghosts could act wrongly or be morally evil.[27] These examples invoke only one way of partitioning the natural from the non-natural, but it seems like

[25] See Burge (1979). I discuss this analogy in more detail in my manuscript, 'Moral Rationalism and Rational Amoralism'.

[26] One might be forgiven for suspecting that a non-cognitivist analysis of 'natural' and 'naturalism' has something going for it.

[27] One of the best reasons to doubt the existence of the sort of god postulated by most Americans is precisely that no such god could have all of the other features commonly supposed and still have acted rightly. Such a god would be very bad (contrary to what most believe), but the truth of that claim relies on their being properties that could be had by nonexistent non-natural entities which would not be natural if there were any.

the point is pretty general. For any other coherent sort of thing that you might think there could have been but as a matter of fact is not, it seems like features of these things might be morally relevant.

The point here is really not specific to evaluative properties, since I think there are other properties that could be instantiated in a world of only natural items, but might also be instantiable in non-natural worlds. Existing in close proximity to five other distinct things seems like a candidate. So if you think that naturalism is incompatible with properties of this sort I think you should conclude both that my story is not compatible with naturalism, but also that naturalism is false. That option creates no problem for my account.

Sometimes in metaethics when people suggest that they are naturalists they don't seem to mean to rule out very much. They admit that the possible extension of moral properties would go beyond any very substantive delineation of the natural, admitting ghosts, forces, or entities not substantiated by natural science, or anything else we can come up with. What these folks seem to care about is that moral properties not be distinct from their supervenience bases. Given the argument above, these supervenience bases will include an awful lot of pretty weird stuff—stuff that seems pretty unnatural to me. So the concern here is different from the concern that motivates those who emphasize that naturalism is a contingent hypothesis.

My response to this specification of naturalism is that my favored theory is neutral on it, but that the issue turns on nothing all that special to ethics and much more on general metaphysical principles. If there is reason to identify properties with the same extension as infinitely disjunctive properties, that reason would apply here. Whether that identification is best thought of as a reduction of the higher-level property to the lower is also an interesting question.[28] But whatever the answer to these questions, the sort of view I suggest here seems ready to accommodate the verdict. Moral terms refer to properties about which some of us have knowledge which we express using those terms. If those properties are identical to properties picked out with infinite disjunctions of non-moral terms then our moral terms refer to those. If they are not identical they do not. There is no problem for the account if this sort of naturalism turns out to be true. And in fact insofar as the theory remains able to explain the possibility of open questions it should be welcomed by naturalists who identify moral properties with such constructions from natural properties.

In summary then, the proposal allows us to use a modified version of Boyd's original idea to handle the Moral Twin Earth example. And this

[28] See John Gibbons's as yet unpublished paper, 'Identity without Reduction', for an interesting discussion of the issue.

version also has the virtue of explaining two features of the example that we should wish to explain anyway. First it explains how any identity claims about moral properties can be open for competent speakers. And second, it explains how a very weak sort of internalism—one which postulates a necessary connection between moral judgments and action-guidingness for many members of a community—could be required. And it does both of these things without violating any naturalist strictures that we should be inclined to accept.

REFERENCES

Boyd, Richard, 'How to be a Moral Realist', in G. Sayre-McCord (ed.), *Essays on Moral Realism* (Ithaca, NY: Cornell University Press, 1988), 181–228.

Brink, David, 'Externalist Moral Realism', *Southern Journal of Philosophy*, supplement (1986), 23–40.

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

Burge, Tyler, 'Individualism and the Mental', *Midwest Studies in Philosophy*, iv, ed. P. French (Minneapolis: University of Minnesota Press, 1979), 73–121.

____ 'Concepts, Definitions, and Meaning', *Metaphilosophy*, 24/4 (1993), 309–25.

Copp, David, 'Milk, Honey and the Good Life on Moral Twin Earth', *Synthese*, 124 (2000), 113–37.

Dreier, James, 'Internalism and Speaker Relativism', *Ethics*, 101/1 (1990), 1–26.

Geirsson, Heimir, 'Moral Twin-Earth: The Intuitive Argument', *Southwest Philosophy Review*, 19 (2003), 115–24.

Gibbard, Allan, *Wise Choices, Apt Feelings* (Cambridge, Mass.: Harvard University Press, 1990).

Hare, R. M., *The Language of Morals* (Oxford: Oxford University Press, 1952).

Horgan, T., and Timmons, M., 'New Wave Moral Realism Meets Moral Twin Earth', *Journal of Philosophical Research*, 16 (1991), 447–65.

____ and ____ 'Troubles on Moral Twin-Earth: Moral Queerness Revived', *Synthese*, 92 (1992*a*), 212–60.

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

____ and ____ 'Nondescriptivist Cognitivism: Framework for a New Metaethic', *Philosophical Papers*, 29 (2000), 121–53.

Kalderon, Mark, 'Open Questions and the Manifest Image', *Philosophy and Phenomenological Research*, 68 (2004), 251–89.

Kripke, Saul, *Naming and Necessity* (Cambridge, Mass.: Harvard University Press, 1972).

____ *Wittgenstein on Rules and Private Language* (Cambridge, Mass.: Harvard University Press, 1982).

Lewis, David. K., 'New Work for a Theory of Universals', *Australasian Journal of Philosophy*, 61 (1983), 343–77.

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

Nozick, Robert, *Philosophical Explanations* (Cambridge, Mass.: Harvard University Press, 1981).

Ogden, C. K., and Richards, I. A., *The Meaning of Meaning* (New York: Harcourt Brace & Jovanovich, 1923).

Putnam, Hilary, 'Explanation and Reference', in G. Pearce and P. Maynard (eds.), *Conceptual Change* (Dordrecht: Reidel, 1973), 199–221.

—— 'The Meaning of Meaning', in K. Gunderson (ed.), *Language, Mind and Knowledge, Minnesota Studies in the Philosophy of Science*, vii (Minneapolis; University of Minnesota Press, 1975).

Salmon, Nathan, *Frege's Puzzle* (Atascadero, Calif.: Ridgeview, 1991).

Sayre-McCord, Geoffrey, '"Good" on Twin Earth', *Philosophical Issues*, 8 (1997), 267–92.

Smith, Michael, *The Moral Problem* (Cambridge: Blackwell, 1994).

—— 'Internal Reason', *Philosophy and Phenomenological Research*, 55 (1995), 109–31.

Soames, Scott, *Beyond Rigidity* (Oxford; Oxford University Press, 2002).

Sosa, Ernest, 'Water, Drink, and "Moral Kinds"', *Philosophical Issues*, 8 (1997), 304–12.

Stocker, Michael, 'Desiring the Bad', *Journal of Philosophy* (1979), 738–53.

Timmons, Mark, *Morality without Foundations* (Oxford; Oxford University Press, 1999).

van Roojen, Mark, 'Humean and Anti-Humean Internalism about Moral Judgements', *Philosophy and Phenomenological Research*, 65/1 (2002), 26–49.

—— 'Cognitivism vs. Non-Cognitivism', *Stanford Encyclopedia of Philosophy* (Spring 2004 edn.), ed. Edward N. Zalta: <http://plato.stanford.edu/archives/spr2004/entries/moral-cognitivism/>.

This page intentionally left blank

7

Moral Feelings and Moral Concepts

Allan Gibbard

In my 1990 book *Wise Choices, Apt Feelings*, I set out to analyze the narrowly moral concepts MORALLY WRONG and MORALLY REPREHENSIBLE.[1] Wrongness pertains to acts, to what a person does, whereas being blameworthy or reprehensible pertains to the act along with motives and the agent's state of mind.[2] A man in the pangs of grief, imagine, speaks woundingly to a friend who is trying to bring comfort. The act might well be wrong and yet, in light of his wrought-up state, he might not be acting reprehensibly. The concept REPREHENSIBLE I treated as the more basic of the two, defining the word 'wrong' in terms of an action's being reprehensible.

A chief part of my interest in these narrowly moral concepts was, of course, to try to understand what is at issue in questions of right and wrong and in questions of blameworthiness. Another motive, though, was to discern what's at issue in blanket attacks on "morality", as with Nietzsche and Williams.[3] Can we understand morality, in a narrow sense, as something we could be without, and that other cultures perhaps do lack, for better or worse?

Narrowly moral questions, I proposed, concern how to feel about things that people do or might do. For an act to be morally wrong isn't just for it to be inadvisable or against the demands of reason. Passing up an enjoyable

Versions of this paper were presented to the British Society for Ethical Theory, to the First Annual Metaethics Workshop, University of Wisconsin, Oct. 2004, and to audiences at Harvard University and at the Pacific Division meetings of the American Philosophical Association. I am grateful to many people for discussion on those occasions, and also, leading up to the preparation of the paper, to participants in the Seminar on Ethics, University of Michigan, Winter term 2004 and in the Workshop on Emotion, Evolution, and Ethics at Duke University and the University of North Carolina, Apr. 2004

[1] I use small caps to denote concepts. This is a variant of the convention that Horwich adopts (1998).

[2] Ross (1930) distinguishes "acts", apart from motives, from "actions", which are acts with their motives. Brandt (1959) discusses blameworthiness as distinct from wrongness.

[3] Nietzsche (1887); Williams (1985).

outing, for instance, might not be the reasonable thing to do, if no strong considerations tell against going, but still if no one else is let down, the act might not qualify as morally wrong and staying at home might in no way be reprehensible. Acting to enrich oneself by bilking large numbers of people out of their pensions, in contrast, is both wrong and reprehensible. That it is *reprehensible*, I proposed, means that feelings of resentment or outrage over it are warranted on the part of impartial onlookers and feelings of guilt over it are warranted on the part of the person who does it. ('Warranted' wasn't the term I then used, but it now strikes me as best for this purpose.) An act is *wrong*, roughly, if it is to be shunned as—barring some abnormal state of mind—reprehensible.

This last move, though it merits close discussion, won't fall within the scope of this paper. The chief point for my purposes here is this: narrowly moral questions, according to the analysis, are at root questions about what narrowly moral feelings are warranted regarding people and their actions. The narrowly moral sentiments are resentment or outrage, indignation, or reprehension on the part of others and guilt or remorse on the part of an agent. To abandon narrow morality, then, would be to abandon the practice of assessing these narrowly moral sentiments as warranted or not. That leaves scope for many other kinds of normative assessment, but not for a kind that looms large in our own mores. That, in brief, was my view in *Wise Choices, Apt Feelings*.

In this paper, I consider two objections to any such analysis. Both objections are empirical in their basis: the analysis, each objection runs, isn't true to the psychology of moral judgment. I'll therefore be citing psychological experiments and the conclusions psychologists draw from them, and asking what to make of their findings. One objection, due to Shaun Nichols, is that young children acquire their moral concepts before they come to understand guilt, so that the concept of guilt can't be a constituent of the concept of moral wrong. The second is that for children and many adults, it seems implausible that the concept of moral wrong could be constructed in the way I picture. Rather, people are implicit moral realists who, in effect, treat their moral emotions of detectors of a property. I concede a great deal to each of these objections, but end up rejecting this form of moral realism as a theory that can be vindicated and used to assess the truth of naive moral judgments.

Moral Realism and Plans for Feelings

I have been speaking here of "warrant" for feelings, and in *Wise Choices* I spoke of "how it makes sense to feel" about things people do. WARRANT

I regard as the basic normative concept, the conceptual atom that renders molecular concepts like REPREHENSIBLE normative. Much in the account of narrowly moral concepts that I have sketched must hinge on how this concept WARRANT is explained. But mostly, in the rest of this paper, I'll be pursuing questions that arise independently of how we gloss warrant. We could be non-naturalists about warrant, as Ewing was, or dispositionalists about warrant, as we might read philosophers in the ideal observer tradition, such as Firth, Brandt, and Michael Smith, as being.[4] I'll briefly sketch my own account of the concept WARRANT, though, for two reasons. First, it bears on the very point of moral evaluation: narrowly moral questions, if I'm right, are questions of how to feel about things people do or might do. Second, I'll later be asking whether the right view to take of morality is "realist". I need, then, to say in what ways my own view is a form of realism and in what ways it isn't. Joined with a realist account of the concept of WARRANT, the analysis of REPREHENSIBLE I have offered would come out as realist, whereas joined with an irrealist account of WARRANT, this analysis of REPREHENSIBLE would be irrealist—for it would have an irrealist component. As for how to classify my own analysis, the question is tricky; my analysis is what Blackburn calls "quasi-realist".[5]

Acts, beliefs, feelings—each of these sorts of things, it seems, can be warranted or not. We speak of the thing to do, the thing to believe, or the thing to feel. We can ask what it makes sense to do, what it makes sense to believe, and how it makes sense to feel about a person or an action. My account of what's at stake in such questions is this; begin with action: the basic normative question is what to do, and to come to a conclusion on this is to form a plan. On a trip, say, I ask whether it makes sense to explore the city this afternoon or to rest first. I come to an answer by forming a plan. The thing to do, I might conclude, is to rest first; to come to this conclusion is to plan, for this afternoon, first to rest.

The same goes for feelings, I argue. The question of how to feel about something is, in effect, a planning question. I ask myself, say, how to feel about lies told to criminal suspects during questioning. As I work toward an answer, I'm in effect forming a plan for how to feel in my circumstance. I can plan what to do, and in a way I can likewise plan what to believe given certain evidence: planning an experiment, statisticians stress, crucially includes plans for how to analyze the outcome, contingency plans for what to believe should the experimental data turn out in various ways it may turn out. The same goes, I claim, for feelings.

[4] Ewing (1939); Firth (1952); Brandt (1959); Smith (1994).

[5] Blackburn (1993); (1998). I adopt this term in *Thinking How to Live*.

This analysis is in essence the one I gave in *Wise Choices* in 1990, and in my new book *Thinking How to Live* (2003), it is developed this way explicitly. The new book takes further a set of claims I made in the previous one. A planning concept of the kind I analyze would, it argues, act in ways that might have been thought diagnostic of normative realism. We are guaranteed, for instance, that there's a non-normative property that constitutes being the thing to do. The theory starts out, then, with somewhat the same kinds of materials as does Ayer's emotivism—a paradigm, we might have though, of anti-realism in moral theory. My theory differs from Ayer's in important ways, to be sure, but the core of Ayer's theory could, I claim, be put in a form that has exactly the same implications for questions of moral realism as does my own theory. (Indeed that, in effect, is what Simon Blackburn does with his own form of "quasi-realism".) I don't end up denying that moral claims are true or false. I do end up claiming that there are moral properties, and that these are natural properties in a liberal sense of "natural". Indeed these properties could, for all my metanormative theory tells us, be properties subject to empirical investigation by psychologists and social scientists. In crucial ways, then, the theory ends up very much like the non-naturalistic moral realism of G. E. Moore—and it also endorses some of the central claims of naturalistic moral realists like Richard Boyd.

All this, however, is by way of *a priori* investigation into concepts. The new book isn't much about us, the flesh and blood organisms whose concepts these might be. In this present paper, I'll return to psychology, to the bearing of empirical investigations into our nature on metanormative questions, on questions of what our normative concepts might be.

Wrong and Guilt: Empirical Concerns

I'll begin with an objection to my theory by Sean Nichols, in his book *Sentimental Rules* (2003). My book *Wise Choices*, unlike the new book, was highly concerned with the actual human psyche and with narrowly moral concepts. Among other things, as I indicated, it analyzes the concept of wrong in terms of reprehension and guilt. This can't be right, Nichols argues, for young children acquire the narrowly moral concept of wrong long before they have the concept of guilt. Wrong they understand by the age of 4, but guilt not until they are 6. Their moral concepts at the start, then, can't be the ones I depict. That, he argues, discredits my analysis too as it applies to adults.[6] If, after all, our concept wrong isn't one that we

[6] Nichols (2004: 90–6).

share with 4-year-old children, then we aren't agreeing or disagreeing with them when we wield our own concept.

Similar grounds for worry about my analysis stem from anthropology. When I was writing *Wise Choices, Apt Feelings*, I was frustrated not to find much by way of reports on what people actually say to each other when they criticize, gossip, and the like. I likewise couldn't find much on the moral emotions and the like that figure in people's responses to each other. The situation is somewhat better by now, with some anthropologists asking very much the kinds of questions I want them to be asking. A lesson that may be emerging is this. On the one hand, moral outrage or something close to it is a human universal.[7] On the other hand, guilt, as European-rooted cultures know it, may be culturally special.[8] Again, we can ask, if some cultures lack guilt or lack a concept of guilt, are these people just using different concepts from ours? Does that mean that we can't straightforwardly agree or disagree with them when they make outrage-laden judgments?

Talk of guilt and its absence cries out for some explanation of what we should take guilt to be. How does guilt, as we know it and hate it,

[7] Moral outrage seems to be a species of anger. Fessler writes, "It is likely that anger is one of the most universally identifiable emotions .... Although notions of goals, rights, property, and even the definition of a person are culturally variable, howsoever these things are defined, when they are transgressed, people react with anger" (2005). Moral outrage may be disinterested, and it would be good to know whether disinterested anger is universal.

[8] Fessler and Haley (2003) find it an open question whether guilt is a human universal or the upshot of special cultural circumstances. They cite evidence that guilt can be highly effective in promoting cooperation. Still, they caution, "It is questionable as to whether guilt is experienced in all or nearly all societies" (16). Many languages have no term specifically for guilt, they note, and guilt has no uniform involuntary display. Fessler found that in a fishing village in Sumatra, "informants rarely discussed anything resembling guilt, frequently only providing accounts of *regret* (e.g., 'I wish I hadn't cheated because it caused so many problems')" (18–19). The Indonesian term *bersalah*, literally, to be in the wrong, is, they say, primarily associated with fear, prompting avoidance rather than reparations or self-punishment. Whether guilt is a human universal, they conclude, thus needs further study (17). Guilt they characterize in much the way I do here, citing some of the same work. "Guilt focuses attention on the action and the harm that has been done to the other party, inflicts subjective discomfort on the actor via its strongly aversive valence, and motivates the actor to make amends by aiding or otherwise compensating the victim .... The functioning of guilt is thus precisely tuned to identifying and reversing the damage done to a cooperative relationship" (16). Shame and pride do appear to be human universals (18–21); they tie these feelings to others' awareness that one has behaved in a blameworthy or commendable fashion (19). (Pride and shame as we understand it, I'd think, may be tied to a wider range of attributes: one's skill, one's house, one's country. One may take pride in a secret accomplishment or be ashamed of a secret vice. I'm not clear whether shame as Fessler and Haley discuss it is this emotion in a particular kind of context or a different but similar emotion. See also Fessler (1999) for an extensive treatment of pride and shame.)

differ from other kinds of anxiety, such as fear and shame in their various forms. Guilt-absent cultures are by no means free of anxiety; the point is that anxiety at having done something wrong may, in those societies, lack certain special earmarks of guilt. Writers take a range of positions on what these earmarks are, but here is my own rough view of the matter. [9] Guilt normally "meshes" with the resentment of others, in that it tends to arise when one would resent others if positions were reversed and one knew what they had done. The feeling doesn't require that others know what one has done—or, if they do know, that they do in fact resent it or may resent it. It doesn't, moreover, respond just to the bare fact that others do resent one. Rather, it is governed by the same standards as govern outrage or resentment, though from the standpoint of the agent rather than the one affected. In roughly the ways that resentment requires thinking an act voluntary, so does guilt.[10] Finally and crucially, guilt is characterized by its tendencies toward actions and displays that work toward reconciliation. One tends to display one's pain to one's victim and to bystanders, thus neutralizing any advantage one has gained at the victim's expense. Feeling guilty also prompts one, if possible, to find a way to "make it up" to the person wronged.

Guilt can be over lack of due care, but the paradigm is guilt over a deliberate action that one now regrets. In the paradigm case, guilt thus requires a "change of heart", and to display pangs of guilt to others is to display this change of heart. Take a child who pushes another child off a swing so that he can get the swing himself. He won't feel guilty for what he did unless he now regrets it, preferring to have left the child alone and forgone the swing. Even guilt over an accident may display a change of heart on how much care to take to avoid the harm. We can understand guilt, then, only if we understand changes of heart, coming to respect the interests of others in a way that one previously didn't. (Perhaps this is what the 5-year-old doesn't comprehend and the 6-year-old does. We can ask, is this scenario absent from the lore of some cultures?)

[9] My views are closely related to those of Baumeister *at al.* (1994) as well as to Fessler and Haley (2003).

[10] Neither resentment nor guilt invariably require thinking an act voluntary, but the exceptions correspond for the two emotions. "Survivor guilt" notoriously needn't stem from one's voluntary actions. Yet it seems aptly named: the "flavor" is that of guilt, it focuses on a kind of alienation from one's fellows, and it copes with a fact that may elicit their resentment. Guilt over actions of one's group also doesn't seem to require any act of one's own, and so not, in particular, a voluntary act of one's own. But that goes for resentment too, as in ethnic rampages where members of a group are attacked and often killed over acts of other members. Guilt and resentment may well not be warranted in those circumstances or may be misdirected, but these do seem to be recurring emotional patterns.

Guilt is closely tied to anxiety over social exclusion, over alienating those who are important to one.[11] But social exclusion will be disastrous anywhere, and so anxiety over alienating others must no doubt be a human universal. Guilt focuses not directly on the social danger, but on the ways one has come to warrant the resentment of others. Not that guilt isn't highly affected by the social perils of being resented, but with guilt one's emotional focus isn't on these dangers but on one's misdeed.

This emotional style of dealing with one's transgressions may not be found everywhere, and if it isn't, then we can ask what it is about cultures rooted in Europe that inculcates such a style. One evident answer is the practice of apology, often forced on the young child and found painful and humiliating. How, then, absent the pattern I have described, do people deal with their own transgressions and the people they wrong? We need to learn more about this from ethnographers; it strikes me as an important question for understanding what's universal in morality and what's local to certain kinds of cultures.

If a culture lacks guilt, Nichols can say, then it must lack narrowly moral concepts as I analyze them. Now in a way, of course, the conclusion that cultures may lack narrowly moral concepts might fit in nicely with the project of identifying morality in a narrow sense, of identifying a target for the critiques of Nietzsche and Williams. Both these philosophers were concerned with guilt (along with a hyperbolic sense of free will and responsibility that won't be my concern here). If parts of humanity manage without guilt, so much the better for the prospects of weaning us ourselves from guilt. But at the very least, for people with outrage but not guilt in their emotional repertories, we need to ask what ethical issues we still can discuss with them, and what concepts we can draw on to formulate common issues.

Outrage without Guilt

Nichols draws, in his critique, on two sets of findings in developmental moral psychology. First, as Turiel and others have found, quite young children make a conventional/non-conventional distinction.[12] Ask a child, "Suppose the teacher said that hitting was allowed in the classroom. Would it be all right to hit people?" Children by the age of 4 answer no. Ask

[11] Baumgartner *et al.* (1994: 246) speak of guilt as having "two sources: empathetic arousal and anxiety over social exclusion". I perhaps should have included empathetic arousal in my account, but I don't have any special reason to think empathy absent in guilt-absent cultures.

[12] Turiel (1983); Turiel *et al.* (1987).

her, “Suppose the teacher said that talking out of turn was allowed in the classroom. Would it be all right to talk without being called on?” She answers yes. Nichols calls this a “moral/conventional distinction”, and uses it to indicate that young children do have the concept of moral wrong.

This talk of a “moral–conventional” distinction is, I think, somewhat misleading. There’s a moral element, after all, to due authority. Classroom order, for example, serves the goals of being in class at all, and order often requires authority. It would be good to have systematic evidence of whether young children have a concept of legitimacy, a concept of authority as opposed to sheer domination. Will they give the same answers concerning the edicts of a playground bully or a kidnapper as they do the rules set down by a teacher or parent? (My own sense is that children at an early age stand ready to retort to an adult assertion of authority, “You’re not my daddy”, but I don’t know of any systematic investigation of the matter.) The word “conventional” strikes me as wrong for the contrast; it suggests things like the rules of table settings, which might have more to do with social displays of competence than with respecting interests and forwarding common purposes. The distinction Turiel finds isn’t between morality and non-moral convention, but between what is and what isn’t contingent on authority.

Still, Turiel’s findings do show that young children have a concept that shares crucial features with the adult concept of being morally wrong. Since they regard the wrongness of hitting people as authority-independent, it can’t be that as they conceive matters, being “wrong” just consists in being the kind of thing that brings penalties.

Nichols’s second set of findings concerns guilt. Before the age of 6, children don’t seem to understand feelings of guilt. Tell a younger child that Tommy deliberately pushed Jane off the swing, and the child expects Tommy to be happy.[13] Tommy, after all, has achieved what he set out to do. So if the child’s concept of wrong has anything to do with guilt, the child must be oblivious to the connection.

My analysis of narrowly moral concepts, then, can’t hold true for any concept that children under 6 have. I say that the basic narrowly moral concept is being blameworthy or reprehensible. That an action is reprehensible just means that reprehension over it on the part of others and guilt over it on one’s own part are warranted. But children younger than 6 have no concept of guilt, and so warrant for guilt can’t be any part of what they build their concept of wrong from. Likewise for adults from guilt-absent cultures, if there are such—and so, it seems, neither they nor young children could have the moral concepts I define in *Wise Choices, Apt Feelings*.

[13] Nunner-Winkler and Sodian (1988).

Perhaps that's all right, and just shows the peculiarity of our own institution of morality, for better or for worse. If I got our own narrowly moral concepts right, it seems to follow, then we can't discuss questions of narrowly moral right and wrong either with young children or with many exotic adults.

Still, there are closely related issues that we could discuss with them, for all these findings show. It hasn't been shown, after all, that either of these populations lacks a concept of outrage—and indeed it would be surprising if they did. We can share with a wide range of people, then, a set of pared-down narrowly moral concepts. I'll call these "near-moral" concepts. An action is outrageous, in this near-moral sense, just in case outrage over it is warranted on the part of others. Suppose, then, as may well be, that the concept of outrage is a human universal, and that the concept emerges fairly early in a child's development. Then, for all we have seen, the quasi-moral concept of being outrageous may be available to all adults and to all children at a fairly early age. All they need is a concept of reprehension or outrage and a concept of warrant, and they have the ingredients for a near-moral concept of being reprehensible.

For us, the quasi-moral concept of being outrageous might be exactly equivalent to the narrowly moral concept, in the following sense. Here is something that we may perhaps accept as truistic, as not in need of debate and discussion: guilt over something one has done is warranted just in case outrage over it is warranted on the part of impartial observers. If I relax with some light reading instead of grading a set of overdue papers, I'm warranted in feeling guilty over what I have done just in case an impartial spectator would be warranted in feeling outraged over it. Clearly this isn't something that a young child could accept, or an adult from a guilt-absent culture. But perhaps it is something that we in a guilt culture do take for granted.

Now if that is so, then on our own view, an action is outrageous in the narrowly moral sense just in case it is outrageous in the near-moral sense. We think that guilt is warranted just when outrage is—that guilt is warranted on the part of an agent just when outrage is warranted from impartial others. We are committed, then, to this claim: guilt and outrage are warranted in just those cases where outrage is warranted. In other words, an action is reprehensible in the narrowly moral sense just in case it is reprehensible in the quasi-moral sense.

Naive Concepts: Wrong and Red

For anything the guilt critique shows, young children might think in near-moral terms in the way I have pictured. We haven't seen that they lack the psychological concept of outrage or reprehension, and we haven't seen that they lack the normative concept of warrant. The outrage-warrant approach is to build near-moral concepts out of these two elements. Still, it does seem implausible that the goings-on in young children's heads, when they think in near-moral terms, are put together in the complex way the outrage-warrant analysis depicts. Something more direct may be going on. Indeed, the same may well be going on with most adults—even including us philosophers much of the time. I'll sketch my best guess for what this something might be, and then spend the rest of the paper on the upshot if this guess is right.

What do the thinking of young children, of guilt-absent adults, and of the rest of us much of the time have in common? Start again with young children. They, like the rest of us, have emotional reactions to things like hitting. (With adults, indeed, we now know which emotional centers of the brain are active when one thinks, say, of pushing a man in front of a moving trolley.[14]) Emotions, on the story that neurophysiologist Antonio Damasio tells, involve, among other things, perceptions of visceral and other bodily goings-on set off by certain kinds of thoughts.[15] (This part of the story is reminiscent of Hume, and psychologist Jerome Kagan's dustjacket praise on Damasio's first book read "Hume must be smiling.") The thought of one child hitting another, then, has a certain emotional "flavor", as we might put it; the flavor is a matter of what it's like to experience those bodily sensations and "bundle" them, as it were, into one's whole conception of what goes on.

As the word 'flavor' suggests, all this is, in a way, like experiencing a taste of ice cream that one can recognize—though the path from the hitting to the emotional flavor may be more circuitous, presumably, than that from the ice cream to the taste. Instead of talking about flavor literally, though, let me consider another analogy, color vision. Certain patterns of electromagnetic radiation lead to an experience of red. And on seeing a thing as red, we naively attribute a property to the object seen; we take it that the thing *is* red. How exactly all this should be put is of course a perennial philosophical issue, which I won't get much further into. My point here is that much of what can be said about seeing something as red and so thinking it red can be said in the moral case. Being outraged by an act is to thinking it wrong as seeing a thing as red is to thinking it red. That is the parallel I want to explore; it is familiar from Hume, Blackburn, ideal observer theorists, and others.

With color, just as with morals, as we all know, it is hotly debated whether the right story is a form of color realism, color irrealism, or error theory.[16]

[14] Greene *et al.* (2001).

[15] Damasio (1994).

[16] For other views, see Boghossian and Velleman (1991); Johnston (1992; 1996).

I argue for a combination of color realism with error: a sophisticated view of color experience allows that colors are real properties, but that naive thought misconceives them.[17] Naive users, though, do count as saying mostly true things about colors. The moral case, I'll argue, is analogous in some ways and not in others, and the upshot for moral concepts is further from realism than with colors. But as with colors, sophisticates can regard the naive as thinking mostly true things, while perhaps misconceiving what they are doing.

The child who sees a red pillar box as red normally judges it to be red. To accomplish this, note, the child doesn't need a concept of looking red. It's we who observe the child who say, rightly, that the pillar box looks red to the child and that this leads the child to think the box red. Soon, though, as the child matures, she can also be convinced that the occasional thing that looks red isn't really red. A familiar red jacket, for instance, under sodium vapor lights at night may look gray. All this goes for adults too, most of the time in daily life, and perhaps with some adults this is as far as color thinking gets. Young children, and even some adults, may lack an articulate account of what's going on, beyond that red things may not look red when seen in funny light. But we observers could attribute to them an implicit theory—and it's not a bad theory. We are detectors of a property that exists independently of us. We detect redness by seeing things as red. The detection isn't infallible, because sometimes our response isn't set off by the property, or is set off by things without the property. I'll call this the naive realist theory of color—not, as I say, because it need be articulated in our naive moments, but because it offers a rationale for our naive judgments and our ways of making those judgments.

In the case of morals, as a parallel might go, what corresponds to seeing as red is an emotional response, calm or more aroused. The naive young child feels outraged at one child's hitting another, say, and thereby thinks the hitting wrong. Being wrong the child then treats as a property, in that she reasons in terms of wrong in much the way she reasons in terms of a property like being square. At this stage, the child doesn't need a concept of outrage; it's we who say that in feeling outraged the child classes the act as wrong. Eventually, though, the child might be convinced that the occasional thing that seems wrong isn't really wrong, or that the occasional thing that seems all right is really wrong. The same act might seem wrong one hour or one day and not the next. Being wrong is to seeming wrong, then, in this more sophisticated child's implicit conception, as being red is to seeming red. Whereas seeming red is a matter of color experience,

[17] My own view, sketched here, is in my (1996).

seeming wrong is a matter of moral emotions, of feelings of outrage. This we can call naive moral realism. It need not be articulated in our naive moments, but it offers a rationale for our naive judgments. And again, perhaps it's not a bad theory.

What, then, of guilt? As we have seen, young children lack it, and so, perhaps, do many adults—guilt as a distinct kind of anxiety with its own distinctive properties. For those of us who have it, though, guilt might be another detector of the same property. Writers from guilt cultures indeed often talk of conscience as our prime way of knowing right from wrong, where conscience is the faculty of guilt or self-reproach. Naive moral realism might come, then, in two versions, one narrowly moral and the other quasi-moral. The quasi-moral version has one detector for wrongness, namely outrage, whereas the other has two detectors, outrage on the part of observers and guilt on the part of the agent.

Return, though, to the variant of implicit naive realism that invokes outrage alone. Both with red and with wrong, others too are good but fallible detectors, on this theory. The upshot is disagreement. Moral disagreement is perhaps more frequent and surely more important than disagreement over colors. If Jonathan Haidt is right, it's primarily interpersonal moral disagreement that prompts moral reasoning. Ordinary reasoning applies to wrongness, since wrongness acts as a property. That's not to say that reasoning often changes people's minds, at least when compelling emotions stand behind a judgment. We reason as advocates, not as seekers of the moral truth prone to follow the argument where it leads. Still, we treat reasoning as having the power to persuade—and occasionally it does.[18]

Mitigated Realism for Color

Should naive realism be our sophisticated view as well? Start again with colors. One naive impression is mistaken: color seems a basic property of things, the kind of salient property one would have to know about to have much comprehension of the real nature of the surface. It's a surprise, then, that color turns out to be a matter of how uncolored particles interact, and that color is of no explanatory interest apart from the peculiar way that we human beings are constructed. It's a surprise that a scientist from a planet of Alpha Centauri, if there were one, would rightly take no interest in colors and shades of color, as we delineate them, unless she took an interest in us. It's a surprise that for subtle questions of color, such as whether two green

[18] Haidt (2001).

thing are the same shade, there may be no fact of the matter; the two may look the same in direct sunlight but different in northern shadow, and if so, there is no intelligible fact of the matter whether they are the same shade or a different shade.

Ought we then to be error theorists about color? Some philosophers think we should be, and if what I have been saying is right, we must indeed attribute implicit mistakes to a naive view of color. But "error theory" as a term of art in philosophy means a theory according to which property ascriptions in a certain realm are systematically false. I point to a British red pillar box and ask the child, "What color is that?" "That's red," she answers. An error theory for color would be one that claims that the child is mistaken, that she is saying something false. But didn't she give the right answer to my question? Didn't she speak truth? I would say that she did. I would insist, moreover, we rightly take her to be speaking our language, as meaning RED by 'red'.

How should a sophisticated account of color concepts that yields these results go? We know that the straightest of realisms about color won't work. There's no fact of the matter precisely what classifications in the world we're making with our color concepts. There's no fact of the matter just whose responses in what kind of light are the test of whether two surfaces are the same shade of green. On the other hand, our visual systems do have remarkable properties of color constancy, so that broad aspects of how colors are perceived stay amazingly constant through a wide range of normal conditions. The upshot, I argue, is that color attributions are vague, but admissible resolutions of the vagueness are highly constrained. On any such resolution, colors, color shades, and the like are properties that surfaces really have. These are, as we might put it, objective, physical properties that are of human interest only. Centauris have no interest in the property of being crimson, and it's a somewhat vague matter what that property is. But in any resolution of the vagueness, it's a property of the electromagnetic reflectance of the surface, the degrees to which the surface reflects electromagnetic radiation of various wavelengths, a function from wavelength to proportion of light reflected.[19]

As for the child who hasn't thought about vagueness and implicitly treats color as a basic part of the nature of things, we should, as I say, regard her as speaking our language, with certain misconceptions. That's just a matter of the most charitable interpretation. On any admissible resolution

[19] The full story is of course more complex. Reflection from a surface, for instance, must be distinguished from scattering of light in the interior of the material, both of which can give rise to color experience. (I draw this, perhaps in garbled form, from a talk by Mark Johnston some years ago.) I'll skip over the refinements, however.

of vagueness, the pillar box is red. The child reasons in roughly the ways one should reason if thinking about colors and deploying the most sophisticated conceptions of color. As so interpreted, she gets most matters right and gets some of them wrong—and she might, if she thinks in certain directions, ask certain questions for which there is no clear right or wrong answer. Naive color realism is naive indeed, but it counts as a view about color as color. The right view of color, then, and of the child's color vocabulary, is what we might call a mitigated realism.

Moral Concepts

Can we tell the same mitigated realist story for morals? It might seem so. It has been perennially debated, to be sure, how variable moral judgments are from person to person and from culture to culture, and what the causes of variations are. Moral judgments at their best, it seems, somehow standardize our emotional responses. On the hypothesis that moral sentiments are error-prone detectors of properties—properties of interest to human beings in general or to members of particular cultural groups—we expect some variation in response. We come to views of what the conditions are under which the responses are most nearly constant. We use our responses under these conditions as fallible tests of theories of moral properties.

All this fits in well with a powerful empirical argument mounted by the social psychologist Jonathan Haidt. In his review article "The Emotional Dog and its Rational Tail", Haidt argues that emotional judgment is driven primarily by emotion. Argument matters, he agrees, but except for people who are specially trained as philosophers, argument comes only when the emotional intuitions of different people clash. Typically, each person is convinced that arguments support his own intuitions. "Moral reasoning is rarely the direct cause of moral judgment", he concludes, especially when people have strong feelings on a matter.[20] "Reasoned persuasion works not by providing logically compelling arguments but by triggering new affectively valenced intuitions in the listener." Persuasion, when it comes, rarely stems directly from moral reasoning. When people change their minds, morally, on questions they feel strongly about, this is far more likely to be the upshot of new intuitions that arise when the question is "reframed", seen in a new Gestalt.

Haidt's "emotionally valenced intuitions", it seems to me, are just what I call emotions or feelings or sentiments. Feelings are *about* circumstances

[20] Haidt (2001: 815).

and things, *over* matters, and directed *toward* people, things, and the like; they aren't just "raw feels" or sensations. As Antonio Damasio describes in *Descartes' Error*, emotions involve visceral sensations, somehow "bundled" into one's conception of a situation.[21] The term "emotion" might suggest high somatic arousal as a part of this syndrome, but I intend it to cover what Hume called "calm passions" as well as "violent passions".

We might, then, try refining naive conceptions of our knowledge of right and wrong into a mitigated realism, a sophisticated view of us as detectors of a vaguely defined property. As with colors, we can try saying, precisely which property is the property of being morally wrong will be a matter of some vagueness. It might be cause for surprise that right and wrong are of interest only to beings like us, and that the gods wouldn't care about these properties unless they were highly anthropomorphic—or even, perhaps, western European in their cultural roots. All this, however, just means that our realism about morals must be mitigated. It doesn't distinguish morals from colors. Is anything wrong with a story like this?

The problem, we expressivists will insist, is that it yields the wrong account of moral disagreement. My wife sees our wood-frame house as a grayish blue, whereas I see it as a bluish gray. Which is it? (Occasionally, in certain lights, I see it her way, and the experience is strikingly different from my usual one. Physical reflectances form a continuum, but color experience doesn't, I note in my own experience.) Well, there's no interesting answer, beyond that, I gather, more people see it her way than mine. We could sharpen our color concepts in various ways, on some of which my house is grayish blue and on some of which it is bluish gray. If she counts as using one such concept and I as another, we're not disagreeing. Do we count as having different concepts? There's no real fact of the matter; the issue of what color our house is lies in the penumbra of vagueness for color concepts. There's no clear matter at issue between my wife and me.

What, then, if we disagree on whether prostitution is always wrong? We might have different feelings about the matter and clashing moral convictions. Is this because there's nothing clear at issue? To say this seems to miss the point. At issue is whether to be against all prostitution, with a special emotional flavor of being against. It's an issue of whether to reprehend all prostitution—or whether to reprehend it if one takes an impartial standpoint of full emotional engagement with the matter.

In saying all this, I do have to allow for an error committed by common sense. We—or most people, in any case—are convinced that our strongest moral views are supported by irresistible arguments, and that if others

[21] Damasio (1994). I'm putting this in my own terms, and not making some of the distinctions that Damasio might regard as crucial.

would only listen to reason in good faith, they would be convinced. We also think that if the arguments, to our surprise, went the other way, we ourselves would listen to reason and change our minds. Jonathan Haidt and co-workers have experimented with this latter conviction, describing, for instance, a case of incest where none of the normally expectable harms turn out to ensue. His subjects are disturbed to find all of their arguments convincingly refuted, but they don't budge on whether the incest was wrong.[22] Haidt also describes the ordinary conviction that those who aren't convinced by one's arguments aren't in good faith.[23] If we attend to the evidence, then, we have to be error theorists in one sense: we have to accept that palpable errors infect our ordinary views of morality.

Must we be error theorists, though, in Mackie's sense? Should we conclude that ascriptions of wrongness, in the ordinary sense, are uniformly false? No more than with color, I'd argue. Just as with a color term like 'green', there's an account of what we could mean by 'wrong' that vindicates the core of our ordinary judgments and practices. In the case of color, the vindicating theory is a mitigated color realism that treats red and green as physical properties of surfaces, vaguely determined and of human interest only. It treats our visual systems and color experience as fallible detectors of this property. In the case of morals, I maintain, the story to tell is quite different, but there likewise does exist a vindicating story.

I'll say a few things to indicate how that story goes, but for any extensive development I must defer to my two books *Wise Choices, Apt Feelings* and *Thinking How to Live*. Questions of moral right and wrong, I propose, are at base questions of how to feel about things people do or might do. Similarly with questions of what one ought to do, in the sense of what one has most reason to do, all things considered: these are questions of what to do. Convictions on these questions consist in something like plans: plans for what to do and plans for how to feel. This view, I argue, though its starting points are like those of Ayer's emotivism—a paradigm anti-realist view of morality—ends up endorsing many tenets of moral realism. It ends up unclear, then, whether to classify the view as a form of moral realism or a form of moral irrealism. On the one hand, according to the theory, moral concepts are distinct from all naturalistic concepts: wrong can't be defined analytically in terms suitable for incorporation in psychology or social science. Moral beliefs can be understood as contingency plans of a certain kind, or by the ways they are logically tied to such contingency plans. Naturalistic beliefs can't be explained in such a way. On the other hand, anyone who plans what to do and how to feel is committed to a

[22] BjörKlund *et al.* (2000).

[23] Haidt (2001).

thesis of property identity: there is a property of being morally wrong. This is an ordinary property, not what Mackie would call a "queer" property. On any reasonable view, it is, in a broad sense, a natural property—though whether it is a property that should figure in the psycho-social sciences, that cuts our social nature "at its joints" for causal/explanatory purposes, isn't anything that metatheory can establish. These conclusions would all be accepted by clear moral realists like Sidgwick, Moore, and Ross, and by such current moral realists as Nagel, Dworkin, and Parfit. Whether that makes this metatheory a form of "moral realism" may then be no more than a matter of stipulation, a question of what the term "realism" is to mean.

Realism: How Morals and Colors Differ

Is what I have been saying just an error theory for morals in camouflage? Naively, I have agreed, we implicitly accept a view of ourselves as detectors of an objective property—and that view turns out to be untenable. A sharply different and defensible kind of view, I add, matches this naive view well enough that we can, speaking the sophisticated language, charitably interpret ordinary thought. This correct view has many of the earmarks of realism. In particular, moral wrongness, on the sophisticated view, is a property of an ordinary kind. It might, for all the very concept of moral wrongness tells us, be the property identified by a hedonistic rule utilitarianism. Still, the defensible view isn't the naive view, and so, it seems, the naive view must just be false. It's false, if what I've been saying is right, even if a defensible replacement mimics it pretty well. Does honesty demand a confession like this?

I think not. To see why, turn again to the case of color. The difference between red and wrong isn't that the naive view of color is mostly right and the naive view of morals is off base. It's a difference in how the naive view must be refined to be defensible. Both refinements involve surprises. Naively, there seems to be something genuinely at issue in the question of what shade of color a thing genuinely is, apart from how we happen to perceive it. Likewise for the question of who sees things correctly, and under what conditions. All this turns out to be an illusion. With morals, there correspondingly seems to be something genuinely at issue in the question of whether, say, all prostitution is wrong. This time, though, it's no illusion. The best sophisticated view of morals, if I am right, vindicates this as a genuine issue. The conviction that the issue is real is one we can't reasonably give up.

What in our conceptions and practices, then, drives the refinements apart? The two naive, implicit property-detector theories, for wrong and for red, seem parallel. Why would their defensible refinements diverge?

The difference stems from contrasts that I've so far glided past, contrasts in our naive conceptions and practices and the points we can find in them. Color responses and emotional responses differ crucially in that emotions are systematically valenced. To reprehend something is to oppose it, to be against it in one's feelings. Colors too have their valences, from time to time, but not in the same, systematic way. In consequence, the point we can find in thoughts of color is to classify things and to recognize the classifications using our eyes. These classifications are useful for a miscellany of purposes, and nothing much hinges on where exactly we regard the lines between them as lying. What shade of green to see this grass as is not a question of any systematic import. And moreover, how to see the grass isn't up to us. Feelings of reprehension, in contrast, do have a systematic import. To reprehend, as I say, is to be against. That's barring refinements, to be sure: in unusual circumstances, I might feel reprehension at something but not position myself against it. Perhaps I thought the matter over yesterday and discussed it with you, and we ended up not opposing it. But we can't, in any blanket way, divorce feelings of reprehension from being against. Then too, feelings respond to judgments in a way that color experience doesn't. We can ask ourselves how to feel about something, and the feelings themselves are somewhat responsive to the answer. So when we respond differently, say, to prostitution, there's a real question to ask: how to respond. There's something at stake in the answer to this question, namely what to be against.

We start out, then, with similar implicit theories of red and of wrong. At the most naive, the child just regards things as red or wrong, with firm conviction. We observers can note that what leads to these property attributions is, in both cases, the child's responses: seeing a thing as red in the one case, and feeling reprehension over an action in the other. Indeed at this stage, we'll see from our eventual, sophisticated standpoint, the child isn't making any systematic mistake. Most of the things the child thinks red really are red, and most of the things the child thinks wrong really are wrong.

To the theory of what's red and what's wrong, the child might then add the beginnings of an epistemology: I know that it's red because I see that it's red, and I know that it's wrong because I apprehend that it's wrong. (My guess is that children make this move with red but not with wrong; this talk of apprehension sounds not so much child-like as like W. D. Ross.) The property-detector theory this leads to is parallel in the two cases, red and wrong.

Divergence then comes when we ask what could vindicate naive practice. In neither case does a property-detector epistemology straightforwardly do the job. Color theory needs the realization that the precise boundaries are arbitrary, that there's no correspondence in how things are colored, apart from us, that isn't rooted in rough similarities of our color responses. Nothing is systematically at stake in where precisely colors and color shades are to be delineated. With emotional responses, though, something very much is at stake: what to be for and what to be against. Regarding a feeling as warranted or not matters, in that our feelings are responsive to these judgments, and thus what we're for and what we're against responds to these judgments. When responses differ, we can't take the upshot to be mere vagueness in proper classifications, to be resolved arbitrarily if they merit the bother of resolving them at all. We can't do this and follow our normal bent to use judgments of right and wrong as guides to conduct. Moral classification ties in systematically with what to be for and what against.

In both cases, the best vindication of the main lines of ordinary practice turns out to be different from the straight property-detector view that naive thought tends toward. In the case of red, it's a property-detector view mitigated by thoughts of vagueness, arbitrariness, and the like; in the case of wrong, I argue, it's a view that what's at stake in morals is how to feel about actions, and thus what to be for and what to be against. Still, though a naive epistemology is in error in both cases, that doesn't mean that judgments of red or of wrong are mostly false, or that ordinary modes of investigation are off base. In both cases, a sophisticated vindication of ordinary practice can be found. In both cases, naive thought fits in closely enough with a defensible explication that charity demands treating naive thought as mostly getting matters right.

People find the truth about color surprising. If a theory told us there's nothing to be surprised at, that we get along perfectly well in our ordinary thought without all this high-falutin science of color and philosophical analysis, that would discredit the theory. If I am right about the concept WRONG, then like remarks go for morality. The surprising truth, though, isn't that nothing is truly red or green—and it isn't that nothing is truly right or wrong.

REFERENCES

Baumeister, Roy, Stillwell, Arlene, and Heatherton, Todd (1994) 'Guilt: An Interpersonal Approach', *Psychological Bulletin*, 115/2: 143–267.

Björklund, F., Haidt, J., and Murphy, S. (2000) 'Moral Dumbfounding: When Intuition Finds no Reason', *Lund Psychological Reports 1*, 1–15 (Lund: Lund University).

Blackburn, Simon (1993) *Essays in Quasi-Realism* (New York: Oxford University Press).

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

Boghossian, Paul A., and Velleman, J. David (1989) 'Colour as a Secondary Quality', *Mind*, 98/389 (Jan.): 81–103.

____ and ____ (1991) 'Physicalist Theories of Color', *Philosophical Review*, 100/1 (Jan.): 67–106.

Brandt, Richard (1959) *Ethical Theory* (Englewood Cliffs, NJ: Prentice Hall).

Damasio, Antonio (1994) *Descartes' Error* (New York: Grosset/Putnam).

Ewing, A. C. (1939) 'A Suggested Non-Naturalistic Analysis of Good', *Mind*, 48: 1–22.

Fessler, Daniel M. T. (1999) 'Toward an Understanding of the Universality of Second Order Emotions', in Alexander Hinton (ed.), *Biocultural Approaches to the Emotions* (Cambridge: Cambridge University Press), 75–116.

____ (2005) 'The Male Flash of Anger', in J. Barkow (ed.), *Missing the Revolution: Darwinism for Social Scientists* (New York: Oxford University Press).

____ and Haley, Kevin (2003) 'The Strategy of Affect: Emotions in Human Cooperation', in P. Hammerstein (ed.), *Genetic and Cultural Evolution of Cooperation* (Cambridge, Mass.: MIT Press).

Firth, Roderick (1952) 'Ethical Absolutism and the Ideal Observer', *Philosophy and Phenomenological Research*, 12: 317–45.

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

____ (1996) 'Visible Properties of Human Interest Only', in Enrique Villanueva (ed.), Philosophical Issues, vii. *Perception* (Atascadero, Calif.: Ridgeview).

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

Greene, Joshua, Sommerville, R. B., Nystrom, L. E., Darley, J. M., and Cohen, J. D. (2001) 'An fMRI Investigation of Emotional Engagement in Moral Judgment', *Science*, 293 (14 Sept.): 2105–8.

Haidt, Jonathan (2001) 'The Emotional Dog and its Rational Tail: A Social Intuitionist Approach to Moral Judgments', *Psychological Review*, 108/4: 814–34.

____ Bjorklund, F., and Murphy, S. (in preparation) 'Moral Dumbfounding: When Intuition Finds No Reason', University of Virginia.

Horwich, Paul (1998) *Meaning* (Oxford: Clarendon Press).

Johnston, Mark (1992) 'How to Speak of the Colors', *Philosophical Studies*, 68/3 (Dec): 221–63.

____ (1996) 'Is the External World Invisible?', in Enrique Villanueva (ed.), Philosophical Issues, vii. *Perception* (Atascadero, Calif.: Ridgeview).

Keltner, Dacher, and Buswell, Brenda (1996) 'Evidence for the Distinctness of Embarrassment, Shame, and Guilt: A Study of Recalled Antecedents and Facial Expressions of Emotion', *Cognition and Emotion*, 10/2: 155–71.

Nichols, Shaun (2004) *Sentimental Rules: On the Natural Foundations of Moral Judgment* (Oxford: Oxford University Press).

Nietzsche, Friedrich (1887) *Zur Genealogie der Moral.* Tr. by Walter Kaufmann and R. J. Hollingdale as *On the Genealogy of Morals* (New York: Vintage Books, 1967).

Nunner-Winkler, Gertrude, and Sodian, Beate (1988) 'Children's Understanding of Moral Emotions', *Child Development*, 59: 1323–38.

Ross, W. D. (1930) *The Right and the Good* (Oxford: Clarendon Press).

Smith, Michael (1994) *The Moral Problem* (Oxford: Blackwell).

Turiel, Elliot (1983) *The Development of Social Knowledge: Morality and Convention* (Cambridge: Cambridge University Press).

____ Killin, M., and Helwig, C. (1987) 'Morality: Its Structure, Functions, and Vagaries', in J. Kagan and S. Lamb (eds.), *The Emergence of Morality in Young Children* (Chicago: University of Chicago Press), 155–244.

Williams, Bernard (1985) *Ethics and the Limits of Philosophy* (Cambridge, Mass.: Harvard University Press).

This page intentionally left blank

8

Negation for Expressivists: A Collection of Problems with a Suggestion for their Solution

James Dreier

1. The Negation Problem

Crucial to the solution of various problems for expressivism is the development of a coherent semantics for negation. Nicholas Unwin (2001) explained why Allan Gibbard's program in *Wise Choices, Apt Feelings* was not up to the task. Briefly, the problem is that there are three ways to 'negate' a sentence like

(M) Miss Manners thinks one must write thank you notes by hand.

The three are:

(N_1) Miss Manners believes one must *not* write thank you notes by hand.

(N_2) Miss Manners believes it is *not* so that one must write thank you notes by hand.

(N_3) Miss Manners does *not* believe one must write thank you notes by hand.

But the resources available in Gibbard's scheme seem to allow the construction of only two, most plausibly (N_1) and (N_3), with no way to interpret (N_2). From a logical point of view, we can think of requirement and permission as operators on sentences (or on infinitives or participles). The

This paper benefited enormously from discussion at the Madison Metaethics Workshop, Oct. 2004, and also from audiences at the University of Maryland and Princeton University. I thank especially Jim Pryor for good counter-examples and Nick Unwin for some emailed comments on an early draft.

operators are duals, so that to define one operator in terms of the other we need not just internal negation, but external as well. Compare the problem of defining possibility in terms of necessity: we can say that $\Box A$ is defined as $\neg\Diamond\neg A$, but only if we already know what it means to negate the possibility operator. If, on the other hand, we know independently what necessity and possibility are, then we can define external negation: $\neg\Box A$ is defined as $\Diamond\neg A$ and $\neg\Diamond A$ is defined as $\Box\neg A$.

The present problem is similar. We can use $\mathcal{R}$ and $\mathcal{P}$ as operators for requirement and permission, and then note that $\mathcal{R}A$ is (defined as) $\neg\mathcal{P}\neg A$; more to the point, $\neg\mathcal{R}A$ is (defined as) $\mathcal{P}\neg A$ and $\neg\mathcal{P}A$ is (defined as) $\mathcal{R}\neg A$. Here the externally negated sentences are explained in terms of other sentences whose negations are all internal; that is to say, the negation signs occur only *inside* the operators, and no operators are negated. These dual definitions are useful. Each sentence with a primary operator (an operator with scope over the rest of the sentence) expresses an attitude, expressivists tell us. The nature of the operator gives the nature of the attitude. If Miss Manners tells you that you must write thank you notes by hand, her sentence ("You must write thank you notes by hand") expresses an attitude, signaled by 'must', toward writing thank you notes by hand. If she tells you that you must *not* write thank you notes by hand, then she is expressing that same 'must' attitude again toward *not* writing thank you notes by hand. The negation is internal, so no new explanation is necessary. But what state of mind is expressed by the statement "It is not so that one must write thank you notes by hand"? If she said that, Miss Manners would be giving voice to the belief described by (N_2). It's all very well to say, as one might, that Miss Manners expresses the negation of the 'must' attitude (the attitude of 'requiring', let's say) toward the same kind of action (writing thank you notes by hand). But how are attitudes negated? Expressivists need no special story about how to negate propositions, or actions, since those negations are not specific to expressivism. But there is no ordinary language or commonsense notion of the negation of an attitude.

It is tempting to think that the negation of an attitude might be the absence of that attitude. Compare belief: besides the 'internal negation' that takes a belief that A to a belief that $\neg A$, there is also the external kind taking belief that A to *failing to believe that A*. Could the negation of *requiring* something be the attitude of *not requiring* it? Not in our context. True, (N_3) is the negation of (M). It is the real negation, the contradictory of (M). But (N_3) is not (N_2), which is what we wanted.

I said that the dual definitions are useful; here's how. If we know what attitude is expressed by '$\mathcal{R}$' (call it *requirement*) and we know what attitude is expressed by '$\mathcal{P}$' (call it *toleration*), then an account of 'negated attitudes' follows naturally. The negation of a tolerant attitude toward A, so to speak, will be the attitude of requiring ¬A, and conversely the negation of requiring A is tolerating ¬A. So if we can just help ourselves to the two attitudes, negation falls out. Since those two attitudes are perfectly ordinary, commonsense attitudes, one approach would be simply to take them as primitive.[1]

But in fact, this approach is problematic. It leaves mysterious why the two attitudes, toleration and requirement, are related to one another logically. Why is there any incoherence in tolerating something and also requiring its contradictory? That is what we are supposed to explain. It's no good just to posit that they are incoherent. How is it, anyway, that attitudes are logically related? As Bob Hale (1993) asked, *can there be a logic of attitudes*? This question, I think, is the same as the question of negation.

2. Disagreeing with Plans

In *Thinking How to Live*, Gibbard introduces a revised approach to norm expressivism. The new view differs from the old in a number of ways, although the core ideas are all still in place. Central to the new approach is the development of the idea of *disagreement*, which Gibbard thinks is the key to all sorts of problems. Among these problems is Unwin's problem of negation. If expressivists can appeal to a notion of disagreeing with someone's attitude (maybe one's own), then negation flows naturally into the scheme. How? Rather than take toleration and requirement as primitive attitudes, we can think of tolerating ¬A as disagreeing with requiring A. External negation corresponds to disagreement. The attitude expressed by the external negation of a sentence (with a primary operator) is disagreement with the attitude expressed by the sentence. In this way, the attitudes are related to each other structurally, as negations ought to be, and the factor by which they are related, disagreement, looks to have the right features to correspond to negation.

Suppose, then, that *you* judge that one must write thank you notes by hand. Your state of mind, according to Gibbard, is one of planning to write thank you notes by hand (if the occasion arises). Miss Manners disagrees. (N_2) says that she does. Her state of mind is: disagreeing with your plan. One way she might disagree, of course, is by definitely planning not to write thank you notes by hand, in which case (N_1) is also true, but another way is for her to plan indiscriminately to write thank you notes by hand, by email, by dictation, in which case (N_1) is not true but (N_2) still is.

[1] I think this is what Simon Blackburn (1988) had in mind.

Because she disagrees with you, (N3) is also true, but it could also be true if (*per impossibile!)* Miss Manners had no plans whatsoever involving thank you notes.

Is Gibbard really entitled to appeal to a sufficiently robust notion of disagreement? The suspicion is that the notion of a state of mind with which one might coherently agree or disagree is a notion that has to be explained. Of course, it is quite intuitive. When you believe in God and I do not, we disagree; when you have a headache and I do not, we do not disagree; when I hope the Yankees win and you hope they lose, we disagree; when you recognize a Schumann lied by the style and I do not, we don't disagree. It is quite intuitive, too, that when you plan always to write thank you notes by hand and I plan to write them indiscriminately by hand and by email, we disagree in plan. Supporting this intuition is the further intuition that if I change from my former, indiscriminate plan to your stricter one, I have *changed my mind*, whereas when your headache goes away you have not changed your mind. In *Thinking How to Live*, Gibbard simply takes the notion of disagreeing with a state of mind (and the notion of a state of mind with which one might disagree) as a primitive notion. But this is worrisome. It's not that absent an account of disagreement we should worry that there is no such thing. Rather, the worry is that if we have no explanation for why some states can be disagreed with and others cannot, then for all we know it could be that the correct explanation is not amenable to expressivism. For instance, it could be (for all we know) that the states with which it makes sense to disagree are the ones that are really representations of independent fact. No expressivist could then let his account of negation be grounded in an appeal to the idea of disagreement in attitude.[2]

3. Hyperplans and Completeness

In *Thinking How to Live*, Gibbard gives us a new device: a *hyperplan*, namely, a fully specified plan for every conceivable contingency. Hyperplans are complete. They are analogous to possible worlds, which are completely

[2] Why not? I think this is a complicated matter. Here's the short version. A sophisticated expressivist might be happy to agree that whenever one attitude disagrees with another, the two attitudes are representations of contrary contents. But no expressivist should be happy using this way of talking in an explanation. When you and I disagree in plan, then we disagree about what one ought to do; no problem there. But contents like *one ought to* φ have to be explained without appeal to *ought*-laden facts or properties. So the final explanation of disagreement in (normative) attitude cannot be in terms of the representational contents of the attitudes. The direction of explanation must be the other way around.

specific ways that the world might be. Your opinions can be represented by a set of possible worlds, and the limiting case of such a set is a single world. Someone who was unimaginably opinionated would represent the world as being exactly this way or that, as being precisely this possible world or that one. Similarly, someone whose plans were unimaginably detailed, perfectly so, would have a hyperplan. Since hyperplans are complete, every act whose contradictory is not required by a given hyperplan is thereby permitted. So, in the context of hyperplans, permission is defined in terms of requirement, without appeal to external negation. External negation, then, can be defined in terms of permission and requirement. Hyperplans solve the problem of primitive disagreement and so the problem of external negation. How exactly do they manage to do this? By being complete.

Earlier we rejected the idea of defining external negation by appeal to absence of an attitude. The point was that (N_3) says that Miss Manners lacks the requirement attitude toward writing thank you notes by hand, but that doesn't mean that she tolerates writing thank you notes by hand, since she may have no view about the question one way or the other, so the absence of requirement can't be what (N_2) attributes. But there are no hyperplans that 'have no view' about a question. So a hyperplan that fails to require A a fortiori permits ¬A.

Now appeal to completeness of formal objects, or of the states of an idealized agent, is tricky in this context, as Unwin (2001) explains in discussing Gibbard (1990). We have to ask, what is the difference between a system that is silent on the question of thank you notes and a system that permits all sorts of ways of writing them? This question is a bit hard to grasp. It helps if we keep our down-to-earth, unidealized, human perspective in mind. First, following Unwin, a traffic code is silent about thank you notes but does not say that email thank you notes are permissible, so silence and permission aren't the same. Second, the important question is not primarily about a formalism, but rather about people and what they are doing when they think this or that, what attitude they are expressing when they say one thing or another. For us theorists to speak of complete formal systems or plans, then, is not enough: we are supposed to say systematically how to connect a state of mind to a normative sentence. Formal objects like normative systems can help with bookkeeping, but they do not, in themselves, tell us about states of mind.

An example will illustrate. Suppose Mr. Manners has a carefully considered and quite permissive view about thank you notes: any old way of writing them, he thinks, is permissible. Officer Lopez, on the other hand, has no views about thank you notes at all. Her firmly held normative views are about behavior in traffic. Lopez and Manners are alike in this way: each lacks the requiring attitude toward hand-written notes, toward emailed notes, toward engraved notes, and so forth. Our formalism is capable of the intuitively correct representation of their normative attitudes, of course. We want to represent Mr. Manners's attitudes by the collection of all complete normative systems that permit each type of thank you note (really we want a proper subset, since he'll have lots of other normative attitudes that 'rule out' many, but for all that's been said so far the full set will do). Officer Lopez's state will include all of those but also the complete systems that forbid emailed notes, those that require engraved notes, and so forth. Officer Lopez is undecided, and so she has not ruled out extremely strict systems of manners, as Mr. Manners has. But what is it about the states of mind of Lopez and Manners as they are that tells us which collection of formal objects properly models each of their states?

Think of it this way. It does not follow from a plan's failing to forbid emailing that the plan permits emailing. That permission *does* follow from the plan's failing to forbid emailing, plus the completeness of the plan. So permissions are fully defined in terms of restrictions, in the context of hyperplans. But what fact about a person's state of mind does the fact of completeness of the formal object correspond to?

I'll first explain how a certain solution might work. The idea will turn on the fact that planning is more 'decisive' than having attitudes might be. Decisiveness of a state is what completeness of plans (considered as formal objects) models. But then I'll say how the new model runs into a 'problem of no mere permissions', where 'mere permissions' arise in cases in which some things are permitted without also being required. The new model seems not to allow for mere permissions. Next, I'll explain how Gibbard makes room for some kinds of mere permissions, but not others. My tentative verdict will be that the missing permissions are ones we can live without. Finally, though, I'll argue that the same old problem returns in a new guise, namely, as the "no preference problem". I have a suggestion for how to solve this problem. If my suggestion works, it makes expressivism negation-safe. But I'm not sure that it does work.

4. The Interpretation of Hyperplans and the Problem of Mere Permissions

Here is how I understand hyperplans. Each hyperplan delivers from each circumstance a particular thing to do. If there is nothing I plan to do in a certain circumstance, then my plans are in that respect *incomplete*; there is no question of my having a very wishy-washy plan that explicitly declines to rule anything out. In this respect, planning is by its nature a 'decisive' activity, as having attitudes is not. A collection of 'requiring' attitudes is not unfinished, we might say, merely because it contains no requirement toward either of a pair of contradictories; a plan *is* unfinished if it contains no plan to do anything in a given circumstance. This means that the indeterminacy between permission and indecision is driven out of the hyperplan model. Planning and generic normative attitude differ in this way. When I do not require one action or another, I might be undecided, but I might be tolerant. If I do not plan to write thank you notes but neither do I plan *not* to write any, I am thereby undecided.

Something is forbidden, in the planning model, when an alternative is planned, and permitted when a complete (hyper)plan doesn't forbid it. This means that in the formalism (as I've given it so far), everything is required or forbidden; there are no (mere) permissions. Once we see this, it is clear why the problem of distinguishing the noncommittal (Officer Lopez) from the tolerant (Mr. Manners) is dissolved: one of the possibilities is eliminated altogether.

Now on the one hand, the eradication of mere permission seems entirely in keeping with the spirit of the planning model, but on the other hand it looks like a problem in its own right. Let's call it "the problem of no mere permissions". It's a problem because we surely want a way to make sense of the normative judgment that a certain course of action is permitted but not required. It is in the spirit of the planning model, because when I am *decided* my plans just say what to do. They do not say what I am permitted to do except insofar as what is permitted is also required. Actually, this is not *quite* true. It suggests a complication to the planning model. The complication will reduce, but not eliminate, the problem of no mere permissions.

I am about to go off on some errands. I will stop and get a coffee, and then I'll go to the supermarket and pick up some dishwashing liquid. Right now, I haven't decided which kind of coffee to get. I'll either get a venti iced decaf Americano, or else a triple grande decaf latte. I just can't make up my mind yet, because it's not a hot day but it's very humid. I'll decide later. I also haven't decided which kind of dishwashing liquid to get, the yellow or the blue kind. That's because I don't think there is any difference between the two kinds except for the color, and I don't care which color I get. Now in each case, one might say that my plans are incomplete, but to my mind the flavor of the incompleteness is very different. In the case of the dishwashing liquid, I won't plan any further, but just grab a bottle when the time comes. It just doesn't matter. One might perfectly well say that my plan is just to grab one, and that I do not plan to grab any particular one is not a hole in my plan. (Compare: I plan to brush my teeth tonight, but I do not plan to start brushing left to right and I do not plan to start brushing right to left.) In the case of the coffee, I am going to make up my mind, or in any case try, because it *does* matter which kind I get. Talking to myself in normative language I might say, "Either the venti iced decaf Americano, or else a triple decaf latte is what I ought to get, but I'm not yet sure which; on the other hand, either color of dishwashing liquid is ok." This difference, between dishwashing incompleteness and coffee incompleteness, is easy to accommodate in the planning model; no surprise since I introduced it from within the planning model. When I just grab the yellow bottle of detergent, I haven't really made up my mind in favor of yellow: I could have picked a blue one without any change of mind. Of course, someone *might* be the same way with respect to coffee, but not me! Here's one way to capture the difference: between colors of dishwashing liquid I am indifferent, and my indifference is adequately expressed in a plan that counts buying either as 'the thing to do', whereas between coffee types I am rather undecided, and can't make up my mind yet which is 'the drink to buy'.[3]

In short, plans, even hyperplans, have room for ties. Options that are tied (for first!) are not severally required, but none of them is forbidden, either. They are therefore each merely permitted. The difference between mere permission and indecision is underwritten, in the planning model, by the difference between the attitude of indifference and the failure to have any worked out preference at all.[4]

5. Hyperplans and the Problem of Supererogation

The model of hyperplans, then, does allow some room for mere permissions. It doesn't allow all the room one might want, though. For there are some mere permissions that are not intuitively accounted for by the relation of ties, or the attitude of indifference. The clearest cases of resistant mere permissions are cases of supererogation. Suppose Carl is on his way to a basketball game, one that he's been looking forward to for months. As he sits on the bus looking out the window, he sees a woman fumbling with some of her belongings, and he watches as she drops her wallet, collects

[3] Compare Gibbard (2003: 45) where Sherlock Holmes has weighed up options and decided that either of two options will do, and when the time comes he opts for one of them (packing): "Packing from preference is different from plumping for packing out of indifference."

[4] There is some room for worry here. It is perfectly clear that there *is* a difference between being indifferent between a pair of options and having failed to form a definite preference at all. Indifference is as definite as preference and entirely different from indecision, as the contrast between the dishwashing liquid case and the coffee case shows. What is somewhat worrisome is the possibility that preference and indifference might ultimately need an explanation in terms of the agent's beliefs about what is better than what. Some of John Broome's work suggests this possibility; see esp. Broome (1993).

the other items she had out and, oblivious, walks on. He cries out, but she doesn't hear him. The bus is at a stop. Carl faces a choice. He could stay on the bus and shout again as it drives by the woman, but that is not very likely to work. Or he could instead sprint out the door, pick up the wallet, and give it to her, but if he does that he will miss the first quarter of the basketball game. What *ought* Carl to do? What do you plan to do in such a circumstance?

Here is what I think we intuitively want to say. Carl is *permitted* to carry out the first plan. Sprinting off the bus is not required. He is, certainly, also *permitted* to carry out the supererogatory second plan: he does nothing *wrong* or *forbidden* by sprinting off the bus and missing the first quarter of the game. But it does not seem quite right to assimilate Carl's example to the example of dishwashing liquid. It's not that it doesn't matter which act Carl performs. It's not that in forming our own plan we find ourselves indifferent between the options. The supererogatory act and the merely obligatory one are not, intuitively, *tied* for first place. The supererogatory act is *better*. It is *preferable*. In my finest planning moments, I plan definitely to save the woman's wallet and, bitterly, miss a quarter of my beloved basketball game, should I find myself in Carl's shoes. But then what to make of the judgment that sitting tight and shouting another time or two is *permissible*?

As I am understanding the planning model, it has room for some mere permissions but no room for the idea of the supererogatory. The model therefore imposes some restrictions on the sorts of fundamental normative outlooks that a person can have. If planning just *is* what's behind normative judgment, then there is no coherent judgment of the form, "A and B are each permissible, though A is definitely better than B." This seems to be a cost.

It may not be much of a cost, though. The example of Carl is deceptive, in a way, because it has a distinctly moral cast to it, as indeed all of the standard examples of supererogation have: supererogation, after all, is an ethical concept. But Gibbard's norm expressivism is in the first instance about the most fundamental category of to-be-doneness, and only derivatively about *moral* norms. In *Wise Choices*, Gibbard called the fundamental category 'rationality' and gave a further, important but detachable, story about how *moral* judgment works. In *Thinking How* Gibbard brackets the question of whether the most fundamental category involved in judgments of what to do is the category of *rationality*, and indeed he's noncommittal about whether the concepts he analyzes are really the ones expressed by the *ought*s of ordinary English, and there is not much mention of morality. Now supererogation is deeply engrained in commonsense thinking about morality, but it is not obviously part of commonsense thinking about what to do, all things considered. Indeed, the coherence of the idea of something's being worse-but-permissible, just from the point of view of rationality, is hotly disputed.[5] So a model of normative discourse that rules out *rational* supererogation, or *final question of what to do* supererogation, may not thereby incur much of a cost.

Ethical supererogation, as I said, is a different story, but it may be that the general sort of approach Gibbard takes to moral discourse is perfectly able to handle supererogation. Judgments of what is morally permissible and what is morally best, for instance, might be judgments of *what to feel*, and moral sentiments might fit together in a structured way.[6] In any case, the plausibility of commonsense *ethical* supererogation is not, for all that's been said, a problem for norm expressivism.

6. The Negation Problem Sneaks Back In

I am sorry to say that, even if I am right that we can live without rational supererogation, the resulting model allows the old negation problem to slip in the back door again. Plans order alternatives; by expelling mere permissions they answer the question of how to distinguish permissions from indecision. Ties, introduced as a way of allowing mere permissions back into the picture, spoil the answer by wrecking the tight relation between permissions and requirements.

Normative claims express planning states, as Gibbard says. Planning states, as I understand them, are something like intentions and something like preferences.[7] Let's look again at our original negations and see what planning states make them true.

(N_1) Miss Manners believes one must *not* write thank you notes by hand.
(N_2) Miss Manners believes it is *not* so that one must write thank you notes by hand.
(N_3) Miss Manners does *not* believe one must write thank you notes by hand.

The first is true iff Miss Manners plans *not* to write thank you notes by hand. Think in terms of preference: Miss Manners prefers *not* writing thank you

[5] I'm thinking of the 'satisficing' literature. See esp. Dreier (2004).

[6] See for instance Gibbard's contribution to this volume.

[7] Gibbard's planning states are not exactly like intentions in the ordinary sense because we can 'plan', as Gibbard uses the word, for circumstances in which we know we will never find ourselves, and we can 'plan' to do things even when we know our 'planning' will not bring about the things we 'plan'. I am not sure exactly how Gibbard's 'planning' differs from preferring.

notes by hand to writing them by hand, according to (N_1). (N_2) attributes to Miss Manners a logically weaker view: it is true iff Miss Manners does *not* plan to write thank you notes by hand, that is, so long as she prefers writing them by hand or is indifferent between writing them by hand and not. This state is logically weaker because it leaves open the possibility that Miss Manners is indifferent among the ways of writing thank you notes, so that, according to her, *not* writing thank you notes by hand is permissible but so is writing them by hand. That possibility is ruled out by (the view attributed to Miss Manners by) (N_1). Finally, (N_3) is about what plan or preference Miss Manners lacks: she does not plan or prefer to write thank you notes by hand. This last is, properly, consistent with Miss Manners having no view at all about thank you notes, and is not, of course, the same as her being indifferent among the ways of writing them.

But *why* is it not the same? That's (what I'll call) the "no preference problem". There is a distinction, I said, between indifference and indecision, the two kinds of "no preference", and it is an intuitive difference, as shown by the contrast between dishwashing liquid (indifference) and coffee (indecision). But how can this difference be made out in an expressivist framework? If you ask me on my way to the market which sort of liquid I prefer, I'll say I'm indifferent; if you ask me which kind of coffee I want, I'll say I haven't decided. But in drawing the distinction, might I be implicitly relying on a more *realist* metaethics that I'm entitled to? I worry that I am.

At the end of section 3, I noted that permissions are defined by the complete facts of requirements *plus* the fact of completeness. Those things are permitted whose complements are not required by complete plans. Then we can define external negation by the duality laws. But this trick works only if we have an independent account of completeness. The model of plans that allow no ties gives us such an independent account, but the model of plans with ties lets it slip away. To see this, imagine that we are interviewing Bob about his preferences. He can tell us when he prefers A to B. Sometimes, when we ask him about a pair of alternatives, he says that he prefers neither. How can we tell whether his lack of preference is the indifference kind or the indecision kind? My question isn't really epistemological. I assume that if we could get our hands on *what fact* it is that determines which way he lacks a preference, then in principle we could *find out* about that fact. In any case, it seems to me, this last question is the important one. If it can be answered, and answered in an expressivist-respectable way, then the problem of negation is solved, by way of the problem of completeness. But it is not clear that it can be answered in an expressivist-respectable way. Maybe the best answer is that the absence of preference is indifference just in case Bob *believes that the alternatives are equally good*, and the absence of preference amounts to indecision just in case Bob *has no belief about which* *alternative is better*. Myself, I suspect the explanation goes the other way around. Bob's beliefs about which alternatives are better are accounted for by Bob's preferences. But in that case, the difference between indifference and lack of preference must be explicable without the help of beliefs about what is better than what.

7. Defined Indifference: A Suggestion

My suggestion for a solution to the no preference problem starts with a piece of technical apparatus, which I'll call *defined indifference*. Before I give the definition, here are a few words of explanation. First, I am not claiming to be defining the ordinary language use of 'indifferent'. Instead I am introducing a technical notion to be used in the expressivist account of negation. I do think the defined notion comes pretty close to capturing *one* sense of 'indifference', at least in that it is close to being necessarily coextensive with the ordinary notion. Second, I will use preference, but not indifference, in the definition. I am taking preference for granted. My problem is to get indifference (or a stand-in) distinguished from the indecisive form of lack of preference, and for this purpose it is acceptable to take preference itself, the definite kind of preference, as a primitive.

> Someone is *defined indifferent* between options A and B iff (i) she does not prefer A to B or B to A, and (ii) for any X, she prefers A to X iff she prefers B to X and prefers X to A iff she prefers X to B.

I want it to turn out that defined indifference has the right logical properties to play the role in the expressivist account of negation that indifference itself was going to play. In the first place, defined indifference is an equivalence relation. Indifference ought to be an equivalence relation, too. But lack of preference is not, because it is not transitive. (I am calling 'lack of preference' the union, or disjunction, of genuine indifference and proper indecision. Proper indecision is plainly not an equivalence relation, because it is not reflexive; nor is it transitive.) I may be undecided between buying an iced Americano for $2 and buying a latte for $3, and also undecided between buying the latte for $3 and the Americano for $2.10, but I definitely prefer buying the Americano for $2 to buying it for $2.10. That is why lack of preference is not transitive.[8] The same example shows how ordinary, intuitive cases of indecision are kept *out* of the domain of defined indifference. For even though I do not prefer the Americano for $2

[8] In general, the union of two transitive relations need not be transitive, but that is not the point here. The transitivity failure of indecision is inherited by lack of preference.

to the latte for $3, nor the latte for $3 to the Americano for $2, my attitudes toward the two do not satisfy the second clause of defined indifference, since there is some X, namely the cheaper Americano, which I prefer to the $2.10 Americano but not to the latte. The intuitive idea is that you cannot be properly indifferent between two options if there is something you prefer (or disprefer) to one and not to the other. To be indifferent is to rank the two together, so that each is below everything the other is below, and each is above everything the other is above. In this way defined indifference resembles indifference.

So far so good, but now comes a hitch. When someone is genuinely indifferent between a pair of options, she will also be defined indifferent between them, but the converse may not be true. There are counter-examples, and they may be serious ones, because they may ruin the prospect of using defined indifference to solve the negation problem.

> **The Thoroughly Indecisive Example:** Lily can't make up her mind about anything. For any pair of options, she has no preference between them. She is not indifferent; rather, she is utterly indecisive.

Lily is, by stipulation, completely indecisive, but she also turns out to be maximally defined indifferent. For each pair of options, Lily is defined indifferent between them. This is because everything she prefers to one she prefers also to the other, for the reason (not intended in our definition) that there *is nothing* she prefers to either; similarly, there is nothing to which she prefers one item but not the other.

Now Lily is a very, very peculiar creature. She is an extreme, degenerate example of a preferring agent. If defined indifference does not match the ordinary notion of indifference for a being like Lily, maybe that is OK. After all, I do not need to insist that defined indifference is indifference, nor even that it exactly matches the extension, for every possible agent, of indifference. Still, the example of Lily is theoretically problematic. I cannot claim that defined indifference represents indifference in an expressivist account of normative judgment. For Lily's attitudes are like those of Officer Lopez (in section 3 above) and not like those of Mr Manners, and we said there must be a difference between those attitudes, so there must be a difference between Lily's preferences and the perfectly indifferent preference structure. But Lily turns out to be defined indifferent between every pair.

If the Thoroughly Indecisive Example were the only counter-example to the suggestion, maybe an expressivist could swallow it; maybe our intuitions deceive us and there is in fact no difference between the normative outlook of extreme limits of the collections of attitudes of Officer Lopez and Mr. Manners. But there is another counter-example.

> **The Example of Momentous and Trivial Choices:** Henry is mainly concerned with affairs of state. He prefers avoiding nuclear war to engaging in it. He prefers maintaining his nation's sovereignty to losing it. And he has many other preferences of enormous national importance. Right now, though, he has to order lunch, and he can't make up his mind. Ham, or tuna? He is undecided.

The problem is that although there are many things that Henry prefers to getting a ham sandwich, like avoiding nuclear war, maintaining sovereignty, and so on, he naturally prefers all of those things also to getting a tuna sandwich; similarly there are many things to which he prefers getting a ham sandwich, like nuclear war and giving up national sovereignty, but he prefers tuna to those things, too. Henry has a Big Important preference structure, and tucked into the middle of it he has a little bubble of indecision regarding unimportant (but not negligible) matters. Henry turns out to be defined indifferent, then, between tuna and ham, although we stipulated that he is undecided between them.

What has gone wrong with my apparatus of defined indifference? It relies on there being a sufficiently rich field of preference to guarantee that something or other will be wedged beneath or above any given item, but very close to it. If there is, then the wedged object will be neither above nor below the item to be paired with the given one. In my leading example, I added a saving of 10 cents to the Americano, plausibly very close in an ordinary person's preference structure, so close as not to make enough difference to yield a definite preference between the money-saving Americano and the latte. For ordinary preferences, it is safe to assume a rich structure, but in theory a person's preference field could be severely impoverished, either overall, as Lily's, or locally, as Henry's. Call this the Problem of the Impoverished Field.

I cannot stipulate that everyone has a rich field of preference, and I cannot rely on the fact that real people do have rich fields. It is not up to me what preferences people can possibly have, and since my suggestion is offered to expressivists as an aid to their explanation for what it *is* to have this or that normative view, I have to be sure that it works in theory for possible as well as actual preferences.

8. Some Untidy Last Thoughts

Fertilizer: inserting lotteries to enrich the field

It is hard to imagine someone with very severely impoverished preference structure. It may be harder to imagine than you think. You know obsessive people, and maybe you can in imagination increase someone's obsessiveness until you imagine them single minded, or like Henry so generally absorbed with enormous issues as to leave sterile little pockets in their field. But we can ask about some artificial objects of preference, objects engineered to enrich the field of preference in just the way we need. These hunks of fertilizer are the so-called 'lotteries' of decision theory fame.

Ask Henry whether there are things he prefers to a ham sandwich that he doesn't prefer to a tuna sandwich, and he can't think of any. He can only think of the momentous affairs of state, all of which he enormously prefers to either sandwich or enormously disprefers. But take one such momentously good object, say the prospect of rapprochement with Cuba, and divide its desirability (to Henry) by a large number. Offer Henry the ham sandwich augmented by a one millionth chance of the rapprochement—that is, tell him (convince him) that if he takes the ham sandwich, the prospects of coming to terms with Cuba will be enhanced by one in a million. Does this tiny extra something get him to prefer ham to tuna? What if the enhancement were a one in a billion enhancement? One in ten to the thousandth power? Henry is defined indifferent, recall, only if there is *nothing* he prefers to a ham sandwich that he doesn't prefer to a tuna sandwich. Plainly any tiny enhancement to the ham prospect will bring it above the plain ham sandwich, so if there is any such tiny enhanced ham prospect that he doesn't prefer to the tuna prospect, Henry must be undecided and not indifferent.

Lottery fertilizer is nice because it comes in continuous quantities. If a certain enhancement is too large, you can always cut it in half. We should keep in mind that there are other enhancements that can act as field fertilizer, even though they do not have the continuity of lotteries. For example, we could offer to donate one penny to Oxfam if Henry gets the ham sandwich. Or, again exploiting a ready-made continuum, we could threaten to postpone one of the Momentous Goods by a day, or a minute, or a tenth of a second, . . . if Henry orders tuna.

Add enough fertilizer to the field and it becomes more and more difficult to imagine Henry being genuinely undecided between the kinds of sandwich even though *no* comparison can separate them. Maybe the bare, unimaginable possibility of that sort of indecision can be shrugged off by an expressivist. Maybe an expressivist can say that in such a case Henry does believe each kind of sandwich is permissible, or that there is simply no determinate fact as to whether he regards each kind as permissible or instead has no view yet. I'm not confident that the fertilizer strategy makes it possible to shrug off the counter-examples, but it does seem to make the bullet easier for an expressivist to bite.

Imputed intentional states

One objection to the fertilizer strategy is that it can simply be stipulated away. Finish the story like this: Henry does not think about lotteries and never considers conjunctions of sandwich prospects with tiny chances of momentous prospects, since those conjunctions are of no possible practical interest and Henry doesn't for a moment think that anybody can *really* make the prospect of rapprochement with Cuba be more likely conditional on Henry's lunch. He therefore has no preferences at all among such things. He is undecided, though not in the sense that he's thought and thought and thought about the sandwiches and just can't make up his mind. And we can similarly stipulate away any thoughts of monetary or temporal commensuration of the momentous and trivial prospects.

There seems to me to be something fishy about this objection. When we impute preferences to agents, we do not in general suppose that the agents have entertained explicitly the comparison we have in mind. I know that my sister prefers a spectacular view of the Grand Canyon to being immersed in icy water for ten minutes. True, I might express this knowledge by saying that she *would* prefer the view to the immersion, but I don't merely know that she *would*, I know that she *does*. By contrast, I know that my son *would* prefer eggplant to caviar if only he were to try them, but he does not now have any such preference.

People have preferences that they've never considered just as they have beliefs they have never entertained, like the belief that 123,000,000,000 is not a prime number. These preferences and beliefs may be analyzable as dispositions to believe or prefer under the right circumstances. If they are, though, they will have to be distinguished somehow from dispositions to believe and prefer things that we do not actually believe and prefer, eggplants to caviar or a moderately obvious but obscure theorem of Euclidean geometry. It may be difficult to distinguish the never-considered-it kind of lack of preference from indifference, in some cases, but perhaps this is no objection to an expressivist program. After all, it may be similarly, analogously difficult to distinguish the state of lacking any normative view about a certain subject from the state of having a quite permissive view about the subject.

9. Conclusion/Recap

The problem of negation, that is, the problem of explaining what in general is expressed by the negation of a sentence expressing a given attitude, is made more tractable by the model of hyperplans proffered in *Thinking How to Live*. On the simplest account, hyperplanning rules out the possibility of mere permissions by its very structure. Allowing ties, thought of as representing indifference, into plans recaptures the possibility of mere permissions. It does not allow for supererogation, but, I've argued, it is not obvious that supererogation enters normative thinking on the ground floor, and the higher tiers might be built up from the foundation in ways that can plausibly account for ethical supererogation. Indifference, however, must be distinguished from indecision, and the task of distinguishing them looks a lot like the original task of explaining the difference between negating an attitude and just lacking it. I suggested that 'defined indifference', picked out by some auxiliary clauses guaranteeing transitivity, would do as an approximation of the ordinary notion of indifference. Some difficulties show up at the margins of possibility, making defined indifference indistinguishable from indecision when the agent's field of preference is radically impoverished, either globally or just locally, but I gave some reasons to think that the difficulties are not too serious.

REFERENCES

Blackburn, Simon, 'Attitudes and Contents', *Ethics*, 98 (1988), 501–17.

Broome, John, 'Can a Humean be Moderate?', in R. G. Frey and Christopher Morris (eds.), *Value, Welfare and Morality* (Cambridge: Cambridge University Press, 1993), 51–73.

Dreier, James, 'Why Ethical Satisficing Makes Sense and Rational Satisficing Doesn't', in Michael Byron (ed.), *Satisficing and Maximizing* (Cambridge, Mass.: Cambridge University Press, 2004), 131–54.

Gibbard, Allan, *Wise Choices, Apt Feelings* (Cambridge, Mass.: Harvard University Press, 1990).

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

Hale, Bob, 'On the Logic of Attitudes', in J. Haldane and C. Wright (eds.), *Reality: Representation and Projection* (Oxford: Oxford University Press, 1993), 337–63, 385–8.

Unwin, Nicholas, 'Quasi-Realism, Negation and the Frege-Geach Problem', *Philosophical Quarterly*, 49 (1999), 337–52.

—— 'Norms and Negation: A Problem for Gibbard's Logic', *Philosophical Quarterly*, 51 (2001), 60–75.

This page intentionally left blank

9

Direction of Fit and Motivational Cognitivism

Sergio Tenenbaum

1. Introduction

The idea of direction of fit has been found appealing by many philosophers. The idea goes back to Anscombe's famous example of the different aims of two agents: a man who is doing his shopping guided by a shopping list, and a detective compiling a list of the man's groceries as he buys them. I can't resist adding myself to the legion of writers who quote the passage from *Intention* in which she points out an important difference between the two directions of fit:

> If the list and things that the man actually buys do not agree, and if this and this alone constitutes a *mistake*, then the mistake is not in the list but in the man's performance ... whereas if the detective's record and what the man actually buys do not agree, then the mistake is in the record.[1]

In both cases of mistake, there's a lack of fit between the world (in particular, what's in the shopping cart), and the agent's mental states (represented by what they have written in their respective lists). But the different "location" of the mistake in the two cases is supposed to show that desires and beliefs

I would like to thank Danielle Bromwich, Philip Clark, Thomas Hurka, Niko Kolodny, Jennifer Nagel, Fred Schueler, Rob Shaver, David Sobel, and two anonymous referees for Oxford University Press for insightful comments on earlier versions of the paper. Versions of this paper were read at the Ethics and Political Workshop at the University of Toronto, at the Metaethics Workshop in Madison, Wisconsin, and at 2005 Workshop on Moral Psychology at Franklin & Marshall College. I would like to thank members of the audience for a lively and very helpful discussion on all these occasions, and in particular, Jeff Seidman, who commented on the paper at the Franklin & Marshall workshop. Research on this paper has been generously supported by a grant from the Social Sciences and Humanities Research Council of Canada.

[1] Anscombe (1963: 56).

have different directions of fit; in the former case the world was supposed to fit the desires (the shopping cart should have been filled in accordance with the shopping list), whereas in the latter case the beliefs were supposed to fit the world (the detective's list should have been made in accordance with what was in the shopping cart). The example, as well as the direction of fit metaphor, seems to many to capture something very important about the different natures of belief and desire, something that might have profound implications for various views about motivation and practical reason. Perhaps the most notable among these supposed implications is the Humean Theory of Motivation, the claim, roughly, that beliefs cannot motivate on their own.[2] But, of course, philosophers haven't rested content with trying to derive these implications from one example and a metaphor. Philosophers have tried to make the notion of direction of fit more precise, or to provide various explanations of the intuition that the detective's and the shopper's attitudes are to be distinguished in terms of directions of fit.[3] There are at least two promising strategies in the literature for spelling out the notion of direction of fit. The first strategy cashes out the metaphor in terms of the different relations of counterfactual dependence between, on the one hand, belief and the world, and on the other hand desire and the world (or on some other attitude that is supposed to track the world). The second strategy appeals to existence of a constitutive relation between truth and belief: belief aims at the truth, whereas desire doesn't. I will argue that the first strategy collapses into the second. However the idea that there is a constitutive relation between belief and truth is itself rather vague, and it is hard to see how it can *explicate* the metaphor of direction of fit, rather than just replace one metaphor (the direction of fit metaphor) with another (the metaphor that belief aims at the truth). The second part of the paper examines whether we can understand the notion of direction of fit in terms of the constitutive relations that belief, but not desire, bears to the truth. I argue that there is indeed a way to cash out the metaphor of direction of fit in these terms; in particular, I argue that the metaphor is best cashed out in terms of the different formal ends guiding the inference from what I call "prima-facie" attitudes to what I call "all-out" attitudes in the theoretical and practical realms.

With this (hopefully) improved understanding of the distinct directions of fit of belief and desire, we can ask whether the fact that desire and belief have these distinct directions of fit can have the rich implications that

[2] See Humberstone (1992) for further philosophical uses of the notion of direction of fit.

[3] Smith, 1994; Humberstone, 1992; Zangwill, 1998; Platts, 1997. For criticisms, see Humberstone, 1992; Copp and Sobel, 2001; Schueler, 1991; Schueler, 2003.

many philosophers have tried to draw from it. Unfortunately, as I argue in section 4, the answer is "no". In particular, I examine whether any notion of direction of fit indeed implies the Humean Theory of Motivation. The Humean Theory of Motivation stands in opposition to the position that I call "motivational cognitivism". Roughly, motivational cognitivism views the relation between (moral) knowledge and moral action as a relation between capacity and exercise; according to the motivation cognitivist, there is some kind of moral knowledge such that its possession guarantees that the agent is motivated to act morally. Is motivational cognitivism compatible with the claim that the theoretical and practical inquiry have distinct formal ends? I argue in section 4 that motivational cognitivism is not only compatible with this claim, but that the idea that theoretical and practical inquiry have different formal ends helps provide an attractive formulation of motivational cognitivism.[4] In particular, I try to show in this section how the notion of direction of fit helps answer a familiar objection to motivational cognitivism. Motivational cognitivism is often presented as the view that accepts the existence of "besires".[5] Besires are supposed to be complex mental states that have the direction of fit of belief towards one content (say *p*) and the direction of fit of desire towards another content (say *q*). However, this very complexity suggests that the motivational cognitivist has simply gerrymandered a state to fit her position. Once we recognize that the mental state is Janus-faced in this way, why couldn't the faces be pried apart? Why aren't we talking about two distinct states that might or might not be co-instantiated by an agent at a time? Our analysis of direction of fit delivers an improved understanding of the kinds of mental states that the motivational cognitivist must postulate and allows her to answer these questions satisfactorily. It allows us to see that the motivational cognitivist is not committed to anything gerrymandered or out of the ordinary; the kind of complexity in question turns out to be no different than the complexity we are committed to accept independently of our views about the nature of moral motivation.

The last section tries to answer an important objection to the argument of the paper. It seems that understood this way, the notion of direction of fit does not fully capture the differences between theoretical and practical reasoning brought to light in Anscombe's example. In particular, it does not account for the fact that there seems to be nothing wrong with

[4] Although it is hard to make claims about past philosophers that are not contentious, I believe Socrates and Kant are clear examples of motivational cognitivists. Among contemporary ethicists, John McDowell seems to be defending such a view in McDowell (1998).

[5] This awkward, but doubtless very useful, label is introduced by Altham (1986).

the propositional attitudes held by the shopper; to the extent that the agent makes a mistake in this case, it, as Anscombe describes it, is a mistake of performance. One might think that it is exactly this feature of the example that should lend support to the Humean Theory of Motivation. As long as this feature is left dangling, one cannot avoid suspecting that even if we have captured *a* notion of direction of fit, we have not captured *the* notion of direction of fit that lends support to the Humean Theory of Motivation. However, I argue that there's something problematic about this characterization of the shopper's attitudes. There *is* something wrong with the propositional attitudes of the shopper; mistakes of performance are best understood as inferential mistakes. What the example reveals, however, is that the content of all-out attitudes in practical reason is always particular, and that, in practical reason, the inference from general to particular is non-trivial in a way that finds no parallel in theoretical reason. And although this is doubtless an important difference between practical and theoretical inquiry, it brings the Humean no support.

2. Working out the Metaphor: The Counterfactual Dependence Strategy

At first, there's something quite intuitive about the distinction between a mind–world and world–mind direction of fit. The basic idea is that a mental state could aim either at tracking the world or at changing it, and to each of these will correspond a different direction of fit; beliefs tend to, or ought to, fit the world, while desires tend to, or ought to, make the world fit them. Explicating these metaphors, however, has been notoriously difficult. One seemingly promising strategy to cash them out is what I call the "counterfactual dependence strategy". The counterfactual dependence strategy tries to explicate the different directions of fit of beliefs and desires by appealing to the fact that beliefs and desires that *p* have different relations of counterfactual dependence to *p* itself or to some third attitude towards *p*. The general idea is to explore the intuition that my belief that *p*, but not my desire for *p*, should be tracking the facts. On the other hand my desire that *p*, but not my belief that *p*, loses its point once *p* has been brought about.

The counterfactual dependence can be either strict or loose. That is, one can claim that the counterfactual dependence always obtains or that it obtains in most or in normal cases. Insofar as one aims to provide an analysis of the notion of belief in terms of direction of fit, appeal to a loose relation of dependence will probably be of no help.[6] But establishing a loose connection might be enough if one has more modest philosophical ambitions, so I'll leave this possibility open.

The straightforward, but certainly hopeless, version of this strategy would be to claim that belief, but not desire, is counterfactually dependent on *p* itself. Given that we're neither omniscient nor infallible, there could be no such strict relation of counterfactual dependence between the belief that *p* and *p* itself. But a looser relation does not fare much better. Let us take a quick look at a possible suggestion:

(1) Under normal circumstances, *S* would not believe that *p* if it were not the case that *p*

This suggestion faces a dilemma. On the one hand, one can specify "under normal circumstances" so as to make sure that (1) will come out true. If, for instance, we were to understand "normal circumstances" as "circumstances under which a believer is reliable", then (1) would indeed be true.[7] But it's hard to see how to spell out the idea that an agent is reliable in a certain context other than by an appeal to the idea that in such a context the agent tends to believe that *p* only if *p*. More generally, there doesn't seem to be any way to spell out the idea of "normal circumstances" that is non-arbitrary and that makes (1) come out true. Certainly "normal" could not be a statistical notion; there are many common circumstances in which one tends to form false beliefs. We could substitute a normative notion, such as "for appropriately formed beliefs" for the notion of "under normal circumstances", but this would encounter a similar dilemma. On the other hand, we could stipulate that only true beliefs are appropriately formed, but this would trivialize condition (1). Or we could, for instance, identify "appropriately formed beliefs" with beliefs that were formed by rational processes. But here it would seem that one could form false beliefs when one is following otherwise rational processes.[8] Faced with these problems, a wise proponent of the counterfactual strategy should opt for trying to find a counterfactual dependence between belief and a mental attitude that, in some way, is supposed to track the world. However it's unclear that such proposals

[6] Humberstone (1992) places universality as a constraint on the notion of direction of fit. However, whether this is a reasonable constraint depends on the aim of invoking a notion of direction of fit.

[7] Of course, even this claim is a simplification, since a context in which a believer is reliable is not necessarily one in which she'll be infallible.

[8] One can avoid this possibility by identifying rational processes with reliable processes of some sort. But to close off this possibility one would need to understand reliable processes as ones that tend to generate true beliefs, again trivializing condition (1).

can overcome the original dilemma. Take, for instance, Michael Smith's account of direction of fit in these terms:

> A belief that *p* tends to go out of existence in the presence of a perception with the content that *not p*, whereas a desire that *p* tends to endure disposing the subject to bring it about that *p*.[9]

"Perception" can be read in two ways. In the first, "perception that *p*" implies a "belief that *p*". In this sense of "perception that *p*" is not much different than a belief that *p* (or perhaps a specific case of a belief that *p*). In a second sense of "perception that *p*", one can have a perception that *p* even when one does not believe that *p*. Let us start with the first sense. Of course if one's aim is to provide an *analysis* of belief, using a notion of perception that presupposes this very notion is not going to be of much help.[10] But even if one is not intending this to be an analysis of belief, the above proposal cannot be very illuminating if perception is to be understood as being a belief. For all the proposal would say is that belief that *p* is incompatible with belief that *not p*, whereas desire that *p* isn't. This no doubt shows that belief and desire are not the same attitude, and perhaps even that typically desire that *p* and belief that *p* tend not to coexist. Understanding the proposal this way makes it come out true, but not very informative. In fact, the same contrast can be drawn between "belief that *p*" and "supposing that *p*", or "suspecting that *p*", or "wondering whether *p*", but this obviously does not show that these attitudes have a distinct direction of fit, let alone that they have the same direction of fit as desire; it just shows that none of these attitudes can be identified with belief. In fact, even if one is not hoping for an analysis of the notion of belief one would hope at least that the notion of direction of fit would throw light on the different ways in which belief and desire are or should be related *to the world* (or to the facts). But understood in this manner, the notion of direction of fit speaks only to the different "mind–mind" relations that beliefs bear to other beliefs and desires.

I take it that Smith himself intends the second reading of the notion of perception, the one according to which "perception that *p*" does not imply "belief that *p*".[11] But it is not clear how to spell out this notion of perception in such a way as to make Smith's proposal come out true. Copp and Sobel summarize the problem nicely:

> It might seem ... unsurprising that we cannot find an introduced state that counts as in some way a perception with the content that not *p*, that is *not*

[9] Smith (1994: 115).

[10] Copp and Sobel (2001) make this point.

[11] In fact, if he did, substituting "belief " for "perception" in the quote above would do just as well, and would be much more perspicuous.

itself a belief, but that interacts with the belief that *p exactly as if* it were an incompatible belief.[12]

After all insofar as a perception that *p* does not imply that one believes that *p*, one could have a perception without forming any tendency to believe. Copp and Sobel give the example of the common optical illusion, in which the asphalt ahead of the driver on a highway might look like a puddle of water. This kind of illusion does not have any tendency to make one believe that there's a puddle on the road; drivers are typically not fooled in any way by it. We can call these optical illusions 'innocent' illusions; although they are cases in which it looks to the subject as if it were the case that *p*, being under this kind of illusion does not make it more likely that the agent will believe *p*. So if the proposal is to come out true, we cannot include this kind of optical illusion as a case of 'perception'. However, it seems hard to find a mental state *M* that satisfies all of the following:

(a) Innocent illusions are not cases of *M*
(b) Being in state *M* with the content *p* does not imply that the subject believes that *p*
(c) The relation between mental states *M* and beliefs with the same content can provide an adequate explanation of the notion of direction of fit.

However, it's not clear that these difficulties are insurmountable. Let us think of a mental state that can be loosely described as "taking *X* as evidence for *p*", or "taking it to be the case that there is reason for *p*", or simply "it appearing that *p*". We can arrive at a somewhat more precise understanding of the state as follows: belief is an "all out" state. That is, believing that *p* is incompatible with believing that *not p*, and there are no states that override one's belief that *p* in the formation of one's unconditional theoretical stance towards *p*. There is no state of, say, "really, I mean it, believing that *p*" for which the belief that *p* provides prima-facie grounds. "Having it appear that *p*" on the other hand is the prima-facie "version" of a belief that *p*. In the absence of countervailing evidence or any reason to think that the appearance is illusory, being in such a state will lead the subject to form the belief that *p*. Using this state in order to understand belief is informative insofar as being in a state such that it appears to the subject that *p* does not imply that the agent believes that *p*. Moreover being in such a state does dispose someone to form the belief that *p*, at least to the following extent: if it appears to a subject that *p* and yet the subject does not believe that *p*, then some explanation is required in terms of countervailing dispositions

[12] Copp and Sobel (2001: 49). See also Schueler (1991) and Humberstone (1992).

or countervailing reasons. Copp and Sobel consider a similar proposal, but they discard it on the grounds that such a state would presuppose a notion of belief. According to them,

> if [the subject] allows that there is evidence for not *p*, she must believe that it counts in favor of believing that *not p*. That is, the judgment that something counts as evidence for *not p* just is a belief.[13]

But there's no reason to think that the state of "appearing that *p*" is such that you are in such a state if and only if you *believe* that there's some X such that X counts as evidence for *p*. For instance, one can feel oneself in the grip of the gambler's fallacy even if, upon reflection, one believes that there is no reason to think that, say, it is less likely that the next coin toss will be heads, given that the last five ones were heads. In other words, even if I don't believe that the previous tosses count as evidence that the coin will land tails next time, I still *take* the tosses to be evidence for the proposition that the coin will land tails next time. One can say here that the agent holds contradictory beliefs, but it seems more plausible to say that it appears to the agent that it's more likely that the coin will land on the tail side, or that she *takes* the previous tosses to be evidence that the coin is more likely to land on the tail side, even if she does not believe the previous tosses constitute a reason to believe that heads is less likely.

It might be useful to take a look back at the case of the puddle illusion, and see how it would be handled by the proposal we're considering. We can think of two possibilities here. First, it might be the case that the best account of the driver's process of belief formation is something like the following: the driver does take the visual perception to be *some* evidence for believing that there is a puddle up ahead on the road, but the evidence in question is overridden by his knowledge that, under those conditions, such perceptions are likely to be misleading, and the absence of any further reason to think that there are puddles on the road. In this case, the state in question is an appearance of the relevant kind (a prima-facie theoretical attitude), but it is no counter-example to the view that such states always dispose the agent to believe the content of the state. The agent in question does form a disposition to believe that there is a puddle ahead on the road, but it's "neutralized" by countervailing dispositions to refrain from forming the belief in such situations. Or perhaps the best account of the process is one in which the agent never takes the visual perception to be evidence for believing that there is a puddle in the road. In this case, the visual perception does not give rise to a disposition to believe its content. But it is also not a case of being an appearance of the relevant kind; it's not a prima-facie attitude at

[13] Copp and Sobel (2001: 49).

all (even if, typically, visual perceptions are prima-facie theoretical attitudes, this is a case in which one isn't). So, also in this case, the puddle illusion is not a counter-example to the proposal in question. In fact, understood this way, we have a close parallel between, on one hand, appearances and beliefs, and on the other hand, desires and actions. Incompatible desires need to be sorted out on the way to action, and, similarly, incompatible appearances must be sorted out on the way to belief.[14]

However, even though this proposal escapes each horn of the dilemma, it faces a different problem. As I pointed out above, desire is in fact analogous to the state of appearing that *p* rather than to belief. In light of this point, we can say more generally that some attitudes are *prima-facie in character* whereas others are *all-out in character*, or "prima-facie" and "all-out" attitudes for short. Appearances and beliefs are, respectively, prima-facie and all-out attitudes in the theoretical domain. On the other hand, desires and actions (or intentions) are prima-facie and all-out attitudes in the practical realm. We can now say that this revised version of the counterfactual strategy accounts for the notion of direction of fit in terms of two distinct pairs of prima-facie and all-out attitudes that belong to the two distinct realm of inquiries: appearance and belief in the case of theoretical inquiry, and desire and action (or intention) in the case of practical inquiry. However the notion of direction of fit was supposed to characterize exactly what was distinctive to each realm. It does not help to notice that there are two corresponding pairs of attitudes, rather than just one pair, that are candidates for this characterization. In fact, as far as the positive characterization goes, so far we simply assumed that there are two pairs of attitudes rather than just one. Tempting as it is to think that desires and actions (or intentions) on the one hand, and appearances and beliefs on the other hand, must form two distinct relations, the prima-facie and the all-out attitudes, we just *assumed* that they are different; we did not provide any characterization of the difference. It is also tempting to invoke the metaphor of direction of fit here to explain the difference between the two pairs of attitudes, but this is obviously circular.

One can however make progress here by trying to identify distinctive features of the relations between the prima-facie attitudes and the all-out attitudes in the different fields. Theoretical inquiry is the search for what is the case, and practical inquiry is the search for how to act; these different kinds of inquiries might dictate different relations between prima-facie and *all-out* attitudes. In particular, one might want to say that, in theoretical

[14] In Tenenbaum (1999), I discuss this parallel in more detail and argue that even the ill-formed belief in the case of the gambler's fallacy finds a parallel in the case of practical reason; I argue there that we should understand *akrasia* in similar terms.

inquiry, prima-facie attitudes are (should be) taken up in all-out attitudes insofar as the agent accepts (should accept) that the content represents how things are, whereas in practical inquiry, prima-facie attitudes are (should be) taken up in all-out attitudes insofar as the agent accepts (should accept) that the content represents what he or she is to do. To say that one's inferences from prima-facie attitudes in theoretical reason to beliefs is or should be guided by *how things are* is to say that the process of belief formation is in some sense guided by the ideal that one's beliefs should be guided by the truth. In other words, this proposal now postulates some kind of constitutive relation between belief and truth; roughly speaking, a belief is an attitude whose formation is, or ought to be, guided by the pursuit of truth. We can say in this case that truth is the *formal end* of inquiry. It's an end that guides, or ought to guide, every instance of engaging in theoretical inquiry.[15] And one might propose that desire and intention, on the other hand, bear a similar relation not to the truth, but to something else. This something else could be 'the good', 'the desirable', 'rational action', 'autonomous action', or something else. I'll assume that what ought to guide us in those transitions in practical reason is 'the good',[16] but again, the argument does not hang on this being the correct choice.

This is indeed a promising proposal. But one must note that this proposal leaves the counterfactual dependence strategy behind; we end up trying to capture the notion of direction of fit in terms of the constitutive relation between belief and truth. The counterfactual dependency strategy collapses into a strategy that tries to capture the direction of fit in terms of, on the one hand, the relation between belief and truth, and, on the other hand (if I am correct about what the formal end of practical reason is), the relation between desire and the good. We can now simply ask what (if any) implications flow from the fact that belief and desire have different formal ends. But before we can answer this question we need to clarify what it means to say that belief is, or ought to be, guided by truth, whereas desire and intention are, or ought to be, guided by the good.

[15] Of course I cannot do full justice to the various issues surrounding the notion of a formal end of inquiry here. I discuss these issues in more detail in Tenenbaum (forthcoming (*a*)). However, a few words of warning might be important. I am not using the notion in the same way as Velleman does (2000*c*), at least insofar as Velleman thinks that specifying the formal end of an inquiry is completely uninformative. If anything, the notion is closer to what he calls there the "constitutive aim" of inquiry or belief. However, it is not quite the same notion either since I am not committing myself to the view that the formal end of inquiry could be *fully* understood apart from its being what constitutes successful inquiry. For an illuminating discussion of these issues, see Clark (2001). The formal end of an inquiry is in my view what Clark calls a "generic object".

[16] Mostly because I think that this is the correct view. See Tenenbaum (forthcoming (*a*)).

3. Aiming at the True and the Good

Although most philosophers agree that belief bears a certain constitutive relation to the truth, the characterization of this relationship is no easy matter. To say that the relation is constitutive is to say that nothing can count as a belief that *p* unless it stands in this relation to the fact that *p*. And, of course, as long as we understand the relation in this manner, we must use some expression like "the aim of belief" in spelling out the condition; a belief that *p* can easily coexist with the fact that *not p*. One can say that believing *p* implies holding *p* to be true. But the truth of this statement depends on what we mean by "holding true". There is a sense in which we hold a statement to be true when we assume something for the sake of argument,[17] and one can certainly believe things without explicitly considering the matter in terms of the truth of a sentence or a proposition. I'm not sure there is any non-trivial way to characterize the relevant sense of "holding true". We can say things like "holding with endorsement" or "holding with acceptance", but if "endorsement" and "acceptance" don't just mean "belief" in this context, it'll be easy to build counter-examples to the claim that to believe is to hold with endorsement or acceptance.[18] For our purposes, it suffices to say that believing *p* implies holding *p* to be true in the sense of "holding true" characteristic of belief. In any case, put this way, this is not quite a characterization of truth as the aim of belief. After all, to say that I hold something true is not to say that I hold it true because I aim to hold it true;[19] it certainly does not follow from the fact that one holds x to be y that one aims to do so. The idea that belief aims at the truth is more robust than the idea that believing *p* amounts to, or implies, holding *p* to be true in a certain way. But how should we understand what the postulation of such a constitutive aim adds to the idea that believing implies holding true?

One way to understand this addition is to think that an agent does not count as believing that *p* unless the agent forms the belief guided by the aim of believing truly. Now this idea needs some refinement. Obviously we do not form beliefs by engaging in explicit instrumental reasoning about maximizing our chances of hitting the truth for every single belief we have. We need to understand having this end in a way that does not

[17] Velleman (2000*b*).

[18] Cf. Van Fraassen's (1980) distinction between "accepting" a theory and "believing" a theory.

[19] This is indeed Velleman's characterization of belief (2000*b*). It'll become clear momentarily why I think that this is not an adequate characterization of belief.

imply anything so obviously false. Various things are often said with respect to belief that could be helpful in understanding better what it is to have such an end. One can say, for instance, that we cannot form beliefs at will,[20] and thus that we cannot have any aim other than believing truly in forming beliefs. Or, one can appeal to Moore's paradox[21] to explain the impossibility of forming a belief in disregard of what one considers to be true. But probably the clearest and most promising way to spell this idea out is to say that belief must respond to evidence;[22] that is, no state counts as belief for *p* if the state is not responsive to the evidence for or against *p*. Again, the notion of "responding to evidence" needs to be clarified here. It's obviously false that all our beliefs are proportioned to evidence, and some of our beliefs, especially unconscious beliefs, beliefs that are the result of wishful thinking, etc., are formed by processes that are in no way truth-conducive.[23] The condition of responsiveness to evidence should be something like the following:

> (RE) An agent counts as believing that *p* only if the agent does not consciously hold the belief due to non-epistemic reasons.

This is a relatively weak requirement. (RE) allows that some beliefs be held for no reason. It also exempts unconscious beliefs, and beliefs that result from self-deception and the like, from its "evidentialist" requirements.[24] Ideally, we would spell out what is meant by "epistemic reasons" and what counts as a belief being due to a reason rather than another. But since spelling out would cost some generality, I will leave this task to the side. The idea that something like (RE) must be true is initially very plausible, but I think it cannot withstand scrutiny. (RE) is particularly plausible if it is understood as part of a *general* condition on an agent *having beliefs*; it is plausible to assume that no one can count as a believer if his beliefs do not satisfy the consequent of (RE) often enough. However, as a specific condition of what makes a *particular* mental state a case of an agent having *a* belief, I think it is false. Here are a few counter-examples to the *specific* condition:

1. Mary is up for the job of her dreams. She looks at the ad, and she's struck by the thought that she is a shoo-in for the position. But Mary thinks that the experience of failing to get the job of one's dreams after expecting that one would is so painful that it's better not to

[20] Williams (1973). [21] Railton (1997).

[22] For instance, Wedgwood (2002). [23] See Shah (2003).

[24] To make matters simpler, I'm also leaving aside the fact that someone defending (RE) would also want to include a further requirement to the effect that an agent does not believe *p* when she is in possession of overwhelming evidence for *not p*.

believe that she'll get the job. She decides to persuade herself that she'll not get the job.

2. Clara's husband has been indicted for a crime. All the evidence points to his guilt. However, Clara doesn't believe that her husband is guilty. When confronted with evidence, she says: "I trust my husband, and to trust somebody involves being committed to believe his innocence, even when the evidence warrants the opposite conclusion."
3. Otto's son is missing. All evidence points to the child's death. Otto acknowledges this fact, but he says: "I can't just let go of him like this. I must continue to believe that he's alive (and thus I believe that he's still alive)."

Now these are cases in which, although the agent in question still seems to hold the propositions in question to be true, they are not cases in which the agent tries to form the belief that *p* only if the evidence warrants the belief. They are also not Pascal-like cases in which the agent forms a plan now to ensure that in the future she'll find warrant for a certain proposition ((1) is the closest to this case).[25] These are beliefs that are not *currently* being sustained by any kind of aim of maximizing the chances that the belief is true or in accordance with the evidence.

These are all cases in which the formation of belief is guided by goals other than truth. To the extent that (1)–(3) are compelling, the idea that belief must be guided by truth is descriptively false.[26] Of course these examples do not conclusively establish this point. One can, for instance, try to explain away these cases as cases in which the agent behaves *as if she believed*, but not cases in which she really believes the statement in question. Or one could say that these are cases in which the agent in question uses some kind of non-standard evidence. The father takes a certain gut feeling as evidence, or the spouse her special acquaintance with her partner as evidence. I don't find these replies promising; I don't think that one can escape the conclusion that one does not always aim at the truth when forming beliefs. Although I can't argue in detail for these conclusions, I hope these examples suffice to give us some reason to think that it would be best to account for the distinctive direction of fit in terms of a *normative*

[25] It is worth noting that faith-based belief in God often seems explicitly to run afoul of (RE). Yet it would be hard to say that people who claim to believe in God while acknowledging that there's no evidence for their existence don't have a proper belief. I owe this example to Fred Schueler.

[26] This needs some qualification. It might be correct to say that I can't count as believing *p* if my belief is supposed to survive overwhelming evidence to the contrary. I'm not going to try to settle this issue.

relation between belief and truth. In particular, we can say that believers are under something like the following normative requirement:

(NR) Believe *p* only if *p* is true.

The kind of normative requirement in question needs to be stated with caution. Cases (1)–(3) do not clearly involve an agent who is doing anything that is, all things considered, wrong or irrational. It might be that trust does require that we override evidence, or that one is better off expecting the worse, or that only a heartless parent would accept anything that far from Cartesian certainty to form the belief that his child has passed away. The normative ideal in question must be an ideal for belief considered solely from a theoretical perspective, insofar as we are engaged in the search for truth abstracted from any other concerns. (NR) is thus best understood as a normative claim about how belief ought to be responsive to evidence *insofar as an agent is engaged in theoretical inquiry*. Of course, the closer one is to accepting the view that belief about *p* should be understood simply as an agent's all-out attitude insofar as she is engaged in theoretical inquiry, the more stringent one's interpretation of "often enough" will be, in our claim above that in order to count as a believer one needs to satisfy the consequent of (RE) often enough. But (1)–(3) should make us suspect that "often enough" cannot become "always".

Obviously there is room for refining (NR), but the simple version of the requirement should suffice for our purposes.[27] To say that (NR) should guide one's belief formation at least insofar as one is engaged in theoretical inquiry is to say that in theoretical inquiry moves from prima-facie and all-out attitudes, as well as, obviously, moves from all-out to all-out attitudes in theoretical reasoning are guided by the ideal of truth, and, roughly, inferences are judged appropriate to the extent that they are truth-conducive.[28] To the extent that practical inquiry has a formal end, a similar thing can be said about it. Moves from prima-facie to all-out attitudes in practical reason are guided by the ideal of the pursuit of the good (assuming, again, that this is the formal end of practical reason), and inferences are judged appropriate to the extent that they are, in some sense, "good-conducive".

We can now refine the understanding of direction of fit proposed in the previous section: as we move from, say, a certain perception to a belief, insofar as what we are engaged in can count as theoretical inquiry, we

[27] See Wedgwood (2002) for a normative account of the relation between belief and truth.

[28] Or likely to be truth-preserving. Of course, one needs to make room for inferences that are not necessarily truth-preserving.

should be guided by the truth-conduciveness of the move. This will count as unsuccessful theoretical inquiry if the belief formed is not true or if the inference was not truth-conducive. On the other hand, when we form intentions on the basis of our desires, we should be guided by the "good-conduciveness" of the move; that is, by the fact that acting (or intending) on the basis of such desires counts as performing actions that are good (or, in other words, as acting well). This will count as unsuccessful practical inquiry if the action was not good (it would have been better to have acted differently), or if the inference was not good-conducive or warranted (the move from one's desires to the action was unwarranted). Ultimately the claim that belief and desire have different directions of fit is best understood as the claim that inferential moves in practical and theoretical inquiry are guided by distinctive formal ends.

One might object that while my description of the relation between belief and truth borders on triviality, the similar relation between intention or action and the good does not obtain. Many believe that one often acts without in any way pursuing what is good, and I have given no argument against their position. This is an important issue and I cannot do full justice to it here.[29] But I hope that the following remarks will show that this objection is not as worrisome as it might appear. First, just as we allowed that in the case of theoretical reason there might be belief formation that is not actually guided by truth, we could also allow that some actions are not guided by the good; perhaps, this is how one ought to understand *akrasia, accidie*, etc. However these actions would be, on the view proposed, in some way defective, by failing to conform to the formal end of practical reason, in the same way that a belief whose formation is not guided by the pursuit of the truth is defective as a piece of theoretical inquiry.[30] One might object here that one can act in a way that is unimpeachable and yet not in the pursuit of anything that one considers to be good. In fact, one might most fully identify with "perverse" pursuits, and feel "alienated" when one is pursuing something that one takes to be valuable.[31] Again, discussing this topic in any detail would lead us far astray. Obviously, if we accept that the good is the formal end of practical reason, we will doubt the coherence of this way of describing any piece of human behaviour. So these claims are best constructed as proposing that the good is not the formal end of

[29] For an extensive defense of the view that the good is the formal end of practical inquiry, see Tenenbaum (forthcoming (*a*)).

[30] Notice that although, as I pointed out above, it might be legitimate to form a belief on pragmatic grounds, it is hard to see how we can have a coherent view of an intention formed by any grounds that do not pertain to practical reason or practical inquiry.

[31] See Velleman (2000*b*, 2000*d*).

practical inquiry; indeed, philosophers who think that this kind of action is a real possibility typically think that something else is the formal end of agency.[32] But, as I said above, the argument of the paper does not depend on taking "the good" rather than "autonomy" or something else to be the formal end of practical reason.

4. Direction of Fit and the Humean Theory of Motivation

Let us distinguish two versions of cognitivism in ethics. Some cognitivist views might accept that the virtuous person's motivational state is *causally* related to a cognitive state, such as an evaluative belief, and deny that being properly motivated is a further cognitive achievement. The motivation itself is a blind disposition that may fail to be effected by the existence of the relevant evaluative belief, even if it is typically effected, or it is effected insofar as the agent is rational.[33] A more stringent form of cognitivism, however, would hold the view that the virtuous person is in a cognitively superior state than the vicious or the akratic person, and that the motivational state of the virtuous person is itself a cognitive state. I will call the latter view "motivational cognitivism"; according to motivational cognitivism, moral motivation stands to our rational powers as exercise to faculty. For motivational cognitivism, if an agent doesn't act as the virtuous agent would, then she cannot be credited with the same understanding of morality that the virtuous agent has. She might fail to have the same beliefs that the moral agents have, or her grasp of the content of the beliefs might be defective, or perhaps she does not fully understand the grounds for forming the relevant moral beliefs.[34] For the motivational cognitivist, differences in motivational states must be fully accounted by differences in cognitive states.

[32] I take it that this is a correct description of Velleman's position in the matter. For Velleman the constitutive end of action (and I take it, a fortiori of practical inquiry) is autonomy, not the good. See Velleman (2000*c*, 2000*d*).

[33] See Smith (1994). Even though Smith claims that an agent, insofar as she's rational, will have her motivation lined up with her values, what makes her a rational agent on Smith's view, as far as I can see, is simply the fact that this causal relation obtains. Smith insists that the agent who suffers from *accidie* or *akrasia* does not necessarily lack any kind of knowledge available to the virtuous agent. Rob Shaver (n.d.) argues that Sidgwick also held a view of this kind.

[34] The possibility that the difference between the virtuous agent and the non-virtuous agent lies in the grounds of their beliefs, instead of the beliefs themselves, is often strangely absent in discussions of the topic. Of course if only beliefs can ground beliefs, then one difference reduces to the other, but if I am right, an adequate analysis of direction of fit presupposes that this claim is false.

I'll assume that if the direction of fit metaphor lends support to the Humean theory of motivation, it must be at least be capable of showing that motivational cognitivism is untenable. One can at least say that if the Humean theory of motivation accommodates motivational cognitivism, there's nothing left distinctively "Humean" about it; there's no concession left to be made to those who stand on the other side of the fence, especially because anti-Humeans are often happy to grant that one can always ascribe some kind of desire to every case in which an agent acts for whatever reason.[35] Motivational cognitivists are often described as those who think that belief alone, and moral belief in particular, can motivate. This is certainly one way in which one could endorse motivational cognitivism: a belief about reasons for action in a certain situation, or an evaluative belief would by itself generate action. We can call this view "belief-based (BB) motivational cognitivism".

However, in focusing on the debate about the truth or falsity of moral judgments one overlooks another possible form of motivational cognitivism suggested by our discussion. Let us grant that, on the side of theoretical reason, the relation between prima-facie and all-out attitudes ought to be guided by the ideal of believing the true—according to inferential patterns dictated by this ideal—and that successful cognition will require a certain kind of non-accidental relation between those attitudes and what is actually true. One could adopt a parallel view about the nature of practical reason. One could say that the relation between prima-facie and all-out attitudes ought to be guided by the ideal of pursuing the good—according to inferential patterns dictated by this ideal—and that successful cognition requires a certain kind of non-accidental relation between those attitudes and what is actually good. If one adopts this latter view and if one thinks that moral action (necessarily) bears the right relation to the good, a relation parallel to the relation between knowledge and truth, one accepts a form of motivational cognitivism that is not committed to the view that beliefs can motivate by themselves. I take it that, for instance, Kant held this kind of motivational cognitivism. Kant maintained a sharp distinction between practical and theoretical reason,[36] taking them to be guided by different and irreducible ideals. Imperatives and maxims are our guides in acting. They're certainly not beliefs, and yet they can be cases of successful (or failed) cognition. As if the previous label weren't enough of a monstrosity, I'll call this view "non-belief-based motivational cognitivism" (NBB). It should be obvious that the relation between belief

[35] See McDowell (1998) on "consequential" desires, Platts (1997) on "trivial" desires, Schueler (1995) on "pro-attitudes", and on "motivated" desires (Nagel, 1970).

[36] See, for instance, Kant (1998: B830–1).

and truth, as I have presented it, poses no threat to NBB motivational cognitivism. After all, this form of cognitivism *advocates*, in this regard, a sharp separation between theoretical and practical reason. But can BB motivational cognitivists accept this analysis of direction of fit? After all, most arguments for the Humean Theory of Motivation target the idea that *belief* could motivate.

If BB motivational cognitivists are right, then some of our beliefs are not only theoretical attitudes but also practical attitudes. As such they would have to have both directions of fit at once. Arguments for the Humean theory of motivation based on the notion of direction of fit try to show that it is either incoherent or very implausible to think that the same attitude could have both directions of fit.[37] Now the Humean might start his argument by claiming that it is incoherent to have an attitude with both directions of fit towards the same content *p*. This claim does not directly contradict any kind of motivational cognitivism. The motivational cognitivist thinks that some beliefs with contents such as "it would be good to help the little child" or perhaps "the little child needs help (and nothing prevents me from helping her)" are inseparable from a motivation to *help the child*. No motivational cognitivist thinks that the moral agent rather has the absurd motivation to bring about the very content of these beliefs. But once we present the motivational cognitivist this way, we seem to provide the advocate of the Humean Theory of Motivation with a powerful argument against BB motivational cognitivism. The mental state that the BB motivational cognitivism postulates turns out to be a rather complex state; what Altham calls a "besire". These mental states are composed of two different contents and two different attitudes, corresponding to each direction of fit, for each of these contents. The agent is supposed, at the same time, to believe the content "it is good to help the child" and be disposed to bring about the content "I help the child". But if this is so, what could be the grounds for claiming that they are inseparable? Why couldn't an agent have one half of the besire but not the other? Moreover isn't this exactly what happens to certain agents, especially agents suffering from motivational disorders such as dejection, *accidie*, or depression? Don't they, say, continue to believe that it would be good to help the child, but fail to garner motivation, or at least sufficient motivation, to bring about that they help the child?

Our analysis of the notion of direction of fit should help us understand why we should not be persuaded by this argument against BB motivational

[37] I have in mind here in particular, Smith's arguments (1987; 1994: ch. 4). However, I am presenting the arguments in a slightly modified form.

cognitivism. It is worth first noting that our discussion suggests that the term "besire" is ambiguous; one could be advocating a view about the existence of any of the following:

(i) a mental state that is both a theoretical and practical all-out attitude;
(ii) a mental state that is both a theoretical and practical prima-facie attitude;
(iii) a mental state that is both an all-out theoretical attitude and a prima-facie practical attitude;
(iv) a mental state that is both an all-out practical attitude and a prima-facie theoretical attitude.

The above argument for the Humean Theory of Motivation is probably at its best when challenging the existence of attitudes described in (i). But if one wants to argue for the impossibility of any attitude that has multiple directions of fit, one has to show that all attitudes described in (i)–(iv) are incoherent; there can be no such "necessary union of direction of fit". However, as we look into all these possibilities, the prospects for making a case for the incoherence of any case of multiple directions of fit become quite dim. Let us look at an example of an attitude that seems to fall squarely into (iv). Intentions seem to be good candidates for being all-out practical attitudes.[38] Now it seems that forming an intention to ϕ serves as grounds for one's belief that one will ϕ,[39] and so here we have a case of (iv). Now various views on the nature of the relation between belief and intention might make it easier to accommodate the view that there are two separable mental states corresponding to the two directions of fit.[40] However it is hard to believe that general considerations about the formal ends of practical and theoretical inquiry should settle among our views about the nature of intentions. Similarly, let us look at the state of being in intense pain. Arguably, being in pain is not a representational state, and thus it does not have any direction of fit. However, it would be a respectable philosophical position to think it is constitutive of this state, at least in the case of human beings, that the following obtain: (*a*) the agent is at least inclined to believe that he is in pain; (*b*) the agent has some motivation

[38] I actually think that only intentions in action are all-out practical attitudes, but this does not affect the argument. See Tenenbaum (forthcoming (*b*)).

[39] Davidson (1980) famously argues that one can intend to ϕ without believing that one will ϕ. But my claim is much weaker. I only claim that intending to ϕ is inseparable from a prima-facie attitude to ϕ.

[40] Some people think that intentions simply are beliefs. See, for instance, Harman (1976). On this view, this is not going to be a case of (iv). But then one's denial of the existence of besires is hostage to a controversial view about the nature of intention. For various problems that the view that intention is belief faces, see Bratman (MS).

not to be in this state.[41] On this view, being in pain is a mental state of kind (ii). Now one can argue against (*a*) or (*b*) being constitutive of intense pain. But it would be bizarre to try to argue that *not both* (*a*) or (*b*) could be constitutive of pain solely on the grounds that theoretical and practical reason have different formal ends. But why would the situation be different with the motivational cognitivist? Why would it be possible to rule out in advance the possibility that an evaluative belief can also be a practical attitude? The postulation of mental states with multiple directions of fit is not an *ad hoc* maneuver on the part of motivational cognitivism; it is something that we might already be committed to in completely different contexts.

It is worth noting that even the most stringent form of motivational cognitivism needs to be committed only to the existence of states of kind (iii). And our analysis should make it clear that accepting states of kind (iii) does not amount to accepting an attitude that is somehow unique or extraordinary. It is easier to make both points if we start from the obvious fact that a belief can serve as evidence for another belief. Take, for instance, Anita's belief that the indentations in the sand that she's observing right now are tiger footprints. It thus appears to Anita that tigers have been around. Now one might suggest that in this case we have an all-out theoretical attitude that is also a prima-facie theoretical attitude of a different content. The belief "the indentations in the sand are tiger footprints" and the "appearance" with the content "tigers have been around", one might argue, are one and the same state. Opposing this suggestion, one might argue that we should keep the two states apart; one might want to insist that it is at least conceptually possible that one forms the belief without having any attitude, prima-facie or all-out, with the content "tigers have been around". I must confess I find it hard to wrap my mind around the idea that this is indeed a conceptual possibility.[42] I don't see how one can have a full grasp of the content "the indentations in the sand are tiger footprints", have a belief with this content, and yet not have it at least *appear* to him that tigers have been around, given the close conceptual connection between "x is a tiger footprint" and "x is the effect of a tiger's paw making contact with the surface". But if one wants to insist that the separation is

[41] Christine Korsgaard's view on the nature of pain, although different from the view described here, does seem to incorporate motivational and cognitive elements as constitutive of pain itself. See Korsgaard (1996: lecture 4).

[42] I am ignoring an irksome complication. One could produce a footprint in the absence of tigers; one could press a severed tiger paw against the sand. Perhaps someone who sees this footprint knowing how it is produced doesn't take this to be any kind of (overridden) evidence that tigers have been around. However, one could complicate the belief so as to rule out this possibility.

conceptually possible, one can just replace this example with one of a closer conceptual connection between the content of the belief and the content of the appearance. Perhaps "John has a sunburn" for the belief, and "John was exposed to the sun" for the appearance. At any rate, it'll be hard to argue against the following general claim of conceptual connection:

> (CC) For some distinct contents X and Y, if a subject S fully grasps X and Y, then it is necessarily the case that if S believes X then it appears to S that Y (S has a prima-facie theoretical attitude with content Y).

I know of no general reason to think that one can rule out that at least some belief states stand in this kind of relation to other beliefs states for which they are evidence. It is also important to note that nothing I said above rules out the possibility that the appearance is conclusive. By a "conclusive appearance", I mean something along the lines of "providing obviously conclusive evidence"; if one has something that counts as obviously[43] conclusive evidence for *p*, and one understands the evidence, and that it is conclusive evidence for *p*, arguably, one necessarily forms the belief that *p*. Similarly, if someone has a prima-facie attitude of content *p* that is (obviously) conclusive, one will necessary form the belief that *p*. One might argue that the belief that John has a sunburn doesn't imply only that it appears that John has been exposed to the sun, but, in fact, the appearance in question leaves no room to doubt that John has been exposed to the sun; in this case, once one believes that John has a sunburn one cannot stop short of the belief that John has been exposed to the sun. Again, here one might think that this is not true for this example, and one might doubt whether it is true for any example. All that I want to note at this point is that one cannot rule out in advance the possibility that having a belief state that X will imply having a belief state in which it appears conclusively that Y.

Now one might say that (CC) does not imply that the belief and the appearance are one and the same mental state; for some reason, one might want to say that they are two states such that one could not be in the former without being in the latter. This might be a plausible move, and since for our purposes this does not make much difference whether the move is made or not, I'll just talk about one state conceptually implying the other, without prejudging whether we have one or two mental states.

[43] This qualifier should make the demand much weaker than a demand for closure. Smith (forthcoming) suggests that the motivational cognitivist is committed to accepting deductive closure. But I hope it'll be clear that motivational cognitivism is not committed to anything that strong.

If one grants (CC), one grants that an all-out theoretical attitude can entail a prima-facie theoretical attitude. As we saw above, we know that it is possible that some practical attitudes imply the existence of some theoretical attitudes. What reason can we have now to deny that a certain all-out theoretical attitude could entail certain prima-facie practical attitudes? Why couldn't the content of the all-out theoretical attitudes of a virtuous person be such as to imply a certain prima-facie practical attitude? That is, why couldn't the relation between the beliefs of the virtuous person and the desire to act in certain way be just the same as the relation between S's belief that John has a sunburn and the fact that it appears to S that John was exposed to the sun? After all, the BB motivational cognitivist need be committed to nothing more than the claim that having the kinds of beliefs that the virtuous agent has will necessarily motivate. A relatively weak version of motivational cognitivism need not say that moral beliefs necessarily lead to action. But in fact, there is no reason to think, purely on the grounds of the nature of these attitudes, that theoretical all-out attitudes could not entail conclusive appearances in the practical realm. If one cannot rule out the existence of these relations within the theoretical realm, why should it be impossible that a similar relation obtain across realms? An example might help make out this point. Suppose one thinks that "John has a sunburn" conceptually entails a conclusive appearance to the effect that John was exposed to the sun. It is now the case that I cannot attribute to Larry a belief, or at least a non-defective belief, with the content "John has a sunburn", unless I am prepared to attribute to him also the belief with the content "John was exposed to the sun". But if this is so in the case of the relation among beliefs, what reasons do we have to rule out the possibility that certain beliefs can be attributed to the agent only if he is prepared to act in certain ways (or form certain intentions)? This still falls short of a commitment to (i), since the moral belief would probably not suffice to give rise to a full-blown all-out attitude; it would probably lack content to specify in detail the actual intention with which the agent acted. However even the most radical motivational cognitivist need not be committed to anything stronger.[44]

One could insist that it is simply implausible to suppose that certain beliefs are capable of inclining the agent to pursue anything. But this is not an *argument* for the Humean Theory of Motivation; it *is* the Humean Theory of Motivation. More plausibly, one can think that states such as *accidie* or depression speak against the fact that moral beliefs can be conceptually connected to the relevant practical attitudes. After all, the

[44] More on this issue in the next section.

agent who suffers from these ills might have the exact same belief as the virtuous agent. The depressed agent, just like the virtuous agent, could believe that it would be very good indeed to help the poor, but just fail to garner the motivation to do it. It would be *ad hoc*, the Humean may say, to deny that the depressed agent has the same belief just because he fails to act in the same way.

But is it *ad hoc*? Our above discussion should suggest that the answer is "no". The defining thesis of BB motivational cognitivism is the claim that the very fact that motivation is not present is what *makes* it the case that we cannot attribute the full-blown moral belief to the agent in question; motivational cognitivism is not committed to the claim that for every case that the motivation is absent we will have an independent reason not to ascribe the full-blown moral belief to the agent. Of course if the central argument of moral cognitivism were the claim that all those who honestly assent to moral claims behave morally, cases of *accidie* and depression would present a serious challenge to the view. But no motivational cognitivist would defend her view in this manner. Motivational cognitivism takes as its starting point the attractiveness of a picture of morality in which moral activity *is* a form of knowledge.[45] So the motivational cognitivist is committed to seeing those motivational failures as in themselves failures to fully grasp the content of one's moral beliefs, or somehow failing to have the same kind of moral beliefs as the moral agent. Of course, one can dispute these claims. But just as in the case of intention or of being in intense pain, what settles the debate is who provides us with the best conception of the virtuous agent, not considerations about the formal ends of theoretical and practical reason.

One should point out that the BB motivational cognitivist is often saddled with a "molecularist" picture that makes her view seems particularly implausible. The BB motivational cognitivist does not need to claim that the difference between the virtuous agent and the one suffering from *accidie* must be present in *each* belief, considered on its own, that fails to motivate the dejected agent.[46] The motivational cognitivist is not committed to the claim that, when the agent suffering from *accidie* says "I should not be just lying in bed", there is really some part of "not" that he doesn't understand. The BB motivational cognitivist thinks that full understanding of the moral

[45] This is true both of historical figures and contemporary philosophers. Kant says that wisdom (*Weisheit*) is primarily a matter of acting. See Kant (1998*b*: G 405). Among contemporary philosophers, John McDowell (1998) explicitly presents the claim that virtue is knowledge as a motivation for his view.

[46] This molecularist interpretation of the BB motivational cognitivist is certainly encouraged by characterizing the position as one that accepts the existence of besires.

facts ensures actions, but the lack of understanding need not be attributed to a belief considered in isolation. In considering the differences between fully virtuous agents and all sorts of other agents, within the confines of BB motivational cognitivism, we can appeal not only to differences in the contents of their beliefs, but also to differences in the relevant prima-facie attitudes, in how they jointly ground the belief in question, in how they cancel other prima-facie attitudes that seem to undermine the belief, etc. We can see now that the BB motivational cognitivist is committed to something weaker than what we've been suggesting; all she needs to accept is that there are some "packages" of all-out attitudes grounded on certain prima-facie attitudes, such that full understanding of how the whole package hangs together is conceptually connected to a prima-facie practical attitude. It is not implausible to think that there is *something* in this package that distinguishes the virtuous agent from, say, the dejected agent. Considerations of the different directions of fit of belief and desire certainly can give us no reason to be suspicious of this commitment; in fact the considerations show that similar relations hold in other domains.

5. The Detective and the Shopper

Anscombe's example seems to suggest a sharp, independently conceived distinction between the two directions of fit, a distinction that does not seem to be captured by the idea that practical and theoretical reason might have different formal ends. And one might suspect that we failed to find an argument for the Humean Theory of Motivation simply because we failed to capture something important in Anscombe's example. In this section, I'll try to lay this suspicion to rest. Let us go back to Anscombe's example. Let us call the detective Jenny and the shopper Leo, and suppose that that's how things look like:

Leo's List	*Shopping Cart*	*Jenny's List*
Grapes	Cherries	Cherries
Apples	Apples	Plums

We can describe what goes wrong with Leo as follows:

(1) Leo wants to buy grapes.
(2) Leo buys cherries.

On the other hand, we can describe what goes wrong with Jenny as follows:

(1) Jenny believes there are plums in Leo's shopping cart.
(2) There are no plums in Leo's shopping cart (there are apples).

We can notice a few things now. First, Leo's mistake can be characterized as an inferential one[47] in a broad sense of 'inferential'. The mistake was moving from an attitude (a desire) to another attitude (acting with an intention) for which the first was supposed to provide grounds. Also, by saying that this was an inferential mistake, I'm not claiming that Leo was irrational, or that his inferential patterns are blameworthy. All that I am claiming is that he moved from an attitude that was unimpeachable to one that was not. If the pattern of inference cannot guarantee that one always move from unimpeachable attitudes to other unimpeachable attitudes, the agent might arrive at mistakes while being perfectly rational. Acting with an intention is not typically characterized as an attitude. However, whether we can count it as an attitude or not is not particularly important for my purposes, as long as one grants that one does things on the grounds of certain desires or intentions; what I've been characterizing as an inferential relation is just the grounding relation between the desire or intention and the action. We can present the inferential relation in our example as follows:

(3) Leo wants to buy grapes.
(4) Leo acts in such a way as to bring it about that these fruits are in the shopping cart.

Note that, if this is the correct characterization of the inferential relation, Leo's mistake is one that can be located in his moving from a prima-facie attitude to an all-out one, and from a desire with a general content to a particular action.

But note that we can also characterize the detective's mistake as an inferential one in this broad sense of 'inferential',[48] in the move from (5) to (6) below:

(5) It appears (perceptually) to Jenny that these fruits are in the shopping cart.[49]
(6) Jenny believes that there are plums in the shopping cart.

Once we think about the differences in these terms, Anscombe's case is also a case in which the difference between the practical and theoretical cases is a difference between the different formal ends that guide the moves from

[47] Someone might protest that this is a mistake in *performance* not an inferential one. I come back to this point in a moment.

[48] This characterization does not rule out the possibility that the belief is 'non-inferential' in a narrower sense of 'inferential'; that is, it is not inferred from other *beliefs*.

[49] I don't mean to imply that content of perceptual experience must be conceptual. It doesn't matter for my purposes if the inference starts from conceptual or non-conceptual content.

prima-facie to all-out attitudes. However it is hard to shake the feeling that there is something different here, that we have not captured the idea that Leo's mistake must be located in how the agent changes the world, rather than in how the agent changes his mind. It is easy to try to dismiss the difference as just the result of the fact that practical reason concerns action; hardly something that any philosopher has missed. Yet we must acknowledge that there is an important disanalogy between theoretical reason and practical reason that comes up in this example that our analysis of direction fails to capture. The move from the general to the particular in theoretical reason in forming a judgment tends to be trivial. Although subsuming a particular under a concept and forming judgments of the form *Fa* is in no way trivial, moving from judgments of the form *(x)Fx* to judgments of the form *Fa* certainly is. That is, leaving aside very complexly formed predicates and other complications, if my judgment that *(x)Fx* is correct, there won't be much room for mistake in moving from the general judgment to the particular one. However the same is not true in the realm of practical reason. Of course, there's no agreed-upon equivalent of universal instantiation in the realm of practical reason. But without trying to work out the details of this proposal, we can think of the move from the general intention to a particular action as a similar move. One example of this kind of inference would be the following: I infer from my thought that, all things considered, actions in which I pause for a moment and draw a circle in the air are desirable to my acting so as to bring it about that I'm drawing a circle in the air right now in a particular way. However, unlike the case of theoretical reason, this move is in no way trivial, for despite the simplicity of the predicate "drawing of a circle", there's no guarantee that I'll succeed in actually drawing a circle in the air; in fact, I'll probably fail. Practical reason allows for mistakes of *performance*,[50] mistakes in trying to execute a flawlessly formed, simple intention.

Arguably all-out attitudes of practical reasons are always particular judgments; they are cases of acting with an intention as described in statements such as (4). Given that an intention does not by itself determine how it will be carried out, an intention that is not an intention in action will always leave room for revision as one tries to carry out the intention in concrete actions. Insofar as practical reasoning aims to issue in some kind of action, forming a general intention is still being in a state that falls short of being an all-out attitude. Any such general intention must have some *ceteris paribus* conditions that could fail to obtain, and thus fail to be the agent's final view about how she shall or should act. Therefore such states are not

[50] This is how Anscombe herself (1963) identifies the mistake of the shopper.

all-out attitudes. If one does not want to go all the way to the Aristotelian view that the conclusion of practical reason is an action, one will need at least to say that the conclusion is a decision to engage in *this particular action*.[51] If this is correct, mistakes of performance are failures that can be coherently ascribed to *any* all-out practical attitude, but to no theoretical attitude. In sum, the difference between the shopper and the detective is best characterized as follows: on the one hand, the shopper makes a mistake in making an inference guided by the formal end of practical reason from a general, prima-facie attitude to a particular, all-out one. On the other hand the detective makes a mistake in making an inference guided by the formal end of theoretical reason from a particular (prima-facie) attitude to a general (all-out) one. Because the move from the general to the particular in the practical realm is non-trivial, all-out practical attitudes are always liable to mistakes of performance.

Now if all mistakes in practical reason were mistakes of performance, we would have an argument for the Humean Theory of Practical Reasons that could probably ground an argument for the Humean Theory of Motivation; after all, the best candidates for non-desire-based reasons are general in character. But this view is obviously false; Leo might be mistaken not only in placing the wrong fruits in the cart, but also in his general intention to see to it that there are grapes in the cart. But couldn't we generate an argument against BB moral cognitivism from the fact that the conclusion of practical reason must be particular in character? After all, moral beliefs are general in character and if one cannot act without making a non-trivial move from the general to the particular, one needs something beyond moral belief to be motivated to engage in any particular action. However this gives us no reason to reject the view that these beliefs can motivate one to act in a particular way in accordance to a general intention. The most we could rule out is that an action could be solely motivated by a moral general belief. This would not necessarily be a problem for the motivational cognitivist. Think for instance about a principle of beneficence such as:

(B) One ought to help others.

Now assume that an agent finds herself in a situation where she could help someone out of the subway (suppose there's a wide gap between the door and the platform). She can do this by either giving the passenger a hand, or by lying down, head inside the train and feet in the platform, so that the passenger could walk over her back in a mildly painful way (she has a

[51] This is admittedly just a sketch of an argument for these claims. I provide more detailed argument in Tenenbaum (forthcoming (*b*)).

strong back, and the passenger is pretty light). I assume that the latter way of helping is, albeit awkward, morally permissible. Thus the following is arguably a consequence of (B):

> The action of helping the passenger by lying between the platform and the train is prima-facie good.

Nonetheless one would not necessarily conclude that such an action was correct or justified.[52] In general, what this shows is that, even in the case of perfectly virtuous action, moral belief alone cannot explain *every single aspect* of the action. But of course no sane form of motivational cognitivism should be committed to the opposing view.

Many attempts have been made to use the notion of direction of fit to expose significant differences between beliefs and desires, or to reveal a deep dissimilarity between theoretical inquiry and intentional action. I have been arguing that direction of fit does not lend itself well to these purposes, and in particular, that it does not lend support to the Humean Theory of Motivation. Surprisingly, attempts to render the notion of direction of fit more precise suggest a picture of reason in which there is in fact a deep similarity between the realm of practical reason and intentional action on the one hand, and the realm of theoretical reason and belief on the other.[53] This is the picture of a natural home for motivational cognitivism, a view in which one employs the same sort of rational faculties, albeit in relation to two different formal ends, in theoretical and in practical reason. Of course, I do not want to let the pendulum swing to the opposite error and argue that my reconceived account of direction of fit can prove the truth of motivational cognitivism; what I intend to convey here is just the sense that this notion may furnish valuable materials for rendering the view more plausible and precise.

REFERENCES

Altham, J. E. J., 'The Legacy of Emotivism', in G. Macdonald and C. Wright (eds.), *Fact, Science and Morality* (Oxford: Blackwell, 1986), 275–88.

Anscombe, G. E. M., *Intention*, 2nd edn. (Ithaca, NY: Cornell University Press, 1963).

[52] Of course what counts as "justified" and "correct" in practical reason, if anything, is the subject of much dispute, but the motivational cognitivist is bound to accept that some actions are justified or correct and others not. No matter how she does it, a reasonable motivational cognitivist should allow that this action might not be justified or correct.

[53] This picture of practical reason and intentional action is developed in much more detail in Tenenbaum (forthcoming (*a*)).

Bratman, Michael, 'Intention, Belief, Practical, Theoretical', unpubl. MS.

Clark, Philip, 'Velleman's Autonomism', *Ethics*, 111 (2001), 580–93.

Davidson, Donald, 'Intending', in *Essays on Actions and Events* (New York: Oxford University Press, 1980).

Humberstone, I. L., 'Direction of Fit', *Mind: A Quarterly Review of Philosophy*, 101 (1992), 59–83.

Kant, Immanuel, *The Critique of Pure Reason*, tr. Paul Guyer and Allen Wood (Cambridge: Cambridge University Press, 1998a).

—— *Groundwork of the Metaphysics of Morals*, tr. Mary Gregor (Cambridge, Cambridge University Press, 1998b).

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

McDowell, John, 'Virtue and Reason', in *Mind, Value, Reality* (Cambridge, Mass.: Harvard University Press, 1998).

Nagel, Thomas, *The Possibility of Altruism* (Princeton: Princeton University Press, 1970)

Platts, Mark, *Ways of Meaning* (London: Routledge & Kegan Paul, 1979).

Railton, Peter, 'On the Hypothetical and Non-Hypothetical in Reasoning about Belief and Action', in G. Cullity and B. Gaut, *Ethics and Practical Reason* (Oxford: Oxford University Press, 1997).

Schueler, G. F., 'Pro-Attitudes and Direction of Fit', *Mind*, 100 (1991), 277–81.

—— *Desire: Its Role in Practical Reason and the Explanation of Action* (Cambridge, Mass.: MIT Press, 1995).

—— *Reasons and Purposes: Human Rationality and the Teleological Explanation of Action* (New York: Oxford University Press, 2003).

Shah, Nishi, 'How the Truth Governs Belief ', *Philosophical Review*, 112 (2003), 447–82.

Shaver, Rob, 'Sidgwick on Moral Motivation', unpub. MS.

Smith, Michael, 'The Humean Theory of Motivation', *Mind*, 96 (1987), 36–61.

—— *The Moral Problem* (Oxford: Blackwell, 1994).

—— 'Is there a Nexus between Reason and Rationality?', in Sergio Tenenbaum (ed.), *New Essays in Moral Psychology* (Amsterdam: Rodopi, forthcoming)

Sobel, David, and Copp, David, 'Against Direction of Fit Accounts of Belief and Desire', *Analysis*, 61 (2001), 44–53.

Tenenbaum, Sergio, 'The Judgment of a Weak Will', *Philosophy and Phenomenological Research*, 49 (1999), 875–911.

—— *Appearances of the Good* (Cambridge: Cambridge University Press, forthcoming (*a*)).

—— 'The Conclusion of Practical Reason', in Tenenbaum (ed.), *New Essays in Moral Psychology* (Amsterdam: Rodopi, forthcoming (*b*)).

Van Fraassen, Bas C., *The Scientific Image* (Oxford: Oxford University Press, 1980).

Velleman, David, *The Possibility of Practical Reason* (New York: Oxford University Press, 2000a).

—— (2000b) 'The Guise of The Good', in Velleman (2000*a*).

—— (2000c) 'The Possibility of Practical Reason', in Velleman (2000*a*).

Velleman, David (2000d) ‘What Happens When Someone Acts’, in Velleman (2000*a*).

Wedgwood, Ralph, ‘The Aim of Belief ’, *Noûs-Supplement: Philosophical Perspectives*, 16 (2002), 267–97.

Williams, Bernard, ‘Deciding to Believe’, in *Problems of the Self* (Cambridge: Cambridge University Press, 1973), 136–51.

Zangwill, Nick, ‘Direction of Fit and Normative Functionalism’, *Philosophical Studies*, 91 (1998), 173–203.

10

Misunderstanding Metaethics: Korsgaard's Rejection of Realism

Nadeem J. Z. Hussain and Nishi Shah

1. Introduction

Contemporary Kantianism in ethics is often thought of not just as a position within normative ethics but also as an alternative to moral realism. We argue that it is in fact not at all clear how contemporary Kantianism can distinguish itself from moral realism. There are of course many Kantian positions. For reasons of space we have chosen to focus here on the position of one of the most prominent, contemporary Kantians, Christine Korsgaard. Officially our discussion is restricted to her version of Kantianism, though we suspect that the lessons learnt here apply elsewhere.

In our experience, it immediately strikes some as implausible that Korsgaard is actually engaged in metaethics. We grant that there are strains in Korsgaard that suggest an attempt to, so to speak, go "beyond" metaethics. We take up such a reading of Korsgaard elsewhere (Hussain and Shah, 2005*b*). Here we simply accept at face value the way in which she repeatedly introduces the Kantian view as an alternative to realism. Crucially, she emphasizes that the realism of concern to her is, as she puts it, "*substantive* moral realism"—that is, a view with specific metaphysical, epistemological, and semantic commitments. It is "the view that there are answers to moral questions *because* there are moral facts or truths, which

We would like to thank Allan Gibbard, Pamela Hieronymi, Elijah Millgram, Tamar Schapiro, Yonatan Shemmer, Kenneth Stalzer, and Sharon Street for very useful conversations about the issues raised in this paper. We would also like to thank the participants of the Annual Metaethics Workshop, University of Wisconsin-Madison, 2004, for their questions and comments. Thanks to Sarah Buss, Alexander George, Kieran Setiya, David Velleman, Allen Wood, and an anonymous referee for written comments on an earlier draft of this paper.

those questions ask *about*" (Korsgaard, 1996: 35).[1] According to realism, moral requirements must be given "some sort of ontological foundation, by positing the existence of certain normative facts or entities to which moral requirements somehow refer" (Korsgaard, 1997: 218). Not surprisingly she does not want to contrast her own position with uses of the word "realism" that merely mark out a contrast with nihilism—realism merely as the general normative view, that is, that there are correct answers to questions about what we should do (Korsgaard, 1996: 35). For the purposes of this paper we take this as sufficient evidence that (i) she is contrasting Kantianism with the metaethical position of realism and that (ii) she takes Kantianism to be the philosophically favoured position of the two. Our claim is that she fails to show either that Kantianism is different or that it is better than realism.

2. The Normative Question(s)

Our general strategy will be to argue that what are supposed to be claims that conflict with realism in fact fail to do so. We will rarely attack the arguments for these claims. What we will attack instead is the argument against realism based on these claims. These claims (and the arguments for them) fail, in general, to undermine realism because Korsgaard fails to show that they actually conflict with realism in the first place. They often fail to conflict because though they may appear to be metaethical claims they in fact are not obviously so and indeed are most charitably interpreted as either claims within normative ethics or normative psychological claims in the philosophy of action, claims compatible with several different metaethical accounts of those same claims including non-reductive normative realism.

We will argue therefore that what explains the failure to distinguish Kantianism from realism is a failure to appreciate all the consequences of the traditional distinction between normative judgements and metaethical interpretations of normative judgements.[2] Thus we begin with a brief review of the differences between normative ethics and metaethics. Within normative ethics, we can distinguish at least two different philosophical tasks. The first is to construct a set of principles that systematize and

[1] Emphases in original.

[2] As we have already noted, one can read Korsgaard as intending to undermine the distinction between normative ethics and metaethics. We take it though that the distinction is supposed to be undermined in part as a consequence of her arguments against realism (and other metaethical views). An argument for the claim that her view is different from, and better than, realism cannot simply presuppose that the distinction has been undermined. For further discussion, see Hussain and Shah (2005*b*).

ground our correct moral judgements. Utilitarianism and various forms of deontology are examples of such theories, expressing competing conceptions of the fundamental moral principle(s) from which correct judgements of moral rightness and wrongness can be derived.[3]

The second task is to place morality within practical reason, explaining whether we have reason to do what morality demands and, if so, whether these reasons are derived from another branch of practical reason. There are two ways of carrying out this task. One is to argue that it follows from the concept of a reason for action or agency or well-being that an agent always has reason to do what is right and avoid doing what is wrong. The debate about the conceptual possibility of the amoralist (someone who judges an action would be right but sees no reason to do it) is about the success of this strategy of placing morality within practical reason. While many philosophers would label this a debate within metaethics, we place it under the heading 'normative ethics' in order to mark the fact that, whichever side one takes, one will not yet have answered the questions that Korsgaard's stated adversary, the realist, attempts to answer. Recall that Korsgaard's "*substantive* moral realism" is a position with specific metaphysical, epistemological, and semantic commitments. Realism is not a position about the relation between some normative concepts (for example, "rightness") and other normative concepts (for example, "reason for action"), but is a position about the nature of normative concepts in general.

The other way of placing morality within practical reason is to show that moral requirements follow from a substantive conception of practical reasons. How one carries out this strategy depends upon one's conception of practical reason. If one thinks that the aim of practical reason is to maximize an agent's desire-satisfaction, then placing morality within practical reason will entail showing that doing what morality demands maximizes the satisfaction of an agent's desires. But if one has some other conception of practical reason, then showing that morality satisfies desires may be beside the point; instead one might be faced with the task of showing that the demands of morality can be derived from something else, for example, the principles of autonomy. Or, there may be no need to show that morality can be derived from anything at all, if according to one's conception of practical reason, the principles of morality are fundamental principles of practical reason.[4]

[3] There are many options here, e.g. upon investigation, one might conclude that there are no deep, exceptionless moral principles (see Aristotelian theories), and that the best we can do is arrive at more-or-less useful rules of thumb.

[4] Note that in this context the claim that moral principles are fundamental principles of practical reason is a substantive claim about the correct conception of practical reason,

A metaethical account offers an interpretation of the normative claims that are offered as answers to these inquiries (for example, that it is morally right to maximize utility or that one has reason to do those actions that are morally right), aiming to tell us what these claims mean, whether they involve metaphysical commitments, if so, what these commitments are, and whether and how we acquire knowledge of these normative claims. Non-reductive realism and non-cognitivism are examples of positions that give competing answers to these questions. The non-reductive realist usually subscribes to a referential semantics (the judgement that x is good expresses a belief that x has a normative property), an ontology of non-natural properties (normative properties are non-natural properties), and an intuitionist epistemology (we come to know basic normative truths by non-sensory, rational intuition). Non-cognitivists, on the other hand, usually reject a referential semantics for moral terms. They claim that moral judgements do not express truth-evaluable beliefs in normative facts, but express non-depictive motivational states such as desires, preferences, or emotions. Non-cognitivists are therefore free to accept an ontology restricted to natural properties and to deny that there is an epistemology needed for moral judgements, since moral judgements, being non-depictive and therefore not truth-evaluable, do not aspire to knowledge.[5]

We do count views that argue that there is, in some sense, no way of getting outside of normative thought to explain it, and that therefore no answers to these questions are possible, as doing metaethics. However, this type of quietism, which claims that no metaethical theories are possible, is not equivalent to merely failing to state a metaethical position. One might pursue normative ethical tasks while ignoring metaethical ones, leaving such questions for others to answer. This acceptance of a division of philosophical labour certainly would not commit one to the quietist claim that metaethics is impossible. Quietism is a bold position in need of justification, whereas the decision to pursue normative ethical questions instead of metaethical ones needs no philosophical defence.[6] The point of metaethics is to give an

not an analytic claim about the relation between the concept of a reason for action and the concept of moral rightness. Our thanks to an anonymous referee for *Oxford Studies in Metaethics* for encouraging us to be clearer about the differences between normative ethics and metaethics. For further clarification, see Hussain and Shah (2005*b*).

[5] We simplify; the non-cognitivist has to give us some account of what we are doing when we claim that we know that murder is wrong.

[6] It is perhaps only relatively recently, *c*. 1903, that philosophers have self-consciously isolated and pursued specifically metaethical questions. Thus, when interpreting ethicists in the history of philosophy who did not explicitly distinguish these questions, we must be very careful not to assume that because of the metaphysical or epistemological sounding labels used to express their views that they are always making metaethical

account of what it is to think a normative thought, or to show that such an account is impossible, not to tell us which normative thoughts to think or to point out which normative thoughts we cannot help but think.

Even in this traditional form of the distinction, normative ethics and metaethics are not completely independent of each other. Since metaethics is an attempt to provide an interpretation of our normative practice, which metaethical theory we end up with will in part be determined by what we think our practices of making normative judgements look like. Furthermore, normative ethics may lead us to think that certain moral claims are true. Ascribing error has its costs and so metaethical theories that allow these judgements to be true will have a defeasible advantage. Similarly certain metaethical theories, reductive realism for example, will entail particular normative claims. One cannot claim that 'right' just means "maximizes utility" without its following that if an action maximizes utility, then it is right.

With this traditional distinction in hand, we turn in the next section to the task of trying to identify Korsgaard's Kantian position by focussing on her insistence on distinguishing her own view from a position she labels 'realism' or 'dogmatic rationalism'. We take the target here to be non-reductive normative realism and ask whether her rejection of non-reductive realism might give us insight into her alternative Kantian view. Our claim is that her central objection to the non-reductive realist reveals the above-mentioned failure to distinguish between different questions about normativity and that the non-reductive realist has a coherent response. This failure to distinguish answers to normative questions from answers to metaethical questions also undermines Korsgaard's attempt to show that her own position is an alternative to non-reductive realism.

We then take a detailed look at her account of instrumental reason. We carefully assess her account of the will, the idea of a constitutive norm, and the role of self-legislation to see if we can identify a positive Kantian position that can respond to the worries about non-reductive realism raised by both her and others. We conclude, however, that it is in fact very hard to see how Kantianism about instrumental reason could represent a position distinct from non-reductive realism.

We finish by assessing whether Korsgaard's constructivism and its account of normative concepts succeeds, as it is apparently supposed to, in establishing an alternative to non-reductive realism. We conclude that, in its

claims. This is why we think it best to avoid simply using the historical labels, such as 'rationalism', 'empiricism', and even 'voluntarism', that Korsgaard uses to describe various ethical theories, as these labels often stand for historical theories in ethics that ran together positions in normative ethics and metaethics.

currently undeveloped state, it does not and that thus in the end Korsgaard leaves us with no distinctive Kantian alternative to non-reductive realism.

3. Sources of Normativity

3.1. *Sources of Normativity*

In this section we argue that in Korsgaard's attempt to delineate "the normative question" in *The Sources of Normativity* (1996) she fails to distinguish the task of placing normative principles and judgements within practical reason from the task of giving a metaethical account of those principles and judgements. This causes her to misunderstand the aims of Prichard and Moore's metaethical views and to reject them on the spurious grounds that they fail to answer questions within normative ethics. We then argue that her own solution to the "normative problem" is infected by this ambiguity, and thus fails to express a distinctive metaethical view, much less one that contrasts with the non-reductive realist views of Moore and Prichard that she rejects.

The failure to distinguish normative from metaethical questions is reflected in a potential ambiguity in Korsgaard's claims to have identified the "source of normativity" or to have "explained normativity". There is a distinction between what *makes* an action wrong or a principle normative, on the one hand, and what *constitutes* the normativity or what the property of being normative itself is, on the other. Thus the fact that brushing my teeth regularly will reduce plaque may *make* brushing my teeth good (for me); however, we do not want to claim, presumably, that the property of goodness itself just is the property of reducing plaque.[7] The ambiguity mentioned above can now easily be seen. There are perfectly understandable senses of these expressions according to which one might well say that one is picking out the source of the normativity of teeth brushing—or explaining why one ought to brush one's teeth—by pointing out that the brushing of teeth reduces plaque. But these claims are best understood as first-order normative judgements about what makes brushing one's teeth good, not as

[7] We are not claiming that the sense of expressions of the form "*make x* wrong" that we are trying to pick out and use here is exhaustive or even central to the ordinary language uses of such phrases. The hope is to use this phrase essentially as a term of art—a now almost standard one—to help keep track of an important philosophical distinction. Note, we do not deny that some such identification of wrongness with what makes things wrong is part of the strategy of certain realists. Our assumption of course is that such an identification is not likely to be part of any metaethical strategy that would be recognizably Kantian. More on this below.

providing a metaethical interpretation of what it means to say that reducing plaque is good, what metaphysical commitments such a judgement involves, or how we come to know that brushing one's teeth is good.

In *Sources*, Korsgaard says that in seeking a philosophical foundation for morality we are not looking for a mere explanation of morality, but for a justification of the claims that morality makes on us (1996: 9; also 16). She says that giving an adequate third-personal explanation of morality, such as might be given by an evolutionary account of morality according to which morally right actions are those that promote the preservation of the species, would not answer "the normative question" because it would fail to justify morality from within the first-personal point of view (14). This suggests that she is seeking to place the claims of morality within practical reason. The normative position that claims that right actions are those that promote evolutionary fitness would be a failed attempt to place the claims of morality since it would not show moral claims to be justified from within practical reason. That some action would promote fitness does not seem at all like a reason to do that action. However, such a position would be an example of a failed theory in normative ethics rather than a metaethical theory.[8]

Korsgaard's description of the "substantive realist" answer to the normative question, however, depicts it as a metaethical position. Then again, the main criticisms that she makes of "substantive realism" seem to presuppose that it is meant to answer a normative question within practical reason. For example, in her discussion of Prichard's response to the question "Why should I do my duty?" Korsgaard assumes that Prichard's answer commits him to the view that moral claims refer to a realm of non-natural, normative properties (1996: 32). But she fails to distinguish this metaethical thesis about the metaphysical commitments of moral judgements from Prichard's normative thesis that moral reasons are foundational or underived. His response to the "why be moral?" question commits him to the latter thesis, not the former. Briefly, his reply is that the question is "improper" or "illegitimate" (Prichard, 2002*a*: 7, 19), because either it is asking for a self-interested reason to do one's duty, in which case it is seeking the wrong

[8] But in a footnote (14) Korsgaard claims that the evolutionary theory reduces normative ideas to natural ones. This suggests that she is interpreting the evolutionary view to be a reductive account of the meaning of 'moral rightness' or a metaphysical reduction of moral rightness to evolutionary utility, rather than a normative account of the right-making property. Our point is not to suggest that a crude evolutionary account of morality escapes Korsgaard's criticisms, but rather that in her discussion of such an account Korsgaard fails to distinguish the metaethical and normative ethical interpretations of such a position, and this suggests to us that her "normative question" itself blurs metaethical questions of the semantics and metaphysics of moral claims and the normative question of how morality fits into practical reason.

kind of justification of morality since morality's claims are unconditional, or it is seeking a moral reason to do one's duty, in which case it is presupposing the very thing it is asking for. Thus, for Prichard, moral reasons are foundational within practical reason, and do not need to be derived from other practical reasons. Of course, if one holds this position, then, in one sense, placing morality within practical reason will be trivial. This is not to suggest that substantive work will not remain. We have to be convinced that the Prichardian is indeed right about the foundational nature of moral reasons; our brief summary of his position is not a complete presentation of his arguments for this conclusion.[9] Furthermore, the normative ethical task of showing which actions are morally right or morally wrong remains.

There are, no doubt, many objections that one might raise to Prichard's response. Instead of illuminating the status of morality within practical reason, he in the end, one might well conclude, merely dogmatically asserts that it is foundational. However, whatever one thinks of his response to the question "Why be moral?", it expresses a position about the status of morality within practical reason and does not by itself commit him to a position about the semantics, metaphysics, or epistemology of moral judgement. That is, accepting the position that the reasons to do one's moral duty are not derived from any non-moral reasons, but are moral through and through, does not commit one to any thesis about what moral claims mean, what moral predicates such as 'duty' express, or whether and how we come to know moral truths.

Prichard allows that one may come legitimately to doubt whether an action one thought was wrong really is wrong but insists that this is not the same thing as granting that an action is wrong and then wondering whether one has reason to do it (Prichard, 2002*a*: 18–20). In describing how one resolves such doubt he does apparently commit himself epistemologically. He claims that moral truths "can only be apprehended directly by an act of moral thinking" (Prichard, 2002*a*: 19). "We do not", he claims, "come to appreciate an obligation by an *argument*, i.e. by a process of non-moral thinking" (Prichard, 2002*a*: 13). However, to the degree that these genuinely are epistemological commitments, they are detachable from the claim that moral reasons are foundational. A reductive realist, for example, might think that moral reasons are foundational within practical reason, but deny that the epistemology involved is at bottom any different than that of the natural sciences.

Later, Korsgaard says that Prichard's way of asking the normative question, "Is this action really obligatory?" can be understood either as

[9] See in addition Prichard (2002*b*: 27–9).

asking whether a moral predicate has been correctly applied or as asking the question she is interested in, which is how any obligation can be normative. She claims that this ambiguity led Prichard to mistakenly believe that by showing that the requirement to perform an action can be derived from the principles of the correct moral theory, and thus that the moral predicate 'duty' correctly applies to the action, one has answered whatever request was posed by that form of words (1996: 39). Whether or not Korsgaard is right that Prichard was misled by a failure to distinguish these questions, we claim that her own normative question is itself susceptible to two different interpretations. "How can any moral obligation be normative?" can either express a request to place moral duty within practical reason, a "justification" of morality, or a request for a metaethics of moral judgement—an explanation of what it is to judge that *X* has a moral duty. By failing to distinguish these questions, Korsgaard gives the mistaken impression that, by showing that "substantive realism" is inadequate as an answer to the former question, she has shown that it is inadequate as an account of anything that might reasonably be requested by asking for "the source of normativity".[10]

Korsgaard's discussion of Moore's famous open-question argument in *Sources* is also infected by her failure to distinguish questions within practical reason from metaethical questions. She claims that the open-question argument derives its power from the pressure of "the normative question": "That is, when the concept of good is applied to a natural object, such as pleasure, we can still always ask whether we should really choose or pursue it" (1996: 43). But, she continues, this should not lead us to conclude, as Moore did, that normative concepts do not have criteria of application. Korsgaard seems to think that Moore, like Prichard, failed to distinguish the question whether a normative concept has been correctly applied from "the normative question", and thus that Moore mistakenly thought that because no naturalistic answer can be given to "the normative question", there can be no naturalistic criteria given to guide the application of a normative concept. But, of course, Moore himself claimed that there were synthetic necessary truths connecting normative and natural properties (e.g. pleasure is good)—that is, he would have accepted a naturalistic account of the normative-making properties. There thus is a sense in which he would have accepted that naturalistic criteria can be given

[10] Furthermore, we will argue below that this confusion leads Korsgaard to think that she is giving a full account of the source of normativity, when in fact she is best interpreted as arguing that a certain set of Kantian claims are the most fundamental normative claims of practical reason, not as giving a metaethical account that tells us what those claims mean, what metaphysical commitments we incur by making them, or how we can come to know them.

for the application of normative concepts, although he would have denied that such criteria constitute analytic definitions of normative concepts or that they allow us to reduce normative properties to natural properties.[11] Moore's non-reductive normative realism, although committing him to the claim that the property of good is not identical to any natural property, did not prevent him from accepting a naturalistic account of good-making properties. We will now argue that Korsgaard's attempt in *Sources* to contrast her own view with non-natural realism is spoiled by her failure to notice that non-natural normative realism is compatible with a naturalistic account of normative-making properties.

Korsgaard claims that the obligations that an agent has spring from what that agent's practical identity forbids, where a practical identity is a description of the agent under which he values himself and sees his life as worth living. Thus, for example, if you value yourself as a psychiatrist, you have an obligation not to violate your patient's confidence, since violating a patient's confidence is incompatible with the job description of a psychiatrist (1996: 101). She also claims that the value of an agent's practical identities depends on the value that he places on his own need for practical identities—his humanity (1996: 121). Furthermore, she argues that rational action is impossible unless agents value their humanity, and that therefore human beings are valuable (1996: 124). We do not want to question the truth of any of these claims, although there is much to contest here; rather, we question whether these claims amount to a metaethical position.[12] The problem is that before we can evaluate the metaethical status of such an account, we need to know what it is to *value* oneself under a description. Is this a belief that something, for example, psychiatry, is valuable? If so, then whether such an account is compatible with non-reductive realism all depends upon whether the belief that something is valuable is a belief in a non-natural property.[13]

[11] In fact, elsewhere Korsgaard herself seems to realize this: describing Moore's position, she writes "Of course it might be *true* that the good is pleasure, or the desirable, or what someone wills" (2003*b*: 103). But then it is not true, contrary to what Korsgaard claims in *Sources*, that Moore thinks that there are no naturalistic criteria for the application of normative concepts.

[12] It might be thought that it is the entire transcendental-style argument for the value of humanity, not the premises taken in isolation, which constitutes Korsgaard's alternative metaethical position. We hope that our discussion of Korsgaard's similar transcendental-style argument for the principle of instrumental reason will make it clear why this is not so. But for specific discussion of the metaethical status of Korsgaard's argument for the value of humanity in *Sources*, see Hussain and Shah (2005*a*).

[13] We are not claiming that valuing something is a belief that something is valuable. After all, there are interesting proposals that valuing something is a matter of having a certain hierarchy of pro-attitudes towards that thing. See e.g. Bratman (2000). We

Unfortunately, Korsgaard's commentary on her account does not help us to understand what metaethical position her account is supposed to yield:

> In one sense, the account of obligation that I have given in these lectures is naturalistic. It grounds normativity in certain natural—that is, psychological and biological—facts. . . . My account does not depend on the existence of supernatural beings or non-natural facts, and it is consistent with although not part of the Scientific World View. In that sense, it is a form of naturalism. (1996: 160)

The second sentence is vitiated by an ambiguity in the term 'grounds'. If 'grounds' just means "depends upon", then the sentence does not imply the absence of non-natural facts. The fact that something is valuable might depend upon natural facts—for example, that it is pleasurable, or that it is the object of an autonomous choice (if this is a natural fact)—but as long as this dependence is not the strong relation of identity, it is left open whether the fact that *x* is valuable is a non-natural fact about *x*. If Korsgaard were instead using 'grounds' in a non-standard way to mean "identical to", then her account would be a form of naturalistic realism. However, this would conflict with her explicit rejection of the kind of naturalism that "identifies normative truth with factual truth" (1996: 161).[14]

Moore himself would also have thought that once one has determined that, for example, pleasure is good—that is, that the property of pleasure is good-making, not that the concept of pleasure and the concept of good pick out the same property—the question whether one has a reason to pursue pleasure has been answered. He thus would not have understood the open-question argument as targeting the question whether we have a reason

explore non-cognitive interpretations of Korsgaard's account of practical identity in Hussain and Shah (2005*b*). Korsgaard does label her account 'constructivism' to contrast it with non-reductive realism. However, as we shall argue later, it is far from clear whether Korsgaard's characterization of constructivism amounts to a metaethical position. Our point for now is that Korsgaard's account of valuing does not by itself yield a metaethical alternative to realism.

[14] Later in this beguiling passage Korsgaard writes: "From outside that (first-personal perspective) standpoint, we can recognize the fact of value, but we cannot recognize value itself " (161). If value cannot be discerned from the empirical third-person perspective, then naturalistic realism is ruled out. A natural way of interpreting the thought in this quotation is that, from the empirical perspective, we can discern the good-making facts, but it is only from the non-empirical, normative point of view of practical reason that we can see that these facts are good-making, because it is only from such a perspective that we can come to know the normative principles that tell us which natural facts are good-making. While this position implies the denial of naturalistic realism, it is perfectly consistent with non-natural normative realism, which says that normative facts are irreducible to natural facts, but which allows that normative facts are dependent on natural facts; that is, it accepts that the only things that have the property of goodness are natural objects/properties.

to pursue our duty or do what is good. Moore aimed the open-question argument at the *semantic* question whether normative concepts can be defined in terms of natural concepts, and concluded from this argument that they cannot. Furthermore, from this semantic result he inferred that normative concepts refer to irreducible normative properties. But none of these conclusions are answers to the question of the place of moral considerations within practical reason, and thus none answer the normative questions, "Why should I do my duty?" or "Why should I pursue what is good?"

Because Korsgaard fails to distinguish this normative question from the metaethical question that Moore was asking, and misinterprets Moore's semantic conclusion that good is indefinable (at least in purely naturalistic terms) as attempting to answer the question of whether (naturalistic) criteria can be given to guide the application of normative terms, she fails to come to grips with, much less argue against, Moore's metaethical non-reductive realism. This is not to say that there is not a legitimate metaethical worry lurking behind Korsgaard's ill-formed objections to non-reductive realism, which is that Moore's account does not provide any illumination: we have no account of what it is for a property to be a normative property, and we have no substantive epistemology that explains how we come to know normative facts. Perhaps Moore is right that no such illumination is possible, but one can sympathize with Korsgaard's inchoate desire for illumination nonetheless.[15] If this is Korsgaard's dissatisfaction with non-natural, normative realism, she fails to express it correctly, because she fails to disentangle the worry that non-reductive realism fails to illuminate and give a substantive epistemology of the normative properties that it claims are expressed by normative predicates from the worry that non-reductive realism fails to illuminate the place of morality within practical reason.

3.2. "The normativity of instrumental reason"

Just as asking for the source of the normativity of duty can be interpreted either as a request to place duty within a conception of practical reason or as a request for a metaethical interpretation of judgements such as "One has the duty to provide for one's children", so too asking for the normative foundation of the principle of instrumental reason can either be interpreted as a request to place the principle of instrumental reason within a conception of practical reason or as a request for a

[15] Another worry that Korsgaard might be trying to express is that the non-natural normative realist cannot explain how moral facts are able to motivate us. Below, we discuss this interpretation of Korsgaard's worry in connection with her discussion of the realist position about the principle of instrumental reason.

metaethical interpretation of the principle (or of particular means–ends normative judgements). In her article, "The Normativity of Instrumental Reason", Korsgaard certainly, once again, can come across as taking a metaethical position. She introduces "the Kantian conception of practical rationality" as a "third and distinct alternative" to be distinguished from, and preferred to, "empiricist" accounts, on the one hand, and "realist" or "dogmatic rationalist" positions on the other (Korsgaard, 1997: 219). Such metaphysical and epistemic sounding labels are once again hard not to read as labels for positions identified by metaethical commitments. And again her setting of her position in contrast to positions with such labels suggests that she takes herself as defending a distinctive metaethical position.

We will approach the question of whether she is indeed expressing a metaethical position in this article in two stages. First, we will consider her arguments against what she calls the realist position. As in the moral case, her failure to distinguish between different questions about normativity confuses the issue here. On the normative reading of her objection to the realist, the realist can deploy a Prichard-style response. Such a style of response shows that the realist position could be coherent. More importantly, as in the moral case, this response has nothing in particular to do with any metaethical position. If we try to read her worry about the realist in metaethical terms, then it is much harder to see what her objection is supposed to be—though we will consider a couple of possibilities. In any case we assume that whatever metaethical objections can be read out of her discussion are supposed to be objections to which her own view of instrumental reason is immune. The discussion in this section will thus set the stage for a consideration in the next section of the apparent positive position expressed in the article—the Kantian conception of practical reason according to which the principle of instrumental reason is constitutive of the will.

3.2.1. The critique of non-reductive realism

She identifies the realist position initially as the view that moral requirements must have "some sort of ontological foundation, by positing the existence of certain normative facts or entities to which moral requirements somehow refer" (Korsgaard, 1997: 218). This is the view adopted by the "dogmatic rationalists", a term that she then uses interchangeably with "realist" for the rest of the article.[16] Given that the epistemic-sounding "dogmatic

[16] We note immediately one potential source of confusion here that arises from mixing ontological and epistemic labels. An empiricist can normally, though not apparently in Korsgaard's idiosyncratic terminology, be a realist. Of course the empiricist conception of the ontology will be such as to fit his epistemology—normative facts must be knowable by empirical means.

rationalism" is the alternative label for what she calls realism, we take it that Korsgaard intends this label to express the position that the relevant normative facts are not empirically accessible and so presumably are also not naturalistic or material facts. The epistemology of this position is to be some kind of intuitionism.[17]

She claims that the difficulty for non-reductive realism "exists right on its surface, for the account invites the question why it is necessary to act in accordance with those reasons, and so seems to leave us in need of a reason to be rational" (1997: 240). As we think is clear from the context, Korsgaard must mean "rationally necessary" as opposed to, say, metaphysically or causally necessary. We also assume that the point here is *not* that, though there can be reasons to ϕ, I perhaps ought not to ϕ because the balance of reasons favours not ϕ-ing. Putting the point more clearly in terms of an ought, then, the question supposedly invited is "Why ought I to do what I ought to do?" The Prichard point is that such a question does not make sense. If I have accepted that I ought to ϕ, then how can it still make sense for me to ask why it is rationally necessary to ϕ?[18] To think that I ought to ϕ just is to think that it is rationally necessary to ϕ. The point has nothing to do with the metaethical issues of what kind of mental state I am in when I think that I ought to ϕ or whether there is a mind-independent fact accessible only by rational intuition that I ought to ϕ.

Thus understood as one kind of normative question it is hard to see what the objection is. Of course, there are questions we can sensibly ask our Prichardian about instrumental reason, including further normative questions. We can ask what makes particular considerations reasons—the reason-making features. The Prichardian might respond in predictable ways.[19] What makes the fact that ψ-ing is a means to ϕ a reason to ψ is that you, say, desire to ϕ.[20] We are not sure about this, but Korsgaard seems to suggest that this would be to derive an "ought" from an "is" (1997: 245).

[17] When she turns to specifying what such a realism about instrumental rationality would look like, we get an additional ontological claim: "truths about reasons . . . exist independently of the will" (219). Finally, when she turns in earnest to the discussion of realism she says that according to realism "there are facts, which exist independently of the person's mind, about what there is reason to do" (240). Note that independence from the will is not identical to independence from the mind. We mention all of this to emphasize how, in one way, the so-called realist target is quite limited—it is a subset of realist positions out there—and how it is very clearly specified in terms of epistemological and metaphysical features.

[18] Cf. Dreier (2001).

[19] Not Prichard himself since he was quite suspicious of the idea that there might be, as he would put it, such "*general* " answers available (Prichard, 2002*c*: 62; 2002*a*).

[20] We do not endorse this version of the principle of instrumental reason. There are several issues here. (i) How should the antecedent be specified? Should it be specified in

But one is no more deriving an "ought" from an "is" in this case than when one says that you should mow the lawn because the grass is tall or that what happened to her is bad because she is in pain. Non-normative facts will *make* certain normative claims true in any non-error-theoretic account. This is just a result of the fact that normative properties rarely, if ever, apply barely.[21]

Our Prichardian might well grant that there is a general normative truth in the background expressible by some version of the following:

(1) For all S, ϕ and ψ, if S (believes that S) desires/intends that S ϕ and (S believes that) S's ψ-ing is a means to ϕ, then S (believes that S) ought to/has reason to ψ.

For all we have said the Prichardian can think of this as another premise that the agent *believes* and then combines in his reasoning with the following beliefs:

(2) My *a*-ing is a means to *b*-ing
(3) I desire to *b*

to reach the conclusion

(4) I ought to *a*

The inference principles relied on are just the ones of theoretical reason.

The alternative, which is also open to the Prichardian, would be to introduce the instrumental principle as a practical inference principle in its own right. He would also add that what makes following that principle correct is precisely that the associated normative claim is true.

In "Realism and Constructivism" Korsgaard argues that the "realist account of the normativity of the instrumental principle is incoherent"

terms of desiring, intending, or willing? Should the antecedent be normative? (ii) Should the principle allow for detachment? (iii) Should it should be a "strict" or a "slack" demand? See Broome (2000) for a discussion of (ii) and (iii). For an extended discussion of the principle of instrumental reason, see Stalzer (2004).

[21] She makes a similar mistake in her discussion of Derek Parfit. She seems to think that Parfit—whom we can treat as a non-reductive realist—has to choose between two views: first, that the complex considerations we use in determining whether an action is right will not explain why the action is right—"It is right because it has the property of rightness"; second, that these considerations "constitute its rightness" (Korsgaard, 2003*a*: 3). Now Parfit should answer that neither is the case. The considerations *make* an action right but do not *constitute* rightness. When the considerations are complicated, then it might well be hard work to come to know that the action in question is right. Indeed an action is right because it has the property of rightness, but we can still ask why it has the property of rightness and this leads us to the considerations that make it right. Talk of constitution though threatens, unless the possibility is explicitly ruled out, to be a reductive view and Parfit wants to avoid that.

(Korsgaard, 2003*b*: 110). The instrumental principle would have to be "some sort of eternal normative verity". "How", she asks, "is this verity supposed to motivate him?" (110). The picture, she claims, is incoherent: "The point is that the instrumental principle cannot be a normative truth that we *apply* in practice, because it ... is essentially the principle of application itself, that is, it is the principle in accordance with which we are operating when we apply truths in practice" (110). Applying the normative truth, the principle of instrumental reason, presupposes that we are already applying the principle.[22]

The point, though, does not have to be put in terms of the principle of instrumental reason. Consider the following more general reconstruction of Korsgaard's point here. The heart of the internalist thought is that normative beliefs are practical. That is, in order to have any normative belief I must be able to act on it. Put in the most general terms, the agent has to be applying or following something like the following normative requirement in order to have any normative beliefs:

(5) $O(BO\phi \rightarrow \phi)$

The symbolism here is basically Broome's: believing that you ought to ϕ requires ϕ-ing. We are proposing to read ϕ generously to allow, for example, $O\phi$ to be a statement of the general instrumental principle like (1): roughly, believing that you ought to be instrumentally rational requires being instrumentally rational.

Now Korsgaard's point is that an agent cannot be motivated by a belief with the content (5) unless he is already applying (5). Imagine giving the agent the following belief

(6) $BO(BO\phi \rightarrow \phi)$

This is just another belief of the form $BO\phi$. No consequences for motivation follow unless the agent is already applying (5). And so perhaps it is not even possible for the agent to have a belief with normative content without already applying (5).

We agree that something like this seems right and similar points will hold for some theoretical norms. It may be true that one does not count as having beliefs unless one is thinking correctly to some extent. If so, then one cannot come to apply a fundamental principle of thinking on the basis of believing

[22] Compare her comment about "goodness in action" earlier on p. 110: "To put the same point another way, goodness in action cannot just be a matter of applying our knowledge of the good—not even a matter of applying our knowledge of what makes action itself good. This is because the ability to apply knowledge *presupposes* the ability to act."

a normative claim, since one would already need to apply the principle in order to believe the normative claim. If this is Korsgaard's point, the realist should happily grant it, as it is perfectly compatible with his central claim that the truth of the normative claim makes it correct to follow the principle. A normative truth can make a certain way of thinking correct even if it cannot be an agent's grounds for coming to think that way. This is just the point that what makes something correct to do and what one's grounds are for doing it need not coincide.[23] In fact, the realist can go further and claim that an agent can come to believe that the normative fact that he ought to be instrumentally rational makes it correct to follow the instrumental principle—which is the principle that he follows in arriving at this very belief—even if the fact that he ought to be instrumentally rational cannot be his initial ground for following the instrumental principle. Thus while Korsgaard may have shown us that normative facts cannot play a certain epistemological role, she has not shown us anything that the non-reductive realist cannot take in his stride.

3.2.2. Interpreting the positive account

To bring out further how Korsgaard's criticism of non-reductive realism misses the mark, and more importantly, to show that Korsgaard does not in fact commit herself to a metaethical position, we will now argue that, for all Korsgaard says in "The Normativity of Instrumental Reason", non-reductive practical realism is compatible with her own "explanation" of the normativity of instrumental reason. This is because what she says only commits her to an explanation of the place of the principle of instrumental reason within practical reason, not to a metaethical interpretation of what it is to think normative thoughts such as that one ought to take the necessary means to one's end.

In trying to reconstruct a positive metaethical position from "The Normativity of Instrumental Reason", we will end up considering several possibilities for such a reconstruction; however, we begin with what naturally comes across as a family of potentially distinctive metaethical claims, namely, the claims that the will or action are supposedly constituted by certain principles or norms.

Now once again her introduction of her position seems to be driven by the normative question about the place of the instrumental principle within practical reason: "The [realist] model, as I said earlier, seems to invite the question: but suppose I don't care about being rational? What then? And in

[23] Thanks to David Velleman for drawing our attention to the relevance of this distinction.

Kant's philosophy this question should be impossible to ask" (1997: 244). It is not clear, as we have emphasized, what is meant by "care". If what Korsgaard is trying to get at is a possibility where someone accepts that ϕ-ing is rational, but then proceeds to ask whether it is rational to ϕ, then the realist can insist that the question does not make sense though, as we have emphasized, the realist's insistence is independent of his metaethical position—his realism.

But what is Korsgaard's central positive claim? "To will an end just is to will to cause or realize the end, hence to will to take the means to the end. This is the sense in which the principle is analytic. The instrumental principle is *constitutive* of an act of the will. If you do not follow it, you are not willing the end at all" (244). The problem, as Korsgaard realizes, is that this does not seem to allow for the possibility of instrumental irrationality. If it is logically impossible to will an end without taking the means to the end, then it is impossible to be instrumentally irrational—to will an end and fail to take what one recognizes to be the means to that end.

To prevent this she makes one negative claim about willing: "So willing the end is neither *the same as* being actually disposed to take the means nor as being a particular mental state or performing a mental act which is *distinct from* willing the means" (1997: 245). And a positive claim about willing:

> [W]illing an end just is *committing* yourself to realizing the end. Willing an end, in other words, is an essentially first-personal and normative act. To will an end is to give oneself a law, hence, to govern oneself. That law is not the instrumental principle; it is some law of the form: Realize this end. That of course is equivalent to 'Take the means to this end'. So willing an end is equivalent to committing yourself, first-personally, to taking the means to that end. (1997: 245)

There is a lot packed in here that one wishes had been laid out a bit more slowly. Our hope is to develop the different possible charitable interpretations of Korsgaard's position and see how far they go.

Put aside for a moment that talk of "laws". Now, much of this actually sounds like many a Prichardian of our acquaintance—though we realize perhaps not yours. To allow for differences to emerge, let us use a term other than "willing" for the kind of attitude or act that the Prichardian wants to claim is directed at the end in the cases in which the principle of instrumental reason applies: we will use the term "intending". So our Prichardian states a string of conceptual truths:

> Intending an end just is *committing* yourself to realizing the end. Realizing the end requires taking the means to the end. So committing yourself to realizing the end is equivalent to committing yourself to realizing the means to your end. *Following* the instrumental principle is committing yourself to realizing the means

> when you commit yourself to realizing the end. So you don't count as committing yourself to realizing the end unless you are following the instrumental principle. The instrumental principle is *constitutive* of intending. If you do not *follow* it, you are not intending the end at all. And, of course, if you're not following it, you're also not being rational.

The plausibility of these conceptual truths turns no doubt on taking "committing", "following", and "intending" as normative concepts. Is this a problem for the Prichardian? It is not at all clear why it would be. The Prichardian, whatever his metaethics, is no error theorist. He is happy to make claims involving unanalyzed normative concepts and he is convinced, let us imagine, by Korsgaard that these sound like good ones to make. But if he can make them too, then these claims seem, unsurprisingly to us, to be merely normative claims and not part of any distinctive kind of metaethical position.[24]

But what about the claim that "willing the end is neither *the same as* being actually disposed to take the means nor as being a particular mental state or performing a mental act which is *distinct from* willing the means" (Korsgaard, 1997: 245). Well, our Prichardian could happily go along with Korsgaard here, but could also deny the claim. He *could* insist that the mental act of willing the end is distinct from the mental act of willing the means. Korsgaard has the odd view that the "dogmatic rationalist conceives willing an end as being in a peculiar mental state or performing a mental act which somehow logically necessitates you to be in another mental state or perform another mental act, namely, willing the means" (1997: 244). But it is hard to imagine our Prichardian saying anything like that for precisely the reason Korsgaard goes on to give: "for no mental state can logically necessitate you to be in *another* mental state or perform another mental act" (1997: 245). The Prichardian would claim that being in one state *rationally, not logically*, necessitates being in the other state or performing the other mental act. So there can be two distinct states or mental acts. It is just that being in one involves a commitment to being in the other.

But what about the Prichardian who is committed to a realist metaethics and a non-reductive one at that? Didn't we grant that Korsgaard might well have legitimate worries about this view? And surely there must be something in Korsgaard's view that is meant to be able to respond, or avoid, precisely those worries that then will also allow us to distinguish her position from the realists. Or at least her claim that the will is constituted by normative principles adds something metaethically distinctive, whatever metaethical label we want to apply to the resulting position. As it turns

[24] There will be particular metaethical positions that might rule them out.

out, nothing Korsgaard says implies anything of the sort. Consider first the epistemological worry that the realist has no account of how we come to know normative facts. As we have just noted though, the normative claims the Prichardian has just endorsed are all apparently conceptual claims. Issues remain. First, arguably the Prichardian will have to step out of the circle of conceptual claims at some point. In the case of instrumental reason, surely at some point the agent will have to will an end. Can the thought that I am willing an end be a conceptual claim? Implausible. So appealing to some supposedly unproblematic epistemology for conceptual claims will not resolve the epistemological issue. The realist thus seems to be left without an illuminating account of how I can come to know that I am willing an end, at least if he takes this to be a normative claim. If I could know that I am willing, then the conceptual claims listed above, plus a means–end belief, should take me the rest of the way—not through action, but to a commitment to taking the means, to willing the means.

But has Korsgaard made progress on the epistemic front? Maybe she thinks that because this kind of normative fact is not mind- or will-independent, there is no need for a substantive tracking epistemology. But we need to be very careful. Again the distinction between normative-making properties and the normative property itself is crucial. To say that I am a bad person because I want to sleep with my neighbour's wife is not immediately to claim that this normative fact is mind-dependent just because the desire is a mental state of mine. Or if it is sufficient to make the normative fact mind-dependent, then there is no "realist" opponent. Wrong-making properties can be mind- and will-dependent and many, if not most of them, are. What is important is whether the normative property is mind-dependent. Merely saying that willing is normative does not make the normative property mind-dependent or will-dependent in any interesting sense. Wrongness is not desire-dependent because my having a certain desire can be wrong. So even if willing is normative, the normative property could be mind-independent. And if it is mind-independent, then it is not obvious how I come to know that it is instantiated.

So perhaps we should take Korsgaard as suggesting that the normative property is not mind-independent. Now, if all Korsgaard adds to this is the brute claim that the normative property *is* mind-dependent then we do not really have a distinctive metaethics, nor one that really provides any help on the epistemic front. As far as the brute claim of mind-dependence goes, our Prichardian can happily keep step—normative properties are mind-dependent but no reason has been given yet to think of them as any less real. No reason not to either of course. The problem is that a brute mind-dependence claim does not get us very far. For all we have said, the view could be the magical one: when I am in a certain mental state the normative property itself somehow comes into existence and not just its particular instantiation. Such metaphysical dependency does not yet imply any particular story of epistemic access. There are complicated issues here, but the main point is that simply declaring mind- or will-dependence does not really get us anywhere.

The problem we may seem to have been circling around though is the question of what willing itself is. Is it something mental? Surely it must be. For Korsgaard willing is "essentially" normative (1997: 245). Perhaps here lies the distinctive sense in which the normative is will- and mind-dependent and perhaps here lies the solution to our epistemological worries.

What does it mean to say that something is essentially normative? Can the "realist" endorse such claims? Well, the realist does make claims such as that murder is essentially wrong. But perhaps this is confused because the realist has to identify something that has the normative property. It is the particular act of killing that has the normative property of being murder. This way of putting it suggests that one can distinguish between the thing of which the normative property is being predicated and the normative property itself. And this might suggest either that we can identify the entity (here an act) without using normative language or, in fact, that it is not essential to the act that it have this normative property—the normative property is not an essential property of the act in question. Perhaps then Korsgaard's distinctive suggestion would be that the act, for her the act of willing, is essentially normative in the sense that it cannot be identified without using the normative language and so the normative property is essential to it in the way that it is not to the act of Jill killing John.

However, there is no reason for a non-reductive realist to grant all this. There may well be acts of killing that are not murder, but the realist can claim that *this* act of Jill killing John *is* murder and essentially so. And he may well deny that there is any way to identify it in terms other than the normative. Similarly, without the normative concepts of belief, desire, intention, we cannot identify an action. A sequence of events standing in causal relations identified without these terms may not, such a view might insist, line up with a description in terms of beliefs, desires, intentions, and actions.[25] This is the sense in which our non-reductive realist's picture could involve a commitment that might capture what Korsgaard is getting at in her talk of the "first-personal". There is no identifying actions in nomological vocabulary. Of course this does not mean that I can only make judgements about myself. Most of our moral practices rely on our ability to make judgements about when others have acted and what they intended.

[25] Cf. Hornsby (1997: 295–6).

We would basically have to give up our existing moral practices if we could not make such judgements—if it literally only made sense for me to judge about willing in the first person.

So there does not look like anything in the talk of willing being essentially normative that the realist has to deny. Now there are puzzles here for the realist, but these, as we shall try to show, are just as much puzzles for Korsgaard.

There is the question of how the level of normative vocabulary "fits" with the level of physical or nomological vocabulary. This is a classic puzzle of course central both to Kant, those inspired by Kant, and to contemporary discussions in the philosophy of action and mind. We do not intend to defend the claim that there is really a difficult problem here or that non-reductive realist approaches will not work. Our point is just that, as far as we can see, Korsgaard says nothing here that contributes to the debate. What she does say could, as we have suggested, be interpreted in non-reductive realist terms. Perhaps this interpretation would lead to a solution—consider the writings of McDowell, Hornsby, or Dancy, or for that matter Davidson's own view. The point though is that Korsgaard has not provided the Kantian with any distinctive way of solving this problem.

Such a non-reductive realist also may have a puzzle when it comes to the question of concept acquisition. How is it that I acquire these normative concepts in order even to have these thoughts? This question is closely related to the worry that the non-reductive realist has no non-trivial account of the content of the normative thoughts. Now, whether concept acquisition is a problem will depend on what answer we give to this question and various background considerations, but again there is nothing Korsgaard says that will help with any of this.[26]

Korsgaard gives no non-trivial account of the content of the relevant normative claims and thoughts beyond the conceptual claims mentioned already. Defending this claim does allow us to fill in a lacuna in our discussion of Korsgaard. We had put aside the talk of laws and legislation right at the beginning of our attempt to give an interpretation of Korsgaard's positive view. It is therefore true that she does say more about willing than we have considered above.

[26] As long as one is not concerned about an error theory or some other kind of confusion lurking in the conceptual scheme, then giving no account of the content is probably fine for practical purposes. We can happily go on using our concepts without any such metaethical account. But this is again a point that does not differentiate between Korsgaard and the non-reductive realist.

> To will an end is to give oneself a law, hence, to govern oneself. That law is not the instrumental principle; it is some law of the form: Realize this end. That of course is equivalent to 'Take the means to this end'. So willing an end is equivalent to committing yourself, first-personally, to taking the means to that end.

We are not quite sure how equivalence relations work between laws, but we are happy to grant that perhaps it is a normative truth—conceptual or not—that the law of the form "Realize this end" is equivalent, in some sense, to a law of the form "Take the means to this end". We are even happy to grant that the command "Realize this end!" is the same as the command "Take the means to this end"—not that we have any special account of the content of commands to provide here. The problem though is that the language of giving oneself a law and governing oneself is surely normative language and so by itself does not help with the metaethics or with providing us with a non-trivial account of the content.

Here is one way to see the point. What is the difference between intending an end and giving oneself a law of the form "Realize this end!"? Korsgaard will want to insist that there is no difference but, as always with such conceptual claims, the account of intending in terms of self-legislation will only be illuminating if we have some understanding of what it is to self-legislate *other* than just intending an end. The problem is that when we shift to some attempt to elucidate what it is to give oneself a law we either have no further elucidation or we end up relying on non-normative reductions that seem quite implausible. So consider how Korsgaard elaborates on the talk of self-legislation: "Then what does it mean to say I take the act of my own will to be normative? Who makes a law for whom? The answer in the case of the instrumental principle is that I make a law *for me*. And this is a law which I am capable of obeying or disobeying" (246). What is required is "that there be two parts of me, one that is my governing self, my will, and one that must be governed, and is capable of resisting my will" (247–8). Does any of this help to elucidate the content of the normative claims? Not really. The problem is that the language being used is both metaphorical and still normative. To the degree that we have a grasp of what it is to make laws, what it is to "give" someone a law and what it is to govern someone, it is because we are competent users of these notions when we talk about kings or legislatures—political sovereigns—issuing laws that their subjects are to follow. In their normal context of usage the concepts here are themselves normative and in philosophical discussions of them their metaethics are contested.[27] The problem is only compounded by the fact that the natural way to read their deployment in Korsgaard's work is surely as metaphor.

[27] As we well know from discussions in political philosophy, it is not at all easy to say what political authority or sovereignty comes to or what the nature of political obligation

There is no political state, no court system, no police *literally* in the self, though there is a long tradition of talking metaphorically as though there were rulers and ruled and so on. It is hard to see how the metaphor says anything more than simply that I can will an end, but fail to will the means even though I should.

Korsgaard's "explanation" of normativity in the end, then, merely uses all of the normative notions that we were hoping for an elucidation of and thus fails to constitute an account of the content of normative claims. This is fine, as we keep emphasizing, if she is doing normative ethics, but of course it prevents her from presenting any metaethical position. Thus whatever worries Korsgaard might have about the non-reductive realist's inability to say more about "*what it is* that [an agent] recognizes" when he recognizes "certain considerations *as* reasons"—to say, in other words, more about normative content—the same worries apply to her view. Indeed the unwillingness to step outside the circle of normative concepts is strikingly similar, however legitimate it might well be. The non-reductive realist at least adds some claims about the nature of normative facts, the semantics of normative language, and the nature of the attitudes that we have towards normative content—he at least says things that look as though they will commit him metaethically. Korsgaard seems to reject these additions of the non-reductive realist. Her remaining claims though are completely compatible with what a non-reductive realist would say. She does not replace the realist's positive metaethical claims with any of her own and so there appears to be no new metaethical position expressed.

Of course it would be a mistake simply to conclude from all of this that Korsgaard and the non-reductive realist have the same metaethical position. The fundamental problem is that her claims are compatible with different metaethical positions or interpretations.

4. Constructivism

Korsgaard claims the banner of constructivism for her view, and she clearly sees constructivism as an alternative to non-reductive realism. We will finish by considering whether her thoughts on constructivism amount to a distinctive metaethical position.

Korsgaard suggests in "Realism and Constructivism" that the difference between the cognitivist position she labels "substantive realism" and her own

is. And as we well know from the discussions in the philosophy of law between legal positivists, natural law theorists, and legal realists, it is not at all clear what indeed calling something a law comes to.

"constructivist" alternative lies in their views of the *function* of normative concepts: The substantive realist thinks that the function of normative concepts is to describe normative reality, whereas the constructivist thinks that the function of normative concepts is to label the solutions to practical problems of what to do (Korsgaard, 2003*b*: 116). But this description of the function of concepts does not yet reveal whether Korsgaard has an alternative metaethics, nor does it set up a contrast with the substantive realist's traditional conception of normative concepts. First of all, if the problems themselves are couched in normative terms, Korsgaard's description of the function of normative concepts as labelling solutions to these problems will not help to establish a metaethical position. The problem is that, in order to understand this function, we would first need to understand the normative concepts that express it. So we would already need an account of normativity before we could use Korsgaard's account to grasp the functional distinction she wants to draw. Second, why should we think that the function of describing normative reality and the function of solving practical problems of what to do conflict? Surely the substantive realist will agree that ethics is about finding the solution to the problem of what to do, adding that normative facts (e.g. action *A* has the property of to-be-doneness) provide the answers to these questions. Correctly describing normative reality, discovering which actions have the property of to-be-doneness, answers the question of what one should do.

Korsgaard does say more about constructivism, or as she also calls her position, procedural realism, in *Sources*.[28] So we will examine her discussion there to see if anything she says establishes a genuine alternative to non-reductive realism. Korsgaard argues that her own view is a form of procedural realism, as opposed to substantive realism, which she is arguing against. Here is Korsgaard's initial characterization of the difference:

> Procedural moral realism is the view that there are answers to moral questions; that is, that there are right and wrong ways to answer them. Substantive moral realism is the view that there are answers to moral questions *because* there are moral facts or truths, which those questions ask *about*. (1996: 35)

But this way of putting things makes substantive realism out to be a species of procedural realism, as the substantive realist agrees with the procedural realist that there are answers to moral questions, and that there are right and wrong ways to answer them. The substantive realist also gives a particular

[28] Although we will argue below that, given the way that Korsgaard uses the terms, procedural realism and constructivism are not equivalent positions. Procedural realism is a broader category of which substantive realism and constructivism are meant to be distinct species.

explanation of what answers moral questions, namely moral facts, but this is not excluded by the description of procedural realism given above. That is, nothing in the above specification of procedural realism excludes the claim that there are procedure-, mind-, or will-independent moral facts. Therefore the specification fails to set up a contrast between procedural realism and substantive realism.

Korsgaard's point, we take it, is that procedural moral realism does not force the acceptance of substantive realism, and might be filled out in a way that makes no commitments to mind-independent intrinsically normative entities. Specifically, it allows that the answers to normative questions are "the results of some constructive procedure" (1996: 35). But of course even a substantive realist can allow that we need procedures for arriving at true moral beliefs, since these procedures are what allow us to track the moral facts. Thus, talk of moral answers being the result of a procedure does not by itself establish the needed contrast between substantive realism and the type of realism Korsgaard wants to advocate.

Does it help to be told that the procedure is "constructive"? Well, what does this mean? It sounds as if it means that the employment of certain procedures creates a normative entity. Thus rather than saying correct procedures track independently existing moral facts, the constructivist claims that the moral facts are created by the employment of these procedures. As Korsgaard puts it:

> The procedural moral realist thinks that there are answers to moral questions *because* there are correct procedures for arriving at them. But the substantive moral realist thinks that there are correct procedures for answering moral questions *because* there are moral truths or facts that exist independently of those procedures, and which those procedures track. (1996: 36–7)

Strictly speaking, what she says here does not entail that procedural moral realism is committed to the claim that moral facts are created by the employment of correct procedures, but it is difficult to understand what else could make sense of the non-tracking relation between the correct procedures and moral facts that she has in mind. In any case, it is fairly clear from Korsgaard's overall position that the relation between correct procedures and moral facts that she intends is one, in some sense, of "construction" or creation.

Do we now have the needed contrast between substantive and procedural realism, or better, the contrast between substantive realism and constructivism, since as we noted "procedural realism" seems to denote the category of which substantive realism and constructivism are species? We cannot answer this question until we have a fuller specification of constructivism. First of all, we need a characterization of the procedures that construct the moral facts. The key question here is whether the specification of these procedures employs normative concepts. For example, Rawls's specification of the procedures that are used to construct justice makes use of the concept of the reasonable. Again, there is nothing wrong with this so long as one is engaged in giving an account of justice within normative ethics, as Rawls is. But this will not do if one is attempting to give a metaethical account of the concept of justice, as it leaves us with the unexplained normative concept of the reasonable. And if we try to give a constructivist account of the concept of the reasonable, we face a regress so long as the procedures used to construct the normative facts expressed using this concept are themselves couched in normative terms.

What if the metaethical constructivist tries to avoid this problem by specifying the correct procedures in non-normative terms? Do we now have a full-blown metaethical position? The problem is that, even if the constructivist is able to specify the relevant procedures in non-normative terms, he is committed to the normative judgement that these are the *correct* procedures. After all, it is only if the procedures are the correct ones that one can use them to construct the relevant normative facts; presumably not any old procedure has the power to create normative facts. But then, what is the constructivist's metaethical account of what it is to judge that a procedure is correct? At this point, all the familiar realist metaethical options appear open. For example, it is open to the reductive realist to say that the relevant normative property (e.g. justice) is identical to the property of being the output of these procedures, specified in non-normative terms. Since a social institution's property of being just and its property of being the output of these procedures are identical, the question of the correctness of the procedure does not come up.[29] But it is also open to the non-reductive realist to say that the procedure has the separate normative property of correctness. Might the constructivist try to give a constructivist account of the judgement that a procedure is correct, going constructivist all the way down, as it were? It is hard to see how this would really provide much elucidation. If the constructivist says, for example, that a certain procedure, call it *x*, creates the facts about which procedure of justice is correct, he must claim that *x* is the correct procedure for constructing facts about correct procedures of justice. And then we are left with an unexplained normative concept, since we still need to be told what it is to judge that procedure *x* is correct.

Might the constructivist specify a procedure that creates its own correctness? It is hard to know what this would involve. If it just means that the

[29] Note that the reductive naturalist could propose this property identity either as an analytic truth or as part of an explanatory empirical theory of the relevant normative discourse.

procedure is self-validating, then it will not single out a unique procedure, since there are many procedures that are self-validating. And then the question arises as to which of these procedures is the correct one. Even if this problem could be avoided, we still would not have an account of what it is to make a correctness judgement, we would just have an account of which correctness judgements are correct. Put it this way: we ask what it is to make a normative judgement. We are told that normative judgements are about normative facts that are created by correct procedures. We then ask what it is to call a procedure correct, and we are told that correct procedures are those procedures that are constructed by a certain procedure, *x*, which is the correct procedure for creating correct procedures. We then ask what it is to judge that *x* is correct. And now what? We never seem to get outside of the normative circle. The objection here is not that the constructivist fails to give us a non-normative reduction of normativity, since we are allowing for positions such as non-reductive realism that commit themselves to sui generis normative properties. The problem is that non-reductive realism or some other metaethical position needs to be added to constructivism in order to turn it from an account of which normative judgements to make into an account of what it is to make a normative judgement. But if this is so, then constructivism is not really a metaethical position at all.

There is another question that needs to be answered before we can assess constructivism as a full-blown metaethical alternative, even if the constructivist can somehow avoid the problem we just laid out: what does it mean to say that the normative facts are *created* by correct procedures? It is important to recall here the point about right-making properties. Imagine that one of our students follows his usual complicated "procedure" for cheating on an exam. That this procedure was followed will make it the case that he has done something bad. In one sense then the normative fact was *created* by the procedure. Emphasizing "created" here though sounds rather dramatic and a non-reductive realist will insist that our student has merely ensured that the relevant normative property is instantiated. He did not create the property itself and indeed the "creation" of the normative fact was not something that, in one important sense, he could control: once he had followed his "procedure", the badness of his action was not up to him. Thus if normative facts were being created by procedures in only this sense, we would not have an alternative to substantive realism. However, when we try to imagine what it would be for normative facts to be created in some more substantive sense, then the view does start to sound quite magical. The mysteriousness of such acts of creation *ex nihilo* seems to be on the same order as the ontological and epistemic mysteries of the non-reductive realist's intrinsically normative entities and our supposed intuitive access to them. Finally, for all that has been said, once these normative facts are created they could be just like the substantive realist's normative facts. It is not obvious here that different ontogeny entails different ontology. Much more clarification would be needed before we could know how different such a view really would be from substantive realism and what kinds of theoretical costs such differences would incur.

5. Conclusion

How is it, then, that Korsgaard has put herself in the awkward dialectical position of framing the Kantian position in opposition to non-reductive realism and yet presenting us with a position whose content is compatible with it? As we have argued throughout the paper, the fundamental problem occurs at the very beginning. In framing her inquiries about the "source of normativity", Korsgaard fails to distinguish the metaethical question of what it is to make a normative judgement from normative questions about which normative judgements to make or even which normative judgements we cannot help but make. This leads her to mistakenly think, for example, that by making a strong case for a Kantian position that certain normative judgements are constitutive of agency, she has given an alternative to the non-reductive normative realist's position about the meaning, metaphysics, and epistemology of these normative claims. Her opposition to non-reductive realism similarly suffers from a misunderstanding of the metaethical tasks this theory seeks to accomplish. This causes her to take them to task for failing to give plausible answers to questions of normative ethics, when in fact, for all she says, her own answers to these questions are available to non-reductive realists. While her positive claims constitute a Kantian position on foundational questions in practical reason, they do not constitute a metaethical position—they do not constitute an alternative to non-reductive realism.

REFERENCES

Bratman, Michael, 'Valuing and the Will', *Philosophical Perspectives*, 14 (2000), 249–65.

Broome, John, 'Normative Requirements', in J. Dancy (ed.), *Normativity* (Oxford: Blackwell Publishers, 2000), 78–99.

Dreier, James, 'Humean Doubts about Categorical Imperatives', in E. Millgram (ed.), *Varieties of Practical Reasoning* (Cambridge, Mass., and London: MIT Press, 2001), 27–48.

Hornsby, Jennifer, 'Agency and Causal Explanation', in A. R. Mele (ed.), *The Philosophy of Action* (Oxford and New York: Oxford University Press, 1997), 283–307.

Hussain, Nadeem J. Z., and Shah, Nishi, 'Is Constructivism an Alternative to Realism?' MS (2005a).

____ 'Metaethics and its Discontents: A Case Study of Korsgaard', MS (2005b).

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

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

____ 'Normativity, Necessity, and the Synthetic a priori: A Response to Derek Parfit', MS (2003a).

____ 'Realism and Constructivism in Twentieth-Century Moral Philosophy', *Journal of Philosophical Research*, APA Centennial Supplement (2003b), 99–122.

Prichard, H. A., 'Does Moral Philosophy Rest on a Mistake?' in *Moral Writings*, ed. J. MacAdam (Oxford and New York: Clarendon Press, 2002a), 7–20.

____ 'Duty and Interest', in *Moral Writings*, ed. J. MacAdam (Oxford and New York: Clarendon Press, 2002b), 21–49.

____ 'Kant's *Fundamental Principles of the Metaphysic of Morals*', in *Moral Writings*, ed. J. MacAdam (Oxford and New York: Clarendon Press, 2002c), 50–76.

Stalzer, Kenneth, 'On the Normativity of the Instrumental Principle', Ph.D., Philosophy, Stanford University (2004).

11

Resisting the Buck-Passing Account of Value

Pekka Väyrynen

1. Introduction

T. M. Scanlon's "buck-passing account" of value continues a long tradition of analyzing value in terms of non-evaluative normative notions.[1] Buck-passers about value hold (speaking roughly for now) that to be valuable is nothing more or other than to have other properties that provide reasons for certain positive responses—namely, certain "pro-attitudes" and/or actions expressive of them—to the bearers of those properties. This is to pass the normative "buck" from value onto other properties: the reasons to favor valuable things are provided not by their value but by the properties that make them valuable (Scanlon, 1998: 97). To illustrate, as the prospects of reaching Mordor turn bleak and Frodo Baggins's spirit falters, Samwise Gamgee tries to lift Frodo's mood with an evaluative claim: "There's some good in this world, and it's worth fighting for."[2] According to the format of analysis favored by buck-passers, the fact that something is worth fighting for would just be the fact that it has other properties that provide

I presented an earlier version as "The Buck-Passing Account of Value (Almost) Refuted" at the 2004 Wisconsin Metaethics Workshop and the *Practical Reason and Moral Motivation* meeting in Rome. (Sometimes the direction of progress with work-in-progress is towards wimpier titles.) I am grateful to these audiences for helpful discussion, and to Russ Shafer-Landau for organizing a wonderful workshop. Many thanks to Jonas Olson, Philip Stratton-Lake, Jussi Suikkanen, and an anonymous referee for comments on earlier drafts which led to numerous improvements. I am indebted to Christian Coons for very helpful conversations during the early stages of the paper.

[1] Scanlon introduces the buck-passing account of value in his (1998: 95–8). Other recent proponents of the view include Parfit (2001), Suikkanen (2004), and Stratton-Lake and Hooker (2006). Rabinowicz and Rønnow-Rasmussen (2004) provide a useful overview of the tradition Scanlon continues.

[2] The line is from the movie *The Lord of the Rings: The Two Towers.*

reasons to fight for it. In this paper, I first clarify my target by addressing questions about buck-passers' format of value analysis, and about its scope in particular. I then build a resistance front to the buck-passing account of value by raising problems for its various forms.

2. What is a Buck-Passing Account of Value?

The basic idea of buck-passing is easy to grasp, but presentations of the view leave it unclear what exactly the view is supposed to be. I begin by clarifying my target on buck-passers' behalf. As we'll see, the buck-passing account comes in many forms which we must address separately.

Scanlon introduces his view by considering the relations between properties that "can be grounds for concluding that [something] is valuable... the property of being valuable, and the reasons that we have for behaving in certain ways in regard to things that are valuable":

> There seem to be two possibilities. The first is [Moore's view] that when something has the right natural properties it has the further property of being valuable, and that property gives us reason to behave or react in certain ways with regard to it... [Contrary to Moore, I believe] that being good, or valuable, is not a property that itself provides a reason to respond to a thing in certain ways. Rather, to be good or valuable is to have other properties that constitute such reasons.... [T]his account takes goodness and value to be... the purely formal, higher-order properties of having some lower-order properties that provide reasons of the relevant kind.... [I]t is not goodness or value itself that provide reasons but rather other properties that do so. For this reason I call it a buck-passing account. (Scanlon, 1998: 97)

Here Scanlon advances two theses about value, one negative and the other positive:

(BP$^-$) Being good, or valuable, isn't itself a reason-providing property. The fact that an object *o* is good, or valuable, isn't itself a reason to respond to *o* in certain favorable ways.[3]

(BP$^+$) Being good, or valuable, *just is* the purely formal, higher-order property of having other property or properties *P* that provide reasons to respond to things having *P* in certain favorable ways.

[3] I understand 'object' broadly to include any type of value-bearer. Facts or true propositions are better candidates than properties for things that have the property *being a reason*. So when Scanlon writes of properties as being what *provide* reasons, I take him to mean that for a property *P* to provide a reason to (say) favor *o* is for the fact that *o* has *P* to be (i.e. to have the property of being) a reason to favor *o*. It is important that the reasons be practical: a thing needn't be valuable merely if there are reasons to investigate or reflect on its properties.

The reasons in question are *justifying* practical reasons. Scanlon (like many others) paraphrases ‘a reason’ for something as “a consideration that counts in favor of it” (1998: 17). He seems to assume that if a thing’s having a certain property *makes* it valuable, thereby giving a reason *why* it is of value, then that property provides a reason *for* certain favorable responses to it.[4] Buck-passing is meant to provide a formal account of value, so it should be silent on *which* properties provide reasons; intuitive candidates include pleasure, health, and knowledge.[5] Whichever these properties are, what matters is that their instances have the property of being reasons.[6] Scanlon leaves it unclear which pro-attitudes constitute the relevant favorable responses, but he notes that they “generally include, as a common core, reasons for admiring the thing and for respecting it” (1998: 95).

Buck-passers often present (BP^-) as essential to their view, apparently because they assume that (BP^-) follows from (BP^+).[7] Scanlon also appears to assume that (BP^+) and the negation of (BP^-)—that is, the view that goodness is a reason-providing property, which Scanlon (correctly or not) attributes to G. E. Moore—exhaust the options. But each of these assumptions is mistaken. Buck-passers should define their view just in terms of its positive thesis.

Take the first assumption first. Given (BP^+), the fact that an object *o* is valuable amounts to the following higher-order fact:

(HOF) *o* has properties (other than being valuable) which provide reasons to respond to *o* in certain favorable ways.

Substituting (HOF) for the fact that *o* is valuable, (BP^-) says that (HOF) doesn’t itself constitute a reason to respond to *o* in the relevant favorable ways. That claim doesn’t follow from (BP^+) alone. We can agree that (HOF) neither gives any *additional* reason to favor *o* beyond the reasons provided by the properties that make *o* valuable nor is what *ultimately* provides the reason to favor *o instead of* the properties that make *o* valuable. But what if I know (perhaps by testimony) that something has properties

[4] See Scanlon (1998: 97) and Stratton-Lake and Hooker (2006: 153). I’ll grant this assumption, but we should note that the relation of making something valuable is distinct from the relation of providing a reason to respond to it in a certain way (cf. Dancy, 2004: 79–80). The former relates (tokens of) value properties to (tokens of) other properties, the latter relates (tokens of) those other properties to attitudes and/or actions.

[5] This is the sense in which pleasure, health, knowledge, and so on, are sometimes said to be “values”.

[6] Whether their instances must be reasons always, or only in certain circumstances, depends on such further issues in the theory of reasons as whether some form of holism about reasons is correct (see s. 4 below).

[7] See e.g. Scanlon (1998: 97), Parfit (2001: 19–20), and Stratton-Lake and Hooker (2006: 149).

that provide reasons to favor it, without knowing what those properties are? One option is to say that (HOF) is a reason to *believe* that there is a reason to favor *o* but is not itself a reason to favor *o*. Another is to say that (HOF) serves as a *derivative* reason to favor *o* which is accounted for by other facts about *o*. The latter idea would be that (HOF) states the sort of thing that can be a reason for action: in the circumstances in question, acting in the light of (HOF) would qualify as acting for a reason, and acting for a reason requires acting on the basis of the sort of thing that can be a reason for action. [8] Since (BP^+) is completely silent on which of these views is correct, (BP^-) doesn't follow from (BP^+) alone (unless it is revised to say that being valuable isn't among the properties that *ultimately* provide reasons).

(BP^+) and the view which (BP^-) negates also don't exhaust the options. One of Scanlon's arguments for the buck-passing account is an argument from intuitions about reasons:

> [W]hen I consider particular cases it seems that the [reasons to choose, prefer, recommend, and admire things that are valuable] are provided by the natural properties that make a thing good or valuable. So, for example, the fact that a resort is pleasant is a reason to visit it or to recommend it to a friend, and the fact that a discovery casts light on the causes of cancer is a reason to applaud it and to support further research of that kind. . . . It is not clear what further work could be done by special reason-providing properties of goodness and value, and even less clear how these properties could provide reasons. (Scanlon, 1998: 97)

Notice that it is perfectly coherent (i) to accept Scanlon's intuition about which sort of properties (ultimately) provide reasons, (ii) to accept that whenever something is valuable, it has the higher-order property of having other properties that provide reasons, but (iii) to hold that this higher-order property is distinct from the property of being valuable. Since Scanlon's argument fails to eliminate any such view, it fails as an argument for (BP^+) even if it succeeds as an argument for (BP^-). [9]

Buck-passers about value should define their view just in terms of their positive thesis, then. But a number of issues remain about how (BP^+)

[8] I am indebted to Michael Smith for making this point in conversation.

[9] This criticism of the argument is essentially Dancy's (2000: 164–5), with inessential simplifications and minor modifications. Scanlon's other argument for (BP^+) is "the fact that many different things can be said to be good or to be valuable, and the grounds for these judgments vary widely. There does not seem to be a single, reason-providing property that is common to all these cases" (1998: 97–8). Stratton-Lake and Hooker (2006: 156–7) show that this argument fails as well. It assumes the plurality of the good whereas both the buck-passing account and its rivals are neutral as between pluralism and monism. If hedonism, for example, were the correct substantive axiology, then the buck-passing account would also imply that all instances of value have in common a single ultimate reason-providing property.

should be understood. To begin, is the thesis meant to be metaphysical or conceptual? Buck-passers tend to go back and forth between describing their view as a doctrine about facts and properties and describing it at the level of language or concepts.[10] But the possibility of defining evaluative concepts in terms of the concept of a reason is consistent with rejecting much of what buck-passers say in their metaphysical mode. Defining one concept in terms of another doesn't always settle the direction of metaphysical priority between what the two concepts are concepts of. Hence I'll take buck-passers' core thesis to be that reasons are metaphysically prior to value, in that the property of being valuable isn't metaphysically independent but is analyzable in terms of the property of having reason-providing properties.[11]

The downside of construing buck-passers' core thesis as a metaphysical claim is that buck-passers tell us very little about how they think of properties.[12] The issue matters. One way to defend (BP^+) would be to argue that the property of being valuable is necessarily co-extensive with the sort of higher-order property that figures in (BP^+) and then appeal to the necessary co-extension test for property-identity: for any properties *A* and *B*, if *A* is necessarily co-extensive with *B*, then *A* and *B* are the same property. Scanlon himself cannot appeal to this argument for (BP^+). He denies that we can identify the property of being valuable with any non-normative property (1998: 96). But it is possible to construct a (possibly infinite disjunctive) property expressed in purely descriptive terms, which is necessarily co-extensive with the property of being valuable (Jackson, 1998: 118–25). Given how sets are individuated, the identity of necessarily co-extensive properties is difficult to avoid if we think of a property as the set of all its actual and possible instances (Lewis, 1999). Thus, if buck-passers accept this conception of properties, they'll have difficulty avoiding the conclusion that the property of being valuable and the property of being reason-providing are identical to properties expressible in purely descriptive terms.[13] Since buck-passers

[10] Scanlon speaks of the view indiscriminately in metaphysical terms and as the conceptual claim that "to call something valuable is to say that it has other properties that provide reasons for behaving in certain ways with regard to it" (Scanlon, 1998: 96). Suikkanen (2004) and Stratton-Lake and Hooker (2006) are similarly undisciplined.

[11] In my usage of 'analysis', analysis is a specification of properties rather than concepts (see e.g. King, 1998).

[12] The buck-passing account allows for deflationary or minimalist conceptions of normative properties. It is also neutral between cognitivist and non-cognitivist accounts of normative judgment.

[13] I doubt we can understand buck-passers' use of the term 'property' as purely *pleonastic:* as taking every meaningful predicate to express a property and two predicates

don't tell us how they think of properties, I'll bracket the issue and ignore arguments for (BP$^+$) premised on the necessary co-extension test for property-identity.[14]

Recent arguments that (BP$^+$), as it stands, is an extensionally inadequate statement of buck-passers' positive thesis are also relevant to interpreting the view. These arguments offer cases where we appear to have reasons to respond favorably to things that aren't valuable. To illustrate this "wrong kind of reasons" problem, imagine that an evil demon is determined to punish me unless I admire him for his determination to punish me. I have a good reason to admire the demon's determination for its own sake, namely that I'll avoid severe pain if I do so, but clearly the reason I have to admire his determination is of a wrong kind to make it valuable.[15] Here I grant that buck-passers can restrict their positive thesis (in some appropriately formal way) to all and only the right kind of reasons.[16] The standard view of higher-order properties is that they are generated by quantification over some set **B** of lower-order "base" properties plus a condition on members of **B** (Kim, 1998: 19–20). In those terms, I grant that there is some extensionally adequate and appropriately formal specification of condition *R* in the following restatement of (BP$^+$):

(BP$^{+\prime}$) Being good, or valuable, just is the property of having some property *P* in **B** such that *R*(*P*), where *R* specifies a condition on members of **B** which is satisfied just by those properties in **B** that provide the right kind of reasons to respond to their bearers in certain favorable ways.

to express different properties if they are non-synonymous. In the pleonastic sense, the property of being valuable is distinct from the property of having other properties that provide reasons, unless (implausibly) 'is valuable' is synonymous with 'has other properties that provide reasons'.

[14] Buck-passers will eventually need to deal with the nature of properties. Suppose e.g. that hedonism turns out to be the correct substantive axiology, so that being valuable and being pleasant are necessarily co-extensive, but that buck-passers are right that being valuable doesn't (ultimately) provide reasons. In that event, nor could the property of being pleasant provide reasons, unless either properties, no matter how they are individuated, provide reasons only under certain descriptions, or else properties are individuated more finely than by necessary equivalence.

[15] Rabinowicz and Rønnow-Rasmussen (2004) discuss examples like this in great detail.

[16] For recent attempts, see e.g. Olson (2004) and Stratton-Lake (2005). An adequate solution should also address certain technical issues about how to formulate the buck-passing account. For example, if something is the lesser of two bads, it has a property that provides a reason to prefer it to the greater bad. Since the lesser bad might still be quite bad, it cannot have positive value simply because it has properties that provide reasons to prefer it. I ignore the issue because it should be possible to state the buck-passing account so as to handle betterness and worseness.

In what follows, I'll take the refinement as understood; below we'll see another purpose it serves.

One test for the plausibility of identifying goodness with the higher-order property in ($BP^{+\prime}$) is whether the latter property would fill the goodness role in the mature evaluative practice of the folk (Jackson and Pettit 1995). Since we don't know how the mature folk valuations will construe the goodness role, confidence that the "role property"—that is, the higher-order property of having the property that plays the goodness role—will be the higher-order property in ($BP^{+\prime}$) would be premature. Notice, for example, that although the supervenience of value properties on non-evaluative properties partly specifies the goodness role, ($BP^{+\prime}$) doesn't entail that practical reasons are ultimately provided only by non-evaluative properties. A further specification of the goodness role might be that something is valuable only if a subject would desire it if she satisfied all rational requirements and other ideals of reason. This, too, fails to support the buck-passing account, since not all rational requirements (for example, those of instrumental rationality) are analyzable in terms of a suitable sensitivity to reasons.[17]

The attractive assumption that *value is normative* is silent as well on whether the higher-order property in ($BP^{+\prime}$) is what fills the goodness role. For value to be normative is for it to make a difference to what one ought or has reason to do. If something is valuable, it merits certain favorable attitudes and there are reasons (at least for suitably situated agents) to adopt those attitudes.[18] I'll assume that value is intimately tied to pro-attitudes and reasons in this way. ($BP^{+\prime}$) explains that intimate tie by reducing value to reasons for the relevant attitudes. But the intimate tie is amenable to other explanations. It might be that something is good when it merits certain favorable attitudes, or when it has properties that provide reasons

[17] See e.g. Broome (2002) and Smith (forthcoming).

[18] The qualification in parentheses hides more than the idea that a reason must always be assigned to an agent who is in a position to act on it. It might turn out further that an agent has a reason only when she satisfies some "internalist" or other subjective condition on justifying reasons. Such conditions raise complications that I have no space to discuss, such as whether the buck-passing account would imply a corresponding subjective condition on value, and whether such a condition would be plausible. In this connection, I should also mention a related structural problem to which buck-passers have yet to give a convincing reply (*pace* Suikkanen, 2004: 531–3). The problem is Dancy's polyadicity objection. Something can be good (or bad) without a specification of the agent, whereas reasons always belong to agents; reasons don't hang around waiting to be assigned to agents. Therefore, no matter how many argument-places goodness has, it is less polyadic than reasons are. But if reasons are polyadic to degree *n*, then the higher-order property of having other properties that provide reasons is also polyadic to degree *n*. Therefore that higher-order property is more polyadic than goodness, in which case the two properties must be distinct. See Dancy (2000: 170).

to respond favorably to it, and yet that goodness isn't reducible to either of these things. Perhaps, for example, we have reasons to have certain attitudes to valuable things because those attitudes are appropriate to the value of the things in question. In that case, value would be something more fundamental that explains the reasons. Again, we cannot assume that specifying the goodness role generates an argument for the buck-passing account of value.

My focal question about buck-passers' positive thesis concerns its *scope:* to just *which* value properties is the buck-passing account of value supposed to apply? Buck-passers tend to speak only of being good, or valuable, but it is natural to wonder why their basic format of analysis shouldn't apply to other value properties as well. For ($\mathrm{BP}^{+\prime}$) is just an instance of the following general schema (where *V* is a value property variable and **B** a set of base properties):

> (BP*) Being *V* just is the purely formal, higher-order property of having some property *P* in **B** such that *R*(*P*), where *R* specifies a condition on members of **B** which is satisfied just by those properties in **B** that provide the right kind of reasons to respond to their bearers in certain favorable ways.

In (BP*) the qualification 'right kind' is more than a placeholder for a solution to the wrong kind of reasons problem. We need it to distinguish distinct value properties from each other. As Scanlon notes, what attitudes the reason-giving properties justify may be different in different cases (1998: 95). Different bearers of a particular value property may call for different attitudes, as with elegance in philosophical argument, elegance in dress, and elegance in chord change.[19] More importantly, instances of different value properties are often to be valued by means of different attitudes, as with being admirable and being trustworthy. Unless the right kind of reasons are those that bear specifically on whether something has a particular value property, there may be distinct value properties for which (BP*) yields the same analysans. Not all value properties, however, bear the kind of analytic connection to the relevant responses which would make it straightforward to analyze, for example, trustworthiness as possession of properties that provide reasons for trust.

To cash out this aspect of the qualification without the circularity in saying that instances of a value property call for those attitudes that are *appropriate* to their value, buck-passers might apply (BP*) in the light of our pre-theoretical views about what responses different value properties call

[19] Indeed, different responses may be apt to an elegant chord change in a jazz tune and in a punk rock song.

for (see Rabinowicz and Rønnow-Rasmussen, 2004: 402). So understood, (BP*) generates a recipe for analyzing different value properties at least partly in terms of the different pro-attitudes which we have reasons to adopt towards their bearers.[20] But *which* value properties?

Applying (BP*) to different sets of value properties generates different forms of the buck-passing account. But, almost without fail, presentations of the view don't specify the intended scope of (BP*). Since (BP*) itself is so schematic that it is hard to know how to argue for or against it, we need to proceed by assessing more concrete forms of the view. One fruitful way to divide the options is to note that buck-passing may be either *all-out* or only *partial.* That is to say, relative to a view of what properties count as value properties in the first place, either *every* value property is the sort of purely formal higher-order property we find in (BP*), or only *some* are.[21] For example, suppose we hold the permissive view that being intrinsically valuable, being morally valuable and being prudentially valuable (and the like), being kind and being generous (and the like), being elegant and being delicate (and the like), and being admirable and being desirable (and the like) all count as value properties in our normative sense. Then all-out buck-passing would hold that every single one of these value properties is the kind of higher-order property we find in (BP*). Different forms of partial buck-passing would restrict the scope of (BP*) only to different proper subsets of these value properties, and treat the rest as some more substantive sort of value properties.

In what follows, I proceed from the premise that the buck-passing account is either partial or all-out in its scope. The scope of the buck-passing account then depends on whether we can draw the kind of distinction among value properties which the truth of partial buck-passing requires, and how we draw it. Section 3 argues that the extant forms of partial buck-passing fail to draw such a distinction; hence they don't succeed in restricting themselves only to some proper subset of value properties. Section 4 builds a cumulative case for resisting all-out buck-passing. Section 5 criticizes a further positive argument for buck-passers' approach to value and offers brief concluding remarks.

[20] Nozick (1981: 429–30) provides a whopping 40-item list of ways of responding to value. As Rabinowicz and Rønnow-Rasmussen (2004: 416) note, the relevant pro-attitudes may have to have a complex intentional content: they may have to consist in favoring an object, in one way or another, on account of some of its properties.

[21] The relativity of the distinction to a set of value properties introduces the complication that materially one and the same view of the scope of (BP*) may be partial relative to one view of what counts as a value property but all-out relative to another. This makes the distinction less neat, but needn't diminish its heuristic value.

3. Resisting Partial Buck-Passing

Most buck-passers advance some form of partial buck-passing. While partial buck-passers rarely make the intended scope of their views explicit, their writings nonetheless contain a few suggestions as to how to restrict the scope of (BP*) only to some proper subset of value properties. I'll argue that, in each case, the form of partial buck-passing in question is unstable.

We have seen that buck-passers tend to talk only of the property of being good, or valuable, in stating their view. A literal interpretation is that (BP*) applies only to a property of being valuable which all and only valuable things have in common, in addition to whatever other value properties they may have—whether they are valuable intrinsically or extrinsically, for their own sakes or instrumentally, whether they are morally or aesthetically valuable, whether they are kind or courageous, or admirable or desirable, and so on.[22] In constructing an analysis of such a wholly generic value property, we must keep in mind that different types of valuable things may be valuable in different ways, in that they may call for different pro-attitudes. Letting **B** be a set of base properties and W range over pro-attitudes, the proposal must be something like this:

> (GBP) For any valuable object x, for x to be of value just is for x to have some property P in **B** such that, for some way of valuing W which x calls for, x's being P is a reason to respond to x in way W.

My objection to (GBP) relies on an assumption about properties which is plausible in the present context: instances of a property should exhibit some substantial commonality. We don't think that a wholly heterogeneous set of things as such makes a difference to what one ought or has reason to do. If (GBP) is to be an adequate analysis of a normative property, the condition it imposes on P should determine a class of ways of valuing that exhibit a substantial commonality. For (GBP) implies that there may be nothing more to being of value than the relevant similarities among the different attitudes with which we have reason to respond to different types

[22] Here I'll let pass the point that I myself find such a wholly generic property of being valuable obscure, for I find it hard to see what substantive commonality all and only the things that are valuable in all these very different ways are supposed to share. The point is akin to Judith Thomson's line on generic goodness (2003 and elsewhere). To be clear, my view is that, whatever Thomson's own intentions may be, her arguments truly target only the claim that there is such a thing as generic goodness, and *not* the claim that there are such properties as being intrinsically good or being valuable for its own sake. For we can treat the latter as ways of being good.

of valuable things. Those things may have nothing else in common, since to provide a reason is always to provide a reason for particular responses. The objection is that there may not be enough similarities to go around among the different ways of responding to valuable things for them to exhibit any substantial commonality. If so, the condition in (GBP) determines only a heterogeneous set of things.

One way to advance this objection is to argue that no sufficiently definite account may be available of what distinguishes positive responses ("pro-attitudes") from the negative ones ("con-attitudes"), and both from responses that are neither. The most promising way to draw these distinctions is based on the idea that pro-attitudes and con-attitudes are distinguished from attitudes that are neither by their involving some conative element and are distinguished from one another by the nature of the conative elements they involve.[23] If all pro-attitudes essentially involved a favorable conative element, (GBP) would determine a substantive conative commonality. But it seems that I can respect or appreciate various valuable things without being moved by them. I can understand that some operas have properties that provide reasons to respect them and various activities involving them, and yet not be irrational or pathological if they fail to engage me conatively. I can admire the way in which a jazz solo moves back to the root chord as neat or nifty, and in that sense appreciate its aesthetic value on the basis of my knowledge of the conventions of jazz, while being unmoved by jazz. Examples like these suggest that there is no guarantee that all pro-attitudes share a common conative core that distinguishes them from con-attitudes.[24] (The problem extends to separating pro- and con-attitudes as a class from the class of responses that are neither.) Thus we cannot assume that the condition on *P* in (GBP) determines a set of things with a substantial conative commonality.[25]

[23] Here I follow Rabinowicz and Rønnow-Rasmussen (2004: 401).

[24] Since my point here concerns attitudes, it doesn't seem to presuppose motivational externalism about reasons-judgments. But if you suspect that it does, note that motivational reasons-judgment externalism may be plausible even if motivational ought-judgment externalism isn't. The judgment that a consideration is normatively relevant in the way that reasons are may sometimes lack the kind of deliberative relevance which motivationally efficacious considerations have. Ought-judgments, by contrast, may well carry greater deliberative relevance.

[25] If the different responses that we have reasons to adopt towards different valuable things lack a common core that would make for a substantial commonality among all and only the valuable things, then (GBP) makes the generic notion of value indefinite. The indefiniteness doesn't result from ordinary phenomena such as the existence of borderline cases in the application of our evaluative language. Rather, (GBP) requires us to assume (controversially, to say the least) that properties themselves can be metaphysically indefinite. A fully generic notion of value may be indefinite in another way as well: many

Analyses can certainly be surprising. The analysans in (GBP) has much more structure to it, however, than one would have thought the analysandum even covertly to possess. Because of this, (GBP) seems more plausible as an analysans of the property of being valuable in a given particular way than of a wholly generic value property. Perhaps the property of being admirable, for example, is analyzable as possession of properties that provide reasons to admire their bearers. If that is the best way to read buck-passers' talk of "being good, or valuable," then what they offer us is a recipe for analyzing particular ways of being valuable: each way of being valuable is to be analyzed as possession of properties that provide (the right kind of) reasons to respond in certain specified ways W. This would broaden the scope of (BP*) so much that it would border on all-out buck-passing. In any case, it won't do for partial buck-passers to restrict themselves merely to (GBP).

The obvious alternative for partial buck-passers is to find some distinction that classifies *kinds* of value property in a way that explains why only some kinds of value property are the sort of purely formal higher-order property we find in (BP*) whereas others are some more substantive kind of value properties. In fact, the writings of buck-passers point to at least *two* kinds of distinctions among value properties which might do the job. I'll discuss these in turn.

Buck-passers tend to present their view by contrasting it with the view that goodness is a reason-providing property, which they (correctly or not) attribute to Moore. Moore shares the tendency to speak of "goodness" and "value," but this is clearly sloppy on his part, as his real concern is with *intrinsic* value. Insofar as buck-passers mean to contrast their view with the view they take to be Moore's, they should claim that the property of being intrinsically valuable is analyzable as an instance of (BP*).[26] Assuming that intrinsic value is value that something has solely in virtue of its intrinsic properties, buck-passers would presumably analyze it as possession of intrinsic properties that provide reasons. But buck-passers' own examples preclude the restriction of (BP*) *solely* to intrinsic value. Scanlon's example of a holiday resort is relevant only if a pleasant resort has some type of value, but whatever type of value it has is presumably not intrinsic.

understand the talk of being valuable as shorthand for the talk of being valuable in some particular way. If asked which things are of value, many of us seek to identify whichever things we identify on the basis of the varied responses we think they merit, not on the basis of some substantive commonality. Of course, this does nothing to enhance the interest of a fully generic notion of value to value analysis.

[26] Curiously, Scanlon often qualifies 'good' and 'valuable' with 'intrinsically' in the discussion that precedes his presentation of the buck-passing account (1998: 88, 90–2), but drops the qualification when presenting it.

We can understand the value of a pleasant resort in one of two ways. A pleasant resort might have *final* value: it might be valuable *for its own sake*. (Final value is distinct from intrinsic value: an object might be valuable for its own sake partly in virtue of its being rare, but being rare isn't an intrinsic property.) More plausibly, a pleasant resort might have *instrumental* value: it might be valuable for the sake of something else. Perhaps the causal and constitutive relations it bears to other things make it conducive to something that is valuable for its own sake, such as pleasurable experiences. Whichever account we adopt of the value of a pleasant resort, we analyze its value in terms of final value, since being instrumentally valuable is analyzable in terms of being finally valuable.[27] Thus, if (BP*) is true of instrumental value, it must be true of final value as well.[28] One motivation to apply (BP*) to final value is that final value has a clear connection to practical reason. Thus one form of partial buck-passing restricts (BP*) to final value, and perhaps intrinsic value, plus any other value properties that are analyzable in terms of those two.[29]

Restricting the buck-passing account to final value suffers from essentially the same problem as (GBP).[30] A reason to favor something for its own sake may be either instrumental or non-instrumental.[31] Thus the buck-passing account of final value should be something like (FBP):

(FBP) For any object *x*, for *x* to be of final value just is for *x* to have some property *P* in the base set **B** such that, for some way of

[27] To some, Scanlon's example of a good resort might suggest that (BP*) applies to what Ross (1930: 65–7) calls "attributive goodness". This is the property of being good of a kind, of satisfying the standards of excellence in a kind. Buck-passers shouldn't apply (BP*) to attributive goodness because being good of a kind isn't necessarily connected to reasons. Consider the property of being a good assassin: for persons to satisfy the standards of excellence in assassinating isn't necessarily for them to have properties that provide reasons to respond favorably to them.

[28] Whether (BP*) is to be applied to instrumental value depends on the controversial issue whether instrumental value is itself a form of value at all, instead of something merely conducive to value. Another notion that seems analyzable in terms of final value is that of being "contributively good", that is, being such as to contribute to the final value of the whole of which it is a part (cf. Ross, 1930: 72). Whether (BP*) is to be applied to "contributive value" depends on the controversial issue whether contributive value is itself a form of value, rather than merely a relation to value (see n. 49). In this respect, contributive value is analogous to instrumental value.

[29] As Jonas Olson pointed out to me, the idea that buck-passing is primarily a view about final value is probably the traditional idea (see e.g. Ewing, 1947: 146).

[30] The parallel objection can be run against forms of partial buck-passing that restrict (BP*) to intrinsic value.

[31] For discussion, see Stratton-Lake (2005) whose solution to the wrong kind of reasons problem exploits this point.

> valuing *W* which *x* calls for, *x*'s being *P* is a non-instrumental reason to respond to *x* in way *W* for its own sake.

To see why (FBP) doesn't support restricting (BP*) to final value (and whatever other value properties are analyzable in its terms), notice that pretty much any attitude that is a possible value for *W* in (FBP) is one that we may have reason to hold for a thing's own sake in some cases but for the sake of something else in others. For example, some things are admirable for their own sakes but others are admirable only for the sake of something else. According to (FBP), they are finally valuable only if they have properties that provide non-instrumental reasons to admire them. These instances of admiration aren't different *in kind*. Indeed, your attitude of admiration is no different in kind if you admire people's keeping their promises but not because you think of promise-keeping as valuable in any way. What then distinguishes the responses that pertain to value from those that don't?

These observations show not that (FBP) is mistaken as such, but that it does little if anything to specify a particular class of different pro-attitudes that we may have non-instrumental reason to adopt to different kinds of things. This is a problem for forms of partial buck-passing built upon (FBP) because if (as argued above) the different pro-attitudes lack a common substantial core, then so do the pro-attitudes quantified over in (FBP). Reasoning that parallels our objection to (GBP) then shows that to apply (BP*) only to final value (and whatever value properties are analyzable in its terms) is merely to apply selectively a general recipe for analyzing different ways of being valuable. If so, (FBP) gives partial buck-passers no independent grounds for restricting (BP*) merely to final value, and the general recipe itself again broadens the scope of (BP*) so much that it borders on all-out buck-passing. An adequate restriction of (BP*) only to some subset of value properties requires some other type of distinction among value properties.

One distinction that we might take to explain why only some kinds of value property are the sort of higher-order property we find in (BP*) appears in Scanlon's response to a tension between his presentation of the buck-passing account and his account of practical reflection on reasons. In discussing the buck-passing account in *What we Owe to Each Other*, Scanlon claims that reasons are typically provided by the natural properties of things rather than their goodness or value (1998: 97).[32] But he also suggests that judgments about reasons involve a distinctively "evaluative element" (1998: 38), that

[32] Thanks to Philip Stratton-Lake for reminding me that Scanlon doesn't adhere to the claim throughout *What we Owe to Each Other*. He claims (although he probably shouldn't) that the property of being wrong is reason-providing, but doesn't regard it

often one's most important reason for doing what would satisfy a desire one has is that it would be worthwhile or honorable (1998: 49), and that the task of practical reflection on reasons is to characterize the concrete forms of value that can be achieved in action (1998: 65–9). Jay Wallace takes this tension to indicate that Scanlon's real concern is with "the relation between general and specific concepts" (2002: 447). He reinterprets the buck-passing account as claiming that general evaluative claims can be "understood as ways of signaling that there is some specific reason for action in the offing, a reason that can be characterized by specifying the particular way in which the action in question would be valuable" (Wallace, 2002: 448). Scanlon endorses this reinterpretation: "My thesis was that goodness is not itself a property that provides reasons, not that the underlying properties that do this are always natural properties... more specific evaluative properties often play this role" (Scanlon 2002: 513).[33]

One form of partial buck-passing then restricts (BP*) to "general" as opposed to "more specific" value properties. Wallace's comment on Scanlon's example of a pleasant resort illustrates the view:

> To say that a resort is "pleasant," for instance, is a way of adverting to the distinctively positive qualities of experience that are enjoyed by a visitor to the resort. It is not merely an evaluatively neutral description of the natural properties of the resort or of the experiences induced by the resort in its visitors, and this is what makes it appropriate to think of pleasure itself as a concrete category of evaluation. (Wallace, 2002: 448)

Wallace doesn't say what he means by 'advert', but he assumes that pleasure is a form of value that provides reasons.[34] It is, however, unclear what makes a value property count as specific or general. (If general value properties are meant to be such properties as being of final value, the view faces the problems discussed above.) The illustration is partly to blame. Being pleasant, like being conducive to pleasant experiences, is a singularly

as a natural property (Scanlon, 1998: 10–12, 147–8). For a discussion of the relation between buck-passing about value and buck-passing about rightness, see e.g. Dancy (2000: 165–7).

[33] Of course, for reasons given in s. 2, I think that the emphasis Scanlon places here on (BP$^-$) is misleading.

[34] On one reading, what Wallace means by 'advert' is that to call something pleasant is to recommend it, perhaps in the sense of ascribing to it a positive value property. But the fact that speakers can use a term to recommend shows neither that it is a value term nor that its referent is a value property. A more plausible sense in which calling something pleasant is a way of adverting to the presence of value is that one pragmatically presupposes or implicates that it instantiates some positive value property, without implying that being pleasant is itself a value property. This happens if e.g. we operate with the substantive but cancellable assumption that pleasure is good.

bad example of a specific value property. *Schadenfreude* and the pleasant experiences that activities such as sadism and genocide induce in some people give us no reason to respond to those activities or experiences favorably (at least, not as parts of these wholes). If so, being pleasant isn't necessarily connected to reasons in the way that buck-passers suppose value to be. It is less plausible to regard being pleasant as a value property than as a property that can make its bearers better (or worse, depending on the context).[35] But the distinction between specific and general value properties is meant to distinguish *among* properties to instantiate which is to *be* valuable (in a particular way), and we cannot do that by appealing to properties that *make* their bearers valuable. The illustration fails to appreciate the distinction between being valuable and making something valuable.

Even if we found better examples of specific value properties, the distinction between general and more specific value properties would have the wrong kind of structure to restrict (BP*) only to some value properties. Whatever the distinction is supposed to be (and this remains unclear), generality and specificity are *relative* and *gradable* notions: one thing (say, beneficence) can be general relative to another (such as kindness) and yet specific relative to a third (such as virtue), and relative generality comes in degrees. As such, the generality/specificity distinction tells us nothing as to where, on the continuum of relative generality vs. specificity, an ascription of a value property is supposed to be an ascription of the sort of purely formal higher-order property we find in (BP*) rather than an ascription of some more substantive value property. The problem, of course, is that the distinction marks only a difference in degree among value properties, whereas any form of partial buck-passing requires a difference in kind between purely formal higher-order value properties and substantive value properties. Hence the generality/specificity distinction as such gives us no grounds not to apply (BP*) throughout the continuum if we apply it anywhere.[36] Any plausible restrictions on the scope of (BP*) must have some other source.

[35] We can interpret Wallace's occasional talk of "particular forms" and "concrete modalities" of value accordingly.

[36] There are other distinctions which partial buck-passers might deploy in lieu of the generality/specificity distinction. For example, one might try passing the buck from properties that mark the genus 'value' onto its species, or from determinable value properties onto their determinates. Both options face the problem of where to draw the line between purely formal and substantive value properties. For example, a property can be a determinable relative to one property but a determinate relative to another: consider being red, being scarlet, and being colored. And species of value may themselves be genera that include more specific value properties. Even if we drew the line by

A different distinction to which partial buck-passers might appeal in this neighborhood is that there is an intuitive sense in which "thick" value properties, such as kindness and generosity, are specific, whereas "thin" value properties, such as being of final value, are general. We might then think that each thin value property is the sort of higher-order property we find in (BP*) whereas thick value properties are some more substantive sort of value properties that are eligible to provide reasons.[37] But this proposal faces a dilemma.

A familiar dispute about thick value properties is whether they can be "disentangled" into distinct non-evaluative and thin evaluative components. Either they can or not. If they cannot, then they are eligible to provide reasons (see below). If they can, then we can analyze, for example, the property of being generous (as a property of persons) as a disposition to act in certain (non-evaluatively specifiable) ways towards others, plus the fact that this disposition has thin value. In that case buck-passers about thin value deny that the property of being generous is eligible to provide (ultimate) reasons and instead take the reason-giving property to be the non-evaluative component of generosity (that is, the disposition to act in certain ways towards others).[38]

On the one hand, then, if the disentanglement claim is true, the normative buck continues onto the non-evaluative components of thick value properties. I suspect this broadens the scope of (BP*) beyond thin value properties. If a buck-passer offered just an account of the thin value component of a thick property, not an account of the property as a whole, the account would have trouble distinguishing different thick value properties from each other. For example, it would have trouble accounting for the differences between the responses for which, say, generosity, bravery, and elegance call without appealing to our notions of generosity, bravery,

restricting (BP*) to those determinable value properties that don't themselves fall under any determinable, or those genus properties that don't themselves fall under a genus of value, these options would face the further problem that neither genus nor determinable properties are, in general, purely formal higher-order properties. Determinables, such as being shaped, mark genuine categories of difference, but not merely in virtue of their determinates; the parallel point goes for genus properties, such as being a mammal. In that case determinable and genus properties wouldn't count as instances of (BP*) simply in virtue of being determinables or of marking a genus. So, neither distinction is structurally cut out to restrict (BP*) only to some proper subset of value properties.

[37] Although the literature usually speaks of thick and thin evaluative *terms* or *concepts*, I'll discuss the property version of the distinction in order to maintain my focus on metaphysical issues. Perhaps the thin/thick distinction is what Wallace and/or Scanlon really have in mind, although if that is the case I wonder why they don't just say so.

[38] I owe this distinction between buck-passers' options on thick properties to Stratton-Lake and Hooker (2006: 152).

and elegance. The account would also be inadequate to the standard characterization of thick value properties as those that satisfy evaluative concepts whose applicability is both world-guided, in the sense of being constrained by non-evaluative criteria, and action-guiding, in the sense of indicating reasons for action.[39] For it would be adequate only to the "action-guiding" conjunct. It would be adequate to the "world-guided" conjunct if it also told us what the properties quantified over in the relevant instance of (BP*) are. (This would also help to distinguish different thick properties from each other.) But an account that captures both conjuncts broadens the scope of partial buck-passing from thin to thick value properties. For it makes thick value properties merely higher-order properties (albeit not purely formal ones). So, we have reason to think that if the disentanglement claim is true, forms of partial buck-passing built on the thin/thick distinction either are inadequate or border on all-out buck-passing.

On the other hand, if thick value properties cannot be disentangled into distinct evaluative and non-evaluative components, then they are eligible to play the reason-giving role. The normative buck won't continue onto a distinct non-evaluative component of a thick property, since the property *has* no distinct non-evaluative component to play the reason-giving role. This might seem like good news to partial buck-passing. Like the generality/specificity distinction, however, the thin/thick distinction marks only a difference in degree along a spectrum of value properties (Scheffler, 1987: 417–18), whereas partial buck-passing requires a difference in kind. Were the disentanglement claim true, we might try holding the thin value component constant and explaining differences in degree in terms of differences in the specificity of the relevant non-evaluative components. But if that isn't an option, the thin/thick distinction will tell us nothing as to where along the spectrum an ascription of a value property is supposed to be an ascription of the sort of purely formal higher-order property we find in (BP*) rather than an ascription of some more substantive value property. Hence the distinction as such gives us no grounds to apply (BP*) anywhere on the spectrum if we don't apply it to thick value properties, and no grounds not to apply it all across the spectrum if we apply it to thin value properties. So, we have reason to think that forms of partial buck-passing built on the thin/thick distinction border on all-out buck-passing.[40]

[39] See e.g. Williams (1985: 129, 140) and Hurley (1989: 11–13).

[40] If any thick value property has a distinct, self-standing non-evaluative component, that component typically is plausibly not analytically distinct from the evaluative component, but rather can be isolated only by substantive normative theorizing. This would seem to be in tension with buck-passers' claim to be advancing a formal account of value which is compatible with any substantive normative and evaluative theory.

So far I have argued that the forms of partial buck-passing surveyed above are unstable. A further source of pressure towards all-out buck-passing is that partial buck-passing cannot claim for itself a putative advantage of the buck-passing approach to value. The attraction is ontological parsimony: if value is analyzable in terms of reasons, what we might have regarded as two separate normative categories are reducible only to one. As Derek Parfit puts it, "in believing that certain aims are good, or worth achieving, [buck-passers] are not committed to normative properties other than the property of being reason-giving, or committed to normative truths other than truths about reasons" (Parfit, 2001: 38). Since partial buck-passers think that there are some substantive value properties, the argument from ontological parsimony can seemingly support all-out buck-passing at best.

If the above forms of partial buck-passing are unstable, this might be because value properties share some characteristics that explain their instability. That would unify the case that partial buck-passing is unstable. Some writers suggest that values have some kind of "unity" that distinguishes them from other values and in virtue of which their components hang together the way they do.[41] If a unity were a structural feature of value properties on many levels of generality, one would expect that either all value properties are purely formal higher-order properties or (more plausibly) none are.[42] The idea that value properties involve a kind of unity, in virtue of which they are structured as they are, is intriguing. It could explain why certain, but not all, possible ways of organizing the various aspects of value properties constitute distinct categories of evaluative difference (Raz, 2003: 133). But as I cannot explicate such a unity to my satisfaction, I rest my claim that partial buck-passing is unstable on my

[41] See e.g. Raz (2003: 39) and Chang (2004: 16). I don't claim that Raz or Chang intend this suggestion to speak against the buck-passing account of value, although at least Raz clearly rejects the buck-passing account.

[42] Typical examples of values that putatively have a unity, such as philosophical talent (Chang, 2004: 16–17), concern "values" in the presently irrelevant sense of properties that make things have their value properties (see n. 5). But the idea that value properties that combine certain constituting qualities without being simply reducible to them have some kind of unity has some intuitive pull. Aristotelian *eudaemonia* is a possible example: being *eudaemôn* collects together its various constituents, such as the virtues as well as certain types of pleasures and honors, and organizes and balances them with respect to each other in an evaluatively distinct way (see e.g. Stocker, 1990: 172). This mode of organization is evaluatively distinct because how well different options satisfy the claims of *eudaemonia* to be protected, aspired to, and so on, isn't simply a matter of how well they satisfy the claims of the constituents, considered merely as separate evaluatively relevant dimensions. What it is even to count as *eudaemonia* is a matter of combining the constituting qualities of *eudaemonia* in the right sort of way, the way exemplified by an excellent life. Perhaps *eudaemonia* has, in this sense, a distinctive sort of unity of its constituting qualities.

grounds for thinking that the particular forms of partial buck-passing discussed above are unstable. Since buck-passers usually take themselves to be partial buck-passers, this result, though inconclusive, is important.

4. Resisting All-Out Buck-Passing

All-out buck-passing strikes many people as incredible: what good reason could we have for thinking that *no* value property *ever* is eligible to provide reasons? One might wonder, though, whether this reaction implicitly begs some crucial question. Here I offer a case for resisting all-out buck-passing which isn't subject to this worry: buck-passers advertise their approach to value as metaethically neutral, but all-out buck-passing turns out to require controversial metaethical assumptions and, in addition, to incur troublesome explanatory debts. Giving an adequate defense of all-out buck-passing therefore requires defending its metaethical commitments and discharging its explanatory debts. The case for resisting all-out buck-passing is the stronger the more demanding this task is.[43]

We have already seen all-out buck-passing to incur one controversial metaethical commitment. It requires that each thick value property be analyzable as a possession of two distinct properties—namely, a certain non-evaluative property that is reason-providing and a "thin" value property—but this disentanglement claim is famously controversial.

While card-carrying buck-passers tend to be non-naturalists about normative and evaluative properties, they aren't non-naturalists qua buck-passers. The basic thrust of their approach to value is neutral between naturalism and non-naturalism. But all-out buck-passing turns out to be incompatible with certain sophisticated forms of evaluative naturalism. Consider, for example, the form of naturalism according to which value properties are clusters of mutually supporting physical, medical, psychological, and social goods unified by homeostatic mechanisms (Boyd, 1988: 203–4; cf. 194–9, 216–17). No value property that is a homeostatic unity is plausibly regarded as the sort of purely formal higher-order property we find in (BP*). According to all-out buck-passing, the reasons connected to the

[43] The fact, noted in s. 2, that the distinction between partial and all-out buck-passing is relative to a conception of what properties count as value properties complicates matters. Restrictive conceptions might count some forms of buck-passing that I construed above as forms of partial buck-passing as forms of all-out buck-passing instead. Permissive conceptions might make all-out buck-passing much more inclusive than its proponents would be willing to grant. My case for resisting all-out buck-passing won't be entirely immune to these complications, but mostly my discussion will require only relatively modest assumptions about what properties count as value properties.

presence of any such property would apparently have to be provided by the non-evaluative properties in the given cluster. We cannot identify the higher-order property of having such properties with the value property because the former leaves out part of the latter, namely the homeostatic unity among the clustered properties.

In response, all-out buck-passers might try to include the relevant homeostatic mechanisms among the reason-providing properties. Suppose I have reason to engage in some co-operative effort. If a value property just is a group of homeostatically clustered goods, the co-operative effort must, if it is to be valuable, support and be supported by other goods, such as friendship and recreation, via the psychological and social mechanisms that contribute to the homeostasis. Apparently what provides me the reason to engage in the effort would have to be either (i) that doing so will tend to foster the realization of these goods and sustain the homeostatic mechanisms on which their unity depends or (ii) that doing so will tend to foster co-operation. Either way, all-out buck-passing is inconsistent with homeostatic naturalism. Given the homeostatic naturalist conception of value properties, (i) implies that, contrary to all-out buck-passing, the reason I have is provided by a value property. Regarding (ii), suppose that engaging in co-operative effort sometimes does but at other times doesn't tend to foster the realization of the goods in question and sustain the homeostatic mechanisms on which their unity depends. If so, the tendency of some activity to foster co-operation sometimes does, but sometimes doesn't, provide reasons to favor it. Given homeostatic naturalism, what the effort fosters when there is reason to favor it is the instantiation of a value property. But in that case we can explain the variability of reasons by reference to value: something is a reason to ϕ in one case because ϕ-ing fosters an instantiation of a value property but isn't a reason to ϕ in another case because ϕ-ing fails to do so. Of course, if a value property explains why certain considerations have the property of being reasons, it cannot be the sort of purely formal higher-order property we find in (BP*).

Let's move on to the explanatory debts of all-out buck-passing.[44] I argued earlier that the trouble partial buck-passers have with finding distinctions among value properties which would explain why only some value properties should be the sort of purely formal higher-order property we find in (BP*) generates internal pressure towards all-out buck-passing. In order for the lack of such distinctions to favor, rather than count against, all-out buck-passing, all value properties must in addition be shown to be

[44] I actually think that all-out buck-passing incurs yet further controversial metaethical commitments, but have no space to argue the point here.

purely formal higher-order properties rather than some more substantive sort of properties. I'll argue that such a project has dubious prospects.

Consider generosity. We might analyze it in part as the disposition to benefit others out of one's own resources without being intrusive or expecting esteem or compensation, and perhaps out of sympathy for their ends. Given that being generous is a value property (to attribute it to people surely is to evaluate them), what bestows the higher-order property of having other properties that provide reasons for certain pro-attitudes on this complex disposition? All-out buck-passers cannot say that the disposition itself does so if we characterize it partly in evaluative terms like 'benefit'. They should specify the goods with which, and the ends in pursuit of which, a generous person is disposed to aid others in the specified kind of way in non-evaluative terms. They might, for example, analyze being generous as (*a*) having the disposition to desire or pursue for others, without being intrusive or expecting esteem or compensation, those resources of one's own which one would desire or pursue for them if one cared for them for their own sakes plus (*b*) the fact that having that disposition provides reasons to take certain pro-attitudes to its bearers.[45]

Analyses of thick value properties along these lines incur serious explanatory debts. Why, for example, are the non-evaluative aspects of generosity related as they are? The properties that make someone beneficent don't provide reasons for the attitudinal responses for which generosity calls when, for example, she also intends to gain others' esteem or expects compensation. What explains why the presence of the latter properties makes this kind of difference between reasons of generosity and reasons of beneficence? One possible explanation is value-based. The way in which generosity organizes its non-evaluative aspects gives it evaluative aspects that beneficence lacks, for we take the two properties to bear differently on agents' moral worth. But then attitudes that are appropriate to generosity are not appropriate to the evaluative nature of esteem-seeking beneficence. But if it is *because* of the distinctive evaluative nature of generosity that its bearers have the higher-order property of having properties that provide the relevant reasons, generosity is distinct from that higher-order property. All-out buck-passers owe us an explanation that is superior to the explanation premised on the assumption that generosity is a substantive value property.[46]

[45] I intend (*a*) as a rough approximation of the buck-passing account of welfare defended in Darwall (2002).

[46] All-out buck-passers cannot discharge this explanatory debt simply by saying that truths about reasons are the basic normative truths. The question of what explains the difference in reasons is perfectly legitimate, and no less legitimate if the notion of a reason for something, paraphrased as a consideration that counts in favor of it, is primitive

All-out buck-passers might seek to undercut the explanation of reasons in terms of value properties by trying to capture the differences between generosity and beneficence in terms of reasons. The most promising strategy seems to be to locate the differences in the reason-providing properties, rather than in the attitudes. Perhaps, for example, the properties that make something beneficent provide reasons to favor their bearers in ways appropriate to beneficence without providing reasons to favor them in ways appropriate to generosity when they are parts of "wholes" whose other parts are properties such as intending to gain others' esteem or expecting compensation.[47] If the ways in which a whole is valuable are determined by its parts in some sense holistically, we might try to capture this evaluative structure in terms of an analogous holism of reasons.[48] One version of this idea is that, even if the beneficence-making properties are pretty much the same as the generosity-making properties, the other parts of a beneficent whole may entail the absence of those background conditions (such as not seeking esteem) which enable the properties that make it beneficent to make it generous. Such a view would require some potentially controversial claims about which properties of generous persons provide reasons for the relevant attitudes and which amount merely to the necessary background conditions for the properties in question to provide those reasons. But the general project would be to emulate the structure of a value property in terms of the conditions under which the base properties quantified over in the relevant instance of (BP*) provide reasons for the attitudes for which the bearers of the value property in question call.

I find this project problematic. Suppose a property which gives reasons to respond to its bearer in a certain way does so only in the presence of some

(Scanlon, 1998: 17). Even if the notion of a reason is a conceptual primitive, it doesn't follow that there is no explanation of why a certain consideration has the property of being a reason. Then it doesn't follow that truths about what sort of differences in the non-evaluative properties of things make what sort of differences to our reasons are primitive truths. Moreover, the buck-passing account as such doesn't entail that truths about reasons are the basic normative truths, for it is consistent with the Humean view that the reasons that agents have are grounded in their desires. Scanlon (1998: 41–9) rejects the Humean view because of his further claim that reasons are the most basic normative elements of practical reason. Here I'll ignore Humean buck-passers about value.

[47] Instead of "wholes" we might speak of objects and their context.

[48] It might be that the value of a whole is determined by the values of its parts "organically", perhaps in the way Moore (1993: 79–81) thought, or that while the value of a whole is some non-organic function of the values of its parts, the values of those parts are contextually conditional (see Dancy, 2003; 2004: 176–84). In the latter case, the relevant sort of holism of reasons would be roughly that of Dancy (2004: 73, 38–43); in the former, it might be roughly that of Ross (1930: 19–20, 41–2). Unfortunately I have no space here for a fuller discussion of these issues.

conditions that enable it to give those reasons. Then the presence of such an enabling condition will play a role in the sustenance of value, thereby giving reason to preserve its presence. For were the condition not to obtain, the property that gives the reason wouldn't do so. If enabling conditions involve properties that provide reasons, all-out buck-passers must either regard them as valuable in some way or explain why these reasons don't ground value. The latter option faces the general problem of explaining why some properties that give reasons, but not others, ground value. (Purely deontological reasons to fulfill promises come to mind as another example of the latter.) The former option faces the problem that it is more plausible to treat the kind of "enabling value" in question as a relation to value than a relational form of value. While a condition that itself has no value cannot contribute value to an object, it may perfectly well play a vital enabling role with respect to the object's value.[49] For example, when an object has final value partly in virtue of being a unique instance of its kind, the non-existence of other instances appears to be a valueless condition, but one that enables the object to have final value.[50] In sum, doctrines about reasons with which buck-passers might seek to emulate the structure of value properties have problematic consequences when conjoined with all-out buck-passing.

Even if all-out buck-passers can defuse these worries, their task won't be finished. Finding doctrines about reasons which mirror the relevant doctrines about value does nothing to settle the question of which are explanatorily prior. All-out buck-passers must further show that the latter doctrines are better explained by the former than vice versa. But the converse direction of explanation is a strong contender. For example, the reasons that enabling conditions provide in virtue of the role they play in the sustenance of value is readily explained in terms of value and without a commitment to regard enabling conditions as having any special relational form of value.

A related explanatory debt of all-out buck-passing is to explain how, for any value property V, what distinguishes V from other value properties is solely a function of the reasons in terms of which it analyzes V. As we saw in section 2, buck-passers can avoid the circularity in saying that

[49] There are related grounds to doubt that "contributive value" (recall n. 28) is a distinct form of value. Philip Stratton-Lake writes: "For something to be contributively valuable is for it to stand in a better-making relation to the whole of which it is a part. Contributive value is, therefore, a relational form of value" (2002: 127). If a part cannot contribute to a whole more value than it actually has as a part of that whole (Dancy, 2003: 630–1), only what is otherwise valuable can be contributively valuable. But this does nothing to show that contributive value is itself a relational form of value, as opposed to a relation to value which it is possible only for valuable parts to instantiate.

[50] For a different sort of example, and a fuller discussion, see Dancy (2003: 634–5).

a value property calls for those attitudes that are appropriate to its value by appealing to our pre-theoretical views about the attitudes for which different value properties call. But this exposes them to the worry that our pre-theoretical views might distinguish finely enough neither between the responses for which closely related but distinct value properties call nor between the properties that provide those reasons. If so, there might be pairs of distinct value properties which are too similar in both respects for either to distinguish the properties. Buck-passers owe us some systematic account of why drawing the right distinctions won't in fact be a problem.

The worry is pressing in the case of value properties that bear no analytic connection to appropriate responses in the way that being admirable or being trustworthy do. As we now conceive all-out buck-passers' strategy, what suffice to raise the worry are mere pre-theoretical possibilities to the effect that two distinct value properties are associated with reasons that are too similar to distinguish the properties. For example, it seems pre-theoretically possible for the correct substantive theory of welfare to imply that we should respond to welfare subjects as if they were friends, that is, respond to them with the same kinds of attitudes, and on the same kinds of grounds, as we respond to friends.[51] (Such a theory wouldn't imply that we should *make* friends with welfare subjects.) Given a view that counts welfare and friendship as value properties, the application of (BP*) to each would in that event deliver the same higher-order property, when it shouldn't. This would be a reason not to identify either property with that higher-order property.

The example presupposes that the pre-theoretical data about welfare and friendship are consistent with the idea that we should respond to welfare subjects as if they were friends. Pre-theoretically, however, the properties that provide reasons of welfare and reasons of friendship do seem similar enough not to distinguish the two properties. For example, insofar as we think (as all-out buck-passers must) that the properties that provide these reasons are non-evaluative, prominent among them are the needs, interests, and desires of friends and welfare subjects. The relevant responses also seem similar enough. Reasons of welfare and those of friendship are reasons to respond to certain individuals, in whatever ways are appropriate, for their own sakes, and it is pre-theoretically possible that the responses are similar enough not to distinguish the two properties. Perhaps, in both cases, the relevant responses are those characteristic of a loving concern. In both cases, then, the relevant reason-providing properties and responses

[51] I am indebted to Christian Coons for suggesting this possibility. Ruling it out with a substantive conception of welfare would violate the spirit of all-out buck-passing as a formal analysis of value.

seem pre-theoretically similar enough not to distinguish between the purely formal higher-order properties with which all-out buck-passers would identify friendship and welfare.[52] The worry appears to generalize.[53]

Towards the end of section 3, I noted that the argument from ontological parsimony seems at most to support all-out buck-passing. As a form of theoretical economy, parsimony is only a defeasible merit: greater parsimony is preferable, but only insofar as all else is at least roughly equal. In this section, I have in effect argued that the other things aren't roughly equal for all-out buck-passing. We have seen that all-out buck-passing requires controversial metaethical assumptions, and that we may doubt whether all-out buck-passers can do better than their opponents in discharging certain explanatory debts concerning value properties and their relation to reasons. Hence ontological parsimony fails, at least for now, to provide any significant source of support for all-out buck-passing. Taken together, the above worries about all-out buck-passing constitute a good cumulative case for resisting it.

5. Conclusion

My resistance to the buck-passing account of value takes the form of a dilemma. Proceeding from the assumption that any form of the account is either all-out or partial in its scope, I first argued that the forms of partial buck-passing I surveyed don't succeed in restricting themselves only to certain proper subsets of value properties, and then built a resistance front to all-out buck-passing. Because buck-passers' basic format of value analysis is so schematic that it can be wielded in a diverse array of ways, my argument strategy against buck-passers has been to spray a buckshot of considerations against particular ways of wielding the format. In closing, I'll criticize a further positive argument for the buck-passing approach to value and offer some tentative positive suggestions.

Suppose buck-passers' opponents (*a*) accept that, whenever something is valuable (in a particular way), it also has the sort of purely formal

[52] Jussi Suikkanen suggested to me that it might be partly constitutive of friends' concern for each other that they have together formed a view of each others' needs, interests, and desires on some shared basis. The same doesn't seem true of an appropriate concern for non-friend welfare subjects, even if we should respond to welfare subjects as if they were friends. I think more needs to be said about how the suggestion is supposed to distinguish friendship and welfare from one another, rather than merely to distinguish the conditions for the presence of welfare-related and friendship-related reasons (which reasons may pre-theoretically be very similar to each other).

[53] Roger Crisp (2005: 82) has independently raised a very similar objection, using grace and delicacy as his example.

higher-order property we find in (BP*), but (*b*) regard this higher-order property as distinct from the property of being valuable (in that way). Philip Stratton-Lake and Brad Hooker argue that buck-passing is a better option for those who agree with the negative thesis that the fact that something is valuable never adds to the reasons provided by the properties that make it valuable. This is because the opposition "leaves unexplained why goodness *cannot* provide us with an additional reason," whereas

> the buck-passing account of goodness explains *why* the fact that something is good never gives us a reason to care about it. On the buck-passing account, the fact that something is good is the fact that it has other properties that provide reasons to care about it, and the fact that it has such properties *cannot* provide an *extra* reason to care about it. (Stratton-Lake and Hooker, 2006: 161)

Stratton-Lake and Hooker in effect claim an exclusive explanatory advantage to buck-passing.

The argument needs refinement, however, given the different possible scopes that (BP*) can take. Stratton-Lake and Hooker should claim that, for any value property *V* to which (BP*) applies, only buck-passers can explain why the fact that something is *V* never ultimately gives us reasons to respond to it in those ways for which its being *V* calls. This claim has a narrower appeal. If the disentanglement claim about thick value properties is false, any thick property is a better candidate for the relevant reason-providing property than the non-evaluative properties co-instantiated with it. Those non-evaluative properties are better candidates only if the disentanglement claim is true. But the opposition has resources to explain why thick value properties would in that event never provide extra reasons to respond to their instances in the relevant ways.

In discussing all-out buck-passing, I mentioned the view that we can appeal to value properties to explain the reasons that are necessarily connected to their instantiation. If the disentanglement claim about thick value properties is true, such a view could explain why the non-evaluative component of generosity (say) provides reasons for certain attitudes to generous things by saying that adopting those attitudes on account of the property in question is a response that is adequate to the way in which those things are valuable. If something's being generous provides an *explanatory* reason why certain of its purely non-evaluative properties give reasons for the attitudes in question, it is reasonable to suppose that being generous (or other thick properties) never provides an extra *practical* reason for those attitudes. For what could be the point of such double duty? The value property would already have made its difference to what we have reason to do. Since Stratton-Lake and Hooker's argument ignores accounts of this sort, it is persuasive only in conjunction with independent arguments against them. Such arguments would, of course, seem to make theirs superfluous.

If we have good reasons to resist the buck-passing approach to value, where should we look for an alternative account of the relation between value and reasons? If value is necessarily connected to reasons but we cannot adequately account for truths about value in terms of truths about reasons, we should doubt that reasons are metaphysically primary. The intuitive default position seems in any case to be that they aren't. It is surely no accident that some considerations but not others count as reasons. Often we can explain *why* a consideration is a reason, and are unsatisfied if we cannot. (The question why a consideration possesses the kind of normative force that is characteristic of reasons is especially natural when its content is non-normative.) We might not always be able to appeal to value properties to explain why the properties that provide us with reasons for certain kinds of responses do so. For we might think that there are deontological reasons that have nothing to do with value. We could, however, try to construct a general schema for explaining reasons which doesn't apply exclusively to value properties.

In many cases the explanation of why a consideration with non-normative content is a reason might well go in terms of a value property. For example, if a sculpture like Bernini's *The Ecstasy of St Theresa* is sublime because of the double-faceted facial expression it portrays, it would seem quite natural to explain why this feature of the sculpture gives us reasons for certain responses by describing its relation to aesthetic sublimity. To say this isn't to deny that the instantiation of the former property ontologically grounds or realizes that of the latter. For that claim doesn't settle the normative question of why, when the latter property is a value property (or some other kind of normative property), the instantiation of the former property should have such relevance to the latter's instantiation. The ontological dependence of an instantiation of a value property on an instantiation of a non-evaluative property is one thing. The normative dependence relation in which an instance of the non-evaluative property stands to that on which it depends for its having the property of being reason-giving (a value property, perhaps) is different. Another illustration of this distinction would be a form of welfarism about reasons which grants that instances of the property of being good for a person are ontologically grounded in certain non-evaluative properties and that all sorts of considerations besides welfare might function as reasons, but holds that any consideration that *does* function as a reason depends for its having the property of being a reason on promotion of welfare. In other cases the explanation of reasons might not proceed in terms of value. It might instead proceed in terms of deontological notions such as rights, fairness, or duty, or etiquettal notions, and so on.

If an explanatory schema of this sort were generally applicable, it would provide us with considerable explanatory gains. The resulting hypothesis about the relation between reasons and value would accommodate the negative insights that buck-passers emphasize. But it would avoid worries about distinguishing reasons that give rise to value from reasons that don't, as well as the other worries I have raised about the buck-passing account. Since I have said very little to develop or support this hypothesis, however, it would be premature for me to endorse it. But the alternative it constitutes to buck-passers' positive approach to value seems worthy of further consideration.

REFERENCES

Boyd, Richard, 'How To Be a Moral Realist', in Geoffrey Sayre-McCord (ed.), *Essays on Moral Realism* (Ithaca, NY: Cornell University Press, 1988), 181–228.

Broome, John, 'Practical Reasoning', in José Luis Bermùdez and Alan Millar (eds.), *Reason and Nature: Essays in the Theory of Rationality* (Oxford: Clarendon Press, 2002), 85–111.

Chang, Ruth, 'All Things Considered', *Philosophical Perspectives*, 18 (2004), 1–22.

Crisp, Roger, 'Value, Reasons and the Structure of Justification: How to Avoid Passing the Buck', *Analysis*, 65 (2005), 80–5.

Dancy, Jonathan, 'Should we Pass the Buck?', in Anthony O'Hear (ed.), *Philosophy, the True, the Good and the Beautiful* (Cambridge: Cambridge University Press, 2000), 159–73.

____, 'Are there Organic Unities?', *Ethics*, 113 (2003), 629–50.

____ *Ethics without Principles* (Oxford: Clarendon Press, 2004).

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

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

Hurley, S. L., *Natural Reasons* (New York: Oxford University Press, 1989).

Jackson, Frank, *From Metaphysics to Ethics* (Oxford: Clarendon Press, 1998).

____ and Pettit, Philip, 'Moral Functionalism and Moral Motivation', *Philosophical Quarterly*, 45 (1995), 20–40.

Kim, Jaegwon, *Mind in a Physical World* (Cambridge, Mass.: MIT Press, 1998).

King, Jeffrey C., 'What is a Philosophical Analysis?', *Philosophical Studies*, 90 (1998), 155–79.

Lewis, David, 'New Work for a Theory of Universals', reprinted in *Papers in Metaphysics and Epistemology* (Cambridge: Cambridge University Press, 1999), 8–55.

Moore, G. E., *Principia Ethica*, rev. edn. (Cambridge: Cambridge University Press, 1993).

Nozick, Robert, *Philosophical Explanations* (Cambridge, Mass.: Harvard University Press, 1981).

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

Parfit, Derek, 'Rationality and Reasons', in Dan Egonsson *et al.* (eds.), *Exploring Practical Philosophy* (Aldershot: Ashgate, 2001), 17–39.

Rabinowicz, Wlodek, and Rønnow-Rasmussen, Toni, 'The Strike of the Demon: On Fitting Pro-Attitudes and Value', *Ethics*, 114 (2004), 391–423.

Raz, Joseph, *The Practice of Value* (Oxford: Clarendon Press, 2003).

Ross, W. D., *The Right and the Good* (Oxford: Clarendon Press, 1930).

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

—— 'Reasons, Responsibility, and Reliance: Replies to Wallace, Dworkin, and Deigh', *Ethics*, 112 (2002), 507–28.

Scheffler, Samuel, 'Morality through Thick and Thin', *Philosophical Review*, 96 (1987), 411–34.

Smith, Michael, 'Is there a Nexus between Reasons and Rationality?', in Sergio Tenenbaum (ed.), *New Trends in Philosophy: Moral Psychology* (Amsterdam: Rodopi, forthcoming).

Stocker, Michael, *Plural and Conflicting Values* (Oxford: Clarendon Press, 1990).

Stratton-Lake, Philip, 'Pleasure and Reflection in Ross's Intuitionism', in Philip Stratton-Lake (ed.), *Ethical Intuitionism: Re-evaluations* (Oxford: Clarendon Press, 2002), 113–36.

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

—— and Hooker, Brad, 'Scanlon versus Moore on Goodness', Terence Horgan and Mark Timmons (eds.), *Metaethics after Moore* (Oxford: Oxford University Press, 2006), 149–68.

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

Thomson, Judith Jarvis, 'The Legacy of *Principia*', *Southern Journal of Philosophy, supplementary vol.* 41 (2003), 62–82.

Wallace, R. Jay, 'Scanlon's Contractualism', *Ethics*, 112 (2002), 429–70.

Williams, Bernard, *Ethics and the Limits of Philosophy* (Cambridge, Mass.: Harvard University Press, 1985).

12

Normativity

Derek Parfit

1

A young Swiss guest of Richard Hare's, after reading a book by Camus, concluded in despair that *nothing matters*. Hare suggested that his friend should ask 'what was the meaning or function of the word "matters" in our language; what is it to be important?' His friend soon agreed, Hare writes,

> that when we say something matters or is important, what we are doing, in saying this, is to express our concern about that something . . . Having secured my friend's agreement on this point, I then pointed out to him something that followed immediately from it. This is that when somebody says that something matters or does not matter, we want to know *whose* concern is being expressed or otherwise referred to. If the function of the expression 'matters' is to express concern, and if concern is always *somebody's* concern, we can always ask, when it is said that something matters or does not matter, 'Whose concern?'[1]

As Hare pointed out, his friend *was* concerned about several things. So was everyone—except a few fictional characters in existentialist novels. People's values differ, and may change. But, since we all care about something, 'it is impossible to overthrow values as a whole'. Hare's treatment worked. 'My Swiss friend ate a hearty breakfast the next morning.'

If someone doubts whether anything matters, it may not help to ask 'Whose concern?' Hare managed to convince his friend

> that the expression 'Nothing matters' in his mouth could only be (if he understood it) a piece of play-acting. Of course he didn't actually understand it.

[1] 'Nothing Matters', in R. M. Hare, *Applications of Moral Philosophy* (London: Macmillan, 1972), 33–4.

There is, I believe, a use of the word 'matters' which Hare does not understand.

When Hare writes that we use such words to *express* concern, he is not, he claims, using 'express' in an 'emotivist' sense. 'I am no more committed to an emotivist view of the meaning of these words than I would be if I said "The word 'not' is used in English to express negation".' Despite this disclaimer, Hare does accept an emotivist or, more broadly, non-cognitivist view. That is why, when Hare's friend concluded that nothing mattered, Hare didn't try to remind him that some things, such as suffering, do matter. As Hare writes:

> My friend ... had thought mattering was something (some activity or process) that things did ... If one thinks that, one may begin to wonder what this activity is, called mattering; and one may begin to observe the world closely ... to see if one can catch anything doing something that could be called 'mattering'; and when we can observe nothing going on which seems to correspond to this name, it is easy for the novelist to persuade us that after all *nothing matters.* To which the answer is, ' "Matters" isn't that sort of word; it isn't intended to *describe* something ...'

On Hare's view, nothing can be truly described as mattering. The truth is only that we care about some things. In saying that these things matter, we are not claiming that they really do matter. Rather, as emotivists claim, we are expressing our concern.

Hare assumes that, in making these claims, he is not denying anything that others might mistakenly believe. There is nothing to deny, he claims, since no other view makes sense. He imagines an objector saying:

> All you have done is to show that people are *in fact* concerned about things. But this established only the existence of values in a *subjective* sense. Now, it may be said, when people talk about the overthrow of values, they do not mean anything so far-fetched as that people should stop being concerned about things ... But values are overthrown if it is shown ... that these subjective feelings of people are all that there is; that values are not (as I have heard it put) 'built into the fabric of the world'. This objection, then, is a challenge to moral philosophers ... to demonstrate what has been called 'the objectivity of values'.

Philosophers, Hare answers, should reject this challenge. There are not two possibilities here, or two genuinely conflicting views. In Hare's words:

> I do not understand what is *meant* by the 'objectivity of values', and have not met anybody who does ... suppose we ask 'What is the difference between values being objective, and values not being objective?' Can

> anybody point to any difference? In order to see clearly that there is *no* difference, it is only necessary to consider statements of their position by subjectivists[2] and objectivists, and observe that they are saying the same thing in different words . . . An objectivist . . . says, 'When I say that a certain act is wrong, I am stating the *fact* that the act has a certain non-empirical *quality* called 'wrongness'; and I *discern* that it has this quality by exercising a faculty which I possess called 'moral intuition'. A subjectivist says, 'When *I* say that a certain act is wrong I am expressing towards it an attitude of disapproval which I have.'

It is true that, as Hare implies, these sentences could be used so that they did not conflict. Hare's objectivist might agree that, when he claims some act to be wrong, he is expressing his disapproval of this act. But this objectivist would mean that he is expressing his belief that this act has the property of being wrong. And, on Hare's view, there is no such property. Acts can't *be* wrong; the truth is only that we disapprove of them. When Hare claims that there is no disagreement here, since these people are saying the same thing, he misinterprets the objectivist's view. He assumes that, when objectivists claim that some acts really are wrong, they cannot mean what they seem to say. They cannot be intending to say something that is, in a strong sense, *true*.

Hare continues:

> We all know how to recognize the activity which I have been calling 'saying, thinking it to be so, that some act is wrong'. And it is obvious that it is to this activity that the subjectivist and the objectivist are both alluding. This activity . . . is called by the objectivist 'a moral intuition'. By the subjectivist it is called 'an attitude of disapproval'. But in so far as we can identify anything in our *experience* to which these two people could be alluding by these expressions, it is the same thing—namely the experience which we all have when we think that something is wrong.

When objectivists claim that certain acts really are wrong, they are not referring to the experiences that we have when we believe something to be wrong. Their claim is about *what* we believe. More exactly, it is about what some of us believe. They would concede that some people—such as some subjectivists, relativists, or sceptics—do not have such beliefs.

Hare might reply that *he* has such beliefs. He is discussing the activity of 'saying, *thinking it to be so*, that some act is wrong.' In thinking that to be so, he believes that this act really is wrong. His point is that such beliefs are

[2] By 'subjectivists' Hare means non-cognitivists, not those cognitivists who believe that normative statements are factual claims about our own attitudes. 'Nothing Matters', 40.

not like ordinary, *descriptive* beliefs. In thinking something to be wrong, we are not believing something to be true, but accepting the imperative 'No one ever act like that!' If Hare gave this reply, however, he would be conceding that there is a disagreement here. According to objectivists, these beliefs *are* descriptive.

Hare then considers another way in which some objectivists explain their view. They claim that, when moral judgments conflict, at least one of these judgments must be mistaken. Subjectivists, they then argue, cannot make that claim. Hare replies that, though such a claim explains objectivity in some other areas, it does not, when applied to morality, draw any 'real distinction'. In his words:

> Behind this argument lies, I think, the idea that if it is possible to say that it is *right* or *wrong* to say a certain thing, an affinity of some important kind is established between that sort of thing, and other things of which we can also say this. So, for example, if we can say of the answer to a mathematical problem that it is right, and can say *the same thing* of a moral judgment, this is held to show that a moral judgment is in some way *like* the answer to a mathematical problem, and therefore cannot be 'subjective' (whatever that means).

That is what it means.

Hare concludes:

> Think of one world into whose fabric values are plainly objectively built; and think of another in which those values have been annihilated. And remember that in both worlds the people in them go on being concerned about the same things—there is no difference in the 'subjective' concern which people have for things, only in their 'objective' value. Now I ask, What is the difference between the states of affairs in these two worlds? Can any other answer be given except 'None whatever'?

The analogy with mathematics, though only partial, also helps here. According to some empiricists, arithmetical truths are contingent. If we ask what makes it true that $5 + 3 = 8$, the answer is that, when people add 3 to 5, they nearly always get the answer 8. This view, we may object, misunderstands the nature of mathematics. Arithmetical truths are not contingent, or empirical, but necessary. Such an empiricist might reply:

> Your talk of necessity adds nothing. Imagine another world which is just like ours, except that in that world mathematical truths are not, as you claim, necessary. In both that world and ours, there would be no difference in the calculations of mathematicians. They would reach just the same answers. What is the difference between these worlds? None whatever.

This would not be a good reply. This empiricist would be right to claim that there is no conceivable difference between two such worlds. But that is because his imagined world is inconceivable. We cannot coherently suppose that 5 plus 3 did not, necessarily, equal 8. Since such truths are necessary, they are true in every possible world. And they are, in every world, necessary.

Hare, similarly, asks us to imagine two worlds. In the objectivist's world, 'values are plainly objectively built'. It is in a strong sense true that, for example, intense suffering is in itself bad, or that we have reason to prevent it, if we can. Such truths are irreducibly normative, and their denial is a mistake. In the subjectivist's world, there are no such truths, since objective values 'have been annihilated'. Everything else, however, is just the same. There is, Hare claims, no conceivable difference between these worlds. That is similarly true, but only because one of these worlds is inconceivable.

I have left it open which world this is. Though no one denies that there are mathematical truths, many deny that there are any normative truths. We shall return to some of the grounds for that denial. Our present question is only whether the idea of normative truths, and of objective values, makes sense.

Hare claims that it does not, as is shown by our inability to describe a difference between his two imagined worlds. But that inability should be explained in a different way. On both of the possible views about the objectivity of values, we cannot coherently imagine both these worlds. Suppose first that, as most objectivists believe, intense suffering really is bad. That, if true, is a necessary truth. There could not be a world in which intense suffering was otherwise just the same, but was not bad. Suppose next that, as Hare believes, it makes no sense to suppose that badness is a property that suffering might have. In his words, 'mattering' is not something that suffering could do. If that is so, there could not be a world in which suffering *was* bad. On neither view could there be two worlds, in only one of which was suffering bad. According to objectivists, such normative truths hold in every possible world. According to subjectivists, they hold in none. That is one difference between these views.

Hare might give a different reply. He might concede that, when objectivists claim that suffering is bad, they mean something different from what subjectivists mean. Hare believes that, if objectivism is put forward as a *moral* view, it is self-defeating. As he writes elsewhere:

> moral judgments cannot be merely statements of fact, and . . . if they were, they would not do the jobs that they do do, or have the logical characteristics that they do have. In other words, moral philosophers cannot have it both ways; either they must recognize the irreducibly

> prescriptive element in moral judgments, or else they must allow that moral judgments, as interpreted by them, do not guide actions in the way that, as ordinarily understood, they obviously do.[3]

As this passage shows, Hare ignores the possibility that there might be normative truths. He claims that, if moral judgments were capable of being true, or of stating facts, they could not guide actions. But, if we judged that we ought to do something, that judgment could guide our acts. So Hare must assume that, even on the view that he is opposing, judgments like 'I ought to do that' could not conceivably be true.

2

Many other writers ignore the possibility that there might be normative truths. And, of those who mention this possibility, many do not take it seriously. According to Brandt, for example, it is 'logically possible' that there are truths about what we have most reason to want. But such truths, he claims, would have less rational significance than facts about what, after informed deliberation, we would want. Brandt could not have made that claim if he had really thought that there might be such truths. Similarly, Gibbard regards this possibility as too fantastic to be worth considering.

There are good reasons to have this attitude. Irreducibly normative truths, if there are any, are most unusual. As many writers claim, it is not obvious how such truths fit into a scientific world-view. They are not empirically testable, or explicable by natural laws. Nor does there seem to be anything for these truths to be *about*. What can the property of badness *be*?

Given these points, it is natural to doubt whether these alleged truths even make sense. If such truths are not empirical, or about features of the natural world, how do we ever come to understand them? If words like 'reason' and 'ought' neither refer to natural features, nor express our attitudes, what could they possibly mean?

Non-reductive realists, as I have conceded, do not give helpful answers to these questions. According to them, we can explain some normative concepts, but only by appealing to others. Thus, in calling suffering bad, we mean that suffering is a state that we have reason to prevent, or relieve, or that we ought to prevent it, if we can. But normative concepts cannot be explained in non-normative terms. Nor can we say much to explain how we understand these concepts, or how we recognize normative truths. And,

[3] R. M. Hare, *The Language of Morals* (Oxford: Oxford University Press, 1952), 195.

when we ask *why* there are such truths, or what *makes* them true, the most that we can do is to explain some of these truths by appealing to others. We soon reach truths for which we can give no further explanation. Many diseases are bad, for example, because they cause suffering; but we cannot say what makes suffering bad.

Though we cannot give helpful answers to such questions, that does not show that there are no normative truths. Normative concepts form a fundamental category—like, say, temporal or logical concepts. We should not expect to explain time, or logic, in non-temporal or non-logical terms. If there are normative truths, these are of a distinctive kind, which we should not expect to be like ordinary, natural truths. Nor should we expect our knowledge of such truths, when we have it, to be like our knowledge of the world around us.

There are some helpful analogies. One example is the category of modal concepts, such as *possible* and *necessary*. Truths are necessary if they could not conceivably be false, or if they hold in every possible world. The concept of necessity cannot be explained in empirical terms, necessary truths are not made true by natural laws, nor is our knowledge of such truths like our knowledge of the actual world.

I shall not try here to defend the view that there are some irreducibly normative truths. My aim will be only to make clearer their distinctive feature: normativity.

One way to make that feature clearer is to describe cases in which normativity is most obviously present. That can be easily done. Two such cases are the badness of suffering, and someone's reason to jump from some burning building. But examples can be misunderstood. Normativity can be confused with other features of the case.

Rather than merely saying where normativity can be found, some writers try to explain what normativity is. But, for the reason I have just given, that cannot be helpfully done. We can ask what normative concepts, such as *ought* and *reason* mean. But there are no answers to these questions that are both interesting and true.

There are some interesting answers, such as those given by naturalists and non-cognitivists. These answers are interesting because they seem to be informative, and, if they were true, they would have important implications. Some of these would be substantive conclusions about what we have reason to want, and to do. Others would be conclusions about the metaphysics and epistemology of ethics, and practical reasoning.

These answers cannot, I believe, be true. Though we cannot explain what normativity is, or what normative concepts mean, we can say what normativity is *not*, and what these concepts do not mean. It could not be true that, as naturalists claim, normative statements mean the same as, or report the same facts as, statements about natural facts. Nor could these statements, as non-cognitivists claim, have merely an emotive or prescriptive sense. For these statements to be normative, they must be capable of being, in a strong sense, true.

Naturalists get one thing right, since they see that there are normative truths. But they mistakenly conflate these truths with the natural facts which, according to these truths, have normative importance. Non-cognitivists avoid this mistake, since they see that normativity cannot be reduced to, or consist in, such facts. They recognize the categorical difference between what is and is not normative. But they mistakenly take this difference to be between facts and the attitudes which they call 'values'.

Non-cognitivists, and many naturalists, get something else right. With their emphasis on motivation, these people see that practical reasoning is not concerned only with beliefs. For us to be fully practically rational, our normative beliefs must motivate us, and, when relevant, lead us to act. But non-cognitivists mistakenly conclude that these beliefs cannot really be beliefs. And both groups reduce normativity to motivating force. If we have most reason to act in some way, or ought rationally to do so, that is not a fact about, or an expression of, some desire or other motivating state.

If we believe in irreducibly normative truths, we are what Korsgaard calls *dogmatic rationalists*. As Korsgaard notes, since these rationalists have little positive to say, they are 'primarily polemical writers', who explain and defend their view by attacking other views.[4] That is what, in this essay, I shall mostly do. As Korsgaard also notes, 'the criticism of an opponent's position is normally the weakest part of a philosophical work'. But, given my beliefs about normativity, I have no alternative.

3

Many writers, I have claimed, ignore the possibility that there might be normative truths. Nowell-Smith, for example, writes: 'Moral philosophy is a practical science; its aim is to answer questions of the form 'What shall I do?' 'But', he warns, 'no general answer can be given to this type of question'. That is an understatement. As Nowell-Smith notes, the word 'shall' is ambiguous. Thus, in saying 'What shall I feel?', we ask for a prediction of

[4] Christine Korsgaard, *The Sources of Normativity* henceforth *Sources* (Cambridge: Cambridge University Press, 1996), 31.

our feelings, which others might correctly give. But, in asking 'What shall I do?', we are not trying to predict our acts. We are trying to make a decision. If moral philosophy had the aim of answering such questions, it could not possibly succeed. Philosophy cannot make our decisions.[5]

Nor can other people. When we ask 'What shall I do?', that is not a question to which even the wisest adviser could give an answer. If I say, 'That's what I shall do', others might say, 'No you won't', or 'No you shan't'. But those would not be conflicting answers to my question. They would be either a prediction, or the expression of a contrary decision—as when a parent says 'You *will do* what I tell you to.'

As these remarks suggest, the question 'What shall I do?' is not normative, nor can it be, as Nowell-Smith claims, 'the fundamental question of ethics'. The fundamental question is: 'What *should* I do?' Since that question *is* normative, it might have answers that philosophy, or other people, could give. There might be truths about what we should do.

Nowell-Smith considers this objection. It may be said, he writes,

> that the fundamental question is not 'What shall I do?' but 'What ought I to do?' and the fundamental concept not decision but obligation.[6]

He replies:

> My reason for treating the 'shall' question as fundamental is that moral discourse is practical. The language of 'ought' is intelligible only in the context of practical questions, and we have not answered a practical question until we have reached a decision.

Though moral discourse is practical, that does not imply that its fundamental question is about what we *shall* do rather than about what we *ought* to do. If we ask moral questions, that may be because we have decided that we shall do, or shall try to do, whatever we conclude that we ought to do. In such cases, in answering these moral questions, we are deciding what to do.

Nowell-Smith might say that, since we may also decide *not* to do what we ought to do, it is still the 'shall' question that is fundamental. The 'ought' question, Nowell-Smith assumes, takes the fundamental concept to be *obligation*. Only the 'shall' question takes that concept to be *decision*.

By tying 'ought' to obligation, Nowell-Smith here restricts the *normative* to the *moral*. But most of our practical decisions do not involve moral thinking; and, in making these decisions, we often ask what we have reason to do, and what we should, ought, or must do. It is true that, in answering

[5] Patrick Nowell-Smith, *Ethics* (Harmondsworth: Penguin, 1954), 319–20.

[6] *Ethics*, 267.

these questions, we may not be deciding what to do. Suppose I come to believe that, since it is the only way to save my life, I should jump from the burning building. After reaching that belief, I must still decide to jump. If I am irrational, I may not make the final moves from 'should' to 'shall', and from there to an act. But that does not show that, in practical reasoning, our fundamental question is whether to make the move from 'I should' to 'I shall'. On the contrary, if we were fully practically rational, we would always make that move, and without any further thought. We would always decide to do, and then try to do, whatever we had concluded that we should do, or that we had most reason to do. Since this move from 'should' to 'shall' would be automatic, we would *never* need to ask 'What shall I do?'

Consider next some remarks by Williams. Like Nowell-Smith, Williams regards practical reasoning as 'radically first-personal', since its central question is 'What shall *I* do?' But Williams assumes that, in deciding what to do, we often ask what he calls the *deliberative question*. We ask what we should do, all things considered, or what we have most reason to do.

Williams's conception of a reason is, however, reductive. He assumes that, when we have a reason to act in some way, that is a fact about this act's relation either to our present desires, or to the motivations that, after informed deliberation, we would have. Williams regards the concept of a reason as, in part, normative. But his conception of normativity is, I believe, too weak. Thus he writes that, when we claim that someone has a reason for acting, we do not mean only that this person is presently disposed to act in some way; but we might mean that he would be so disposed if he knew a certain fact. We would then be adding to, or correcting, this person's factual beliefs, 'and that is already enough', he writes, 'for this notion to be normative'.

When Williams argues that there are no external reasons, he imagines someone who maltreats his wife, and whose attitudes and acts would not be altered by informed and rational deliberation. If we are Externalists, we might claim that, despite this man's motivational state, his wife's unhappiness gives him reasons to treat her better. In rejecting this claim, Williams asks:

> what is the difference supposed to be between saying that the agent has a reason to act more considerately, and saying one of the many other things we can say to people whose behaviour does not accord with what we think it should be? As, for instance, that it would be better if they acted otherwise?[7]

[7] Bernard Williams, 'Internal Reasons and the Obscurity of Blame', henceforth *IROB*, in *Making Sense of Humanity* (Cambridge: Cambridge University Press, 1995), 39–40.

We might answer: 'The difference is that, if we merely said that it would be better if this man acted more considerately, we would not be claiming that, as we believe and you deny, he has reasons to do so.'

Williams's ground for rejecting this claim is that he finds it 'quite obscure' what it could mean. As he writes elsewhere, Externalists do not 'offer any *content* for external reasons statements'. [8] Williams may here be assuming Analytical Internalism. [9] On this view, in claiming that

(1) this man has reasons to treat his wife better,

we would mean that

(2) if he deliberated rationally on the facts, he would be motivated to treat her better.

If (1) meant (2), and we knew that (2) was false, it would indeed be obscure what, in claiming (1), we could mean. *Non*-Analytical Internalists would not find our claim so obscure. Such Internalists believe that, though (1) is true only if (2) is true, these claims have different meanings. These Internalists would understand—though they would reject—the view that, despite this man's motivational state, he has reasons to treat his wife better.

Discussing another, similar example, Williams asks:

> What is gained, except perhaps rhetorically, by claiming that A has a reason to do a certain thing, when all one has left to say is that this is what . . . a decent person . . . would do?[10]

This question seems to assume that, if our claim about A does not have the sense described by Analytical Internalists, there is nothing distinctive left for it to mean. We couldn't mean that, despite A's motivational state, A has a reason to do this thing. If we could mean that, there would be a simple answer to Williams's question. We might be saying something that was both distinctive and true.

Williams continues:

> it would make a difference to ethics if certain kinds of *internal* reason were very generally to hand . . . But what difference would external reasons make? . . . Should we suppose that, if genuine external reasons were to be had, morality might get some leverage on a squeamish Jim

[8] Bernard Williams, *World, Mind, and Ethics*, henceforth *WME*, ed. J. E. J. Altham and Ross Harrison (Cambridge: Cambridge University Press, 1995), 191, my italics.

[9] As he seems to do elsewhere. Thus he writes: 'I think the sense of a statement of the form "A has a reason to *phi*" is given by the internalist model' (*IROB* 40). See also 'Internal and External Reasons', henceforth IER, in *Moral Luck* (Cambridge: Cambridge University Press, 1973), 109–10, and *IROB* 36; both discussed below. On the other hand, see *WME* 188.

[10] *WME* 215.

> or priggish George, or even on the fanatical Nazi? . . . I cannot see what leverage it would secure: what would these external reasons do to these people, or for our relations to them?

These remarks assume that, for external reasons to make a difference to ethics, such reasons would have to get *leverage* on people, by motivating them to act differently. This conception of ethics is, I believe, too utilitarian. When we believe that other people have reasons for caring, or for acting, we do not have these beliefs as a way of affecting those people. Our aim is, not influence, but truth. Similar remarks apply to morality. Someone might say:

> What difference would it make if it were true that the Nazis acted wrongly? What leverage would that moral fact have secured? What would the wrongness of their acts have done to them?

Even if moral truths cannot affect people, they can still be truths. People can be acting wrongly, though the wrongness of their acts does not do anything to them.

After asking what external reasons would do to such people, Williams writes:

> Unless we are given an answer to that question, I, for one, find it hard to resist Nietzsche's plausible interpretation, that the desire of philosophy to find a way in which morality can be guaranteed to get beyond merely *designating* the vile and recalcitrant, to transfixing them or getting them inside, is only a fantasy of *ressentiment*, a magical project to make a wish and its words into a coercive power.[11]

Williams has a real target here. Many philosophers have hoped to find moral arguments, or truths, that could not fail to motivate us. Williams, realistically, rejects that hope.

Note however that, in making these remarks, Williams assumes that claims about reasons could achieve only two things. If such claims cannot get inside people, by inducing them to act differently, they can only designate these people. On the first alternative, these claims would have motivating force. On the second, they would be merely classificatory, since their meaning would be only that, if these people were not so vile, or were in some other way different, they *would* act differently. As before, however, there is a third possibility. Even when such claims do not have motivating force, they could be more than merely classificatory. They could have *normative* force. Perhaps these people *should* act differently.

We should remember next that Externalists need not be Moral Rationalists. Some Externalists would agree with Williams that those who act wrongly

[11] *WME* 216.

may have no reason to act differently. These people are Externalists in their beliefs about prudential reasons. Return to Williams's imagined person who needs some medicine to protect his health, and whose failure to care about his future would survive any amount of informed and procedurally rational deliberation. Such a person, Williams writes, would have no reason to take this medicine.[12] He might ask:

> What would be gained by claiming that this person has such a reason? What would that add to the claim that, if he were prudent, he would take this medicine?

This claim would add what Williams denies. This person, these Externalists believe, ought rationally to take this medicine. He has reasons to care about his future; and, since these are reasons for caring, this person's failure to care does not undermine these reasons. Such claims, I believe, make sense, and might be true.

4

Many other writers conflate normativity and motivating force. For example, Korsgaard writes that, if a certain argument 'cannot motivate the reader to become a utilitarian then how can it show that utilitarianism is normative?' McNaughton writes that, when externalists deny that moral beliefs necessarily motivate, they 'deny the authority of moral demands'.[13] Scheffler writes that, even if wrong-doing were always irrational, that would not give morality 'as much authority as some might wish', since it would not 'guarantee . . . morality's hold on us'.[14] And Railton writes: 'our hypothetical approvals under full information have a kind of motivational force or authority for us'.

Railton also writes:

> there is no need to explain the normative force of our moral judgments on those who have no tendency to accept them and who recognize no significant community with us. For that is not a force that we observe in moral practice.[15]

[12] IER 105–6.

[13] David McNaughton, *Moral Vision* (Oxford: Blackwell, 1988), 48.

[14] Samuel Scheffler, *Human Morality* (Oxford: Oxford University Press, 1992), 95.

[15] Peter Railton, 'What the Non-Cognitivist Helps us to See the Naturalist Must Help us to Explain', in John Haldane and Crispin Wright (eds.), *Reality, Representation and Projection* (Oxford: Oxford University Press, 1993), 286. This describes what non-cognitivists might claim, but Railton, though rejecting non-cognitivism, seems to endorse this claim.

What we can observe, in seeing how people act, is not normative but motivating force. Similarly, Railton writes that, to show how the idea of happiness can have a 'normative role', or have 'recommending force', we can appeal to the fact that it is 'impossible for a person to have the peculiar experience that is happiness and not be drawn to it'. But we cannot, he adds, 'claim it as definitional that happiness matters, i.e. that that which left us indifferent would not, by definition, be happiness'. The concepts *normative*, *recommend*, and *matter* are here conflated with, or reduced to, psychological appeal.

Consider next some remarks of Mackie's. Since Mackie is an *error theorist*, who believes that ordinary moral thinking is committed to peculiar non-natural properties, we might expect that he at least would give a non-reductive account of the normativity that he rejects. Mackie writes that, according to some cognitivists, a moral judgment is 'intrinsically and objectively prescriptive', since it 'demands' some action, and implies that other actions are 'not to be done'. These phrases look normative. But Mackie later writes that, in response to Humean arguments for non-cognitivism, cognitivists might

> simply deny the minor premiss: that the state of mind which is the making of moral judgments and distinctions has, *by itself*, an influence on actions. [They] could say that just seeing that this is right and that is wrong will not tend to make someone do this or refrain from that: he must also *want* to do whatever is right.

If cognitivists made such claims, Mackie continues, they would 'deny the intrinsic action-guidingness of moral judgments', and they would 'save the objectivity of moral distinctions . . . only by giving up their prescriptivity'. Mackie here assumes that, in claiming moral judgments to be action-guiding and prescriptive, we mean that such judgments can, by themselves, *influence* us, or tend to *make* us act in certain ways. So, even when describing the view that he rejects—or the 'objectively prescriptive values' that he calls 'too queer' to be credible—Mackie takes normativity to be a kind of motivating force.[16]

Others make similar remarks. An objective value, Korsgaard writes, would have to be 'able both to tell you what to do and make you do it. And nothing is like that.' And Wittgenstein wrote:

> the absolute good . . . would be one which everybody, independent of his tastes and inclination, would *necessarily* bring about or feel guilty for

[16] J. L. Mackie *Hume's Moral Theory* (London: Routledge & Kegan Paul, 1980), 54–5. For another discussion of this view of normativity, see Stephen Darwall, 'Internalism and Agency', in *Philosophical Perspectives*, 6, *Ethics* (Oxford: Blackwell, 1992).

> not bringing about. And I want to say that such a state of affairs is a chimera. No state of affairs has, in itself, what I would like to call the *coercive power* of an absolute judge.[17]

Normativity, I believe, cannot be, or be created by, any kind of power, not even that of some absolute omnipotent judge.

The most surprising maker of such claims is the young Thomas Nagel. In his introduction to *The Possibility of Altruism*, Nagel wrote:

> Philosophers . . . commonly seek a justification for being moral: a consideration which can persuade everyone or nearly everyone to adhere to certain principles, by connecting those principles with a motivational influence to which everyone is susceptible . . .[18]

This remark conflates justification, persuasion, and motivation.

This conflation was deliberate. When Nagel wrote this book, he regarded ethics 'as a branch of psychology', and was 'in search of principles which belong both to ethics and to motivation theory'. This approach, he admitted, 'may appear to involve an illegitimate conflation of explanatory and normative enquiries'. But the alternative, he thought, was 'to abandon the objectivity of ethics'. If we are to 'rescue' ethics, we must show that ethical requirements are based upon, or provided by, motivational requirements. Normativity, Nagel assumed, must be a kind of motivating force.

This assumption, I have claimed, does not rescue but abandons ethics. It is one example of what Nagel later called 'the perennially tempting mistake of seeking to explain an entire domain of thought in terms of something outside that domain, which is simply less fundamental than what is inside'. Since Nagel is one of those who have done most to challenge this mistake, it is significant that, in his first book, he himself made this mistake. That shows how tempting, and damaging, it can be.

Nagel began by discussing first-person practical judgments, such as

(1) the judgment that we have a reason to act in some way, or that we should do so.

Such a judgment involves

(2) the belief that we have a reason to act in this way, or that we should do so.

[17] (my italics) 'Wittgenstein's Lecture on Ethics', *Philosophical Review*, 74/1 (Jan. 1965), 7.

[18] Thomas Nagel, *The Possibility of Altruism*, henceforth *PA* (Oxford: Oxford University Press, 1970), 3.

But, Nagel argued, (1) involves more than (2). Though such judgments involve beliefs, they include another element, which he called their *motivational content.*

Nagel described this content in several ways. In his most common phrase, such judgments include

(3) 'the *acceptance of a justification* for doing or wanting something'.

(3), straightforwardly understood, gives to 'motivational content' what we can call its *justificatory* sense.

We may ask how (3) differs from (2). When we believe that we have a reason to do something, or that we should do it, are we not thereby believing that we have some justification for doing this thing? Nagel would have replied that, in *accepting* a justification for this act, we are not merely *believing* that it would be justified. Since such judgments are practical, they have 'practical consequences'. When we accept such a judgment, that should affect our motivation.

On the simplest form of this reply,

(1) our judging that we have a reason to do something, or that we should do it,

includes

(4) our being motivated to do this thing.

(4) gives to 'motivational content' what we can call its *motivational* sense.

If Nagel had claimed that (1) includes (4), he would have been defending Belief Internalism. On this view, we cannot believe that we have a reason to do something without being motivated to do it.

Nagel hoped to defend this view. Thus he claimed that, if ethics is to contain 'practical requirements', motivational theory must contain results that are '*inescapable*': there must be 'motivational influences which one *cannot reject* once one becomes aware of them'. And he wrote:

> Internalism is the view that the presence of a motivation for acting morally is guaranteed by the truth of the ethical propositions themselves. On this view . . . when in a particular case someone is (or perhaps merely believes that he is) morally required to do something, it follows that he has a motivation for doing it . . . . The present discussion attempts to construct the basis of an internalist position.

This attempt failed, since the conclusions Nagel reached were not, even in his own terms, internalist. As he wrote, 'a practical judgment can sometimes fail to prompt action or desire'. Such judgments can fail to motivate, he added, even 'without any explanation'.[19]

Though Nagel rejected Belief Internalism, he defended a related view. In his words:

> The belief that a reason provides me with sufficient justification for a present course of action does not necessarily imply a desire or a willingness to undertake that action; it is not a sufficient condition of the act or desire. But it is sufficient, in the absence of contrary influences, to *explain* the appropriate action, or the desire or willingness to perform it. That is the motivational content of a judgment about what one presently has reason to do.

On this more cautious view, practical judgments do not *necessarily* motivate us. What such judgments guarantee is only what Nagel calls 'the *possibility* of appropriate motivation'.

This view may seem trivially true. Who would deny that, when we believe that we have a reason to do something, or that we should do it, we *might* be motivated to do it?

Nagel's view was not, however, trivial. On the Humean theory of motivation, which is now widely accepted, no beliefs can motivate us all by themselves. For some belief to motivate us, it must be combined with some *independent* desire—some desire that is not itself produced by this belief. Suppose that, though we believe that we should do something, we have no such relevant independent desire. On the Humean theory, it is then causally impossible for us to do this thing. Reason by itself is impotent, since beliefs about reasons have no power to motivate us. Nagel argued, I believe soundly, that we should reject this view. And, as he claimed, this rejection has great significance.

Return now to Nagel's view about the content of practical judgments. According to Nagel, in

(1) judging that we have a reason to do something,

we are not having a mere belief, since this judgment has motivational content. (1) includes

(5) being in a state which, though it may not motivate us to do this thing, would be sufficient to explain such motivation.

This claim gives to 'motivational content' what we can call its *explanatory* sense.

[19] *PA* 65.

Nagel's view seems, in one way, inconsistent. If (1) includes (5), that does not show that (1) is not a mere belief. (1) could be the same as

> (2) our believing that we have this reason.

It could be true that

> (6) in having such beliefs, we are in a state which, though it may not motivate us, would be sufficient to explain such motivation.

Humeans would reject (6), since they assume that no belief could by itself motivate us. But, as I have said, Nagel rightly rejected this view.

It might be said that, if such beliefs could by themselves motivate us, they cannot be *mere* beliefs. They would be very special beliefs: ones with motivating force. But this reply misses the point. If practical judgments are beliefs, that makes them mere beliefs in the sense of 'mere' that is relevant here. According to anti-Humeans, beliefs that are in this sense 'mere' could by themselves motivate us.

Remember next Nagel's claim that (1) includes

> (3) our accepting a justification for doing this thing.

This claim also fails to show that practical judgments are not mere beliefs. (3) could be our *believing* that we have this justification.

Nagel's view, I conclude, should have taken a simpler form. He need not have distinguished (1), (2), and (3), since these are all descriptions of the same kind of normative belief. Nor should Nagel have claimed that the *content* of these beliefs is, in part, *motivational*. These beliefs are, in content, *normative*. Nagel's claim should have been only that these beliefs might, by themselves, motivate us.

If Nagel's view had taken this simpler form, it would have been closer to the view that, in his later writings, he so forcefully defends. Practical judgments, he could have claimed, are about irreducibly normative truths.

Nagel did not make that claim, in his first book, because he conflated normativity with motivating force. Though that conflation was in part deliberate, it led him, I believe, astray.

He did not distinguish, for example, between his different senses of the phrase 'motivational content'. Thus, in discussing first-person practical judgments, Nagel wrote:

> the acceptance of such a judgment is by itself sufficient to explain action or desire in accordance with it . . . I have referred to this motivational content as the *acceptance of a justification* for doing or wanting something.[20]

[20] *PA* 109.

This definition conflates what I have called the justificatory and explanatory senses. This conflation is surprising. When we claim that someone's state would be sufficient to explain his doing something, we do not seem to be claiming that this person accepts a justification for doing this thing.

Nagel's failure to draw this distinction had, I believe, some bad effects. For example, he wrote:

> Moral scepticism is a refusal to be persuaded by moral arguments or reasons. The object of persuasion in this case is action or desire, and that differentiates it from epistemological scepticism. The latter is a refusal to be persuaded by certain arguments or evidence, where the object of persuasion is *belief.* To defeat moral scepticism, therefore, it is not sufficient to produce the belief that certain moral statements are true, for this may leave the sceptic unpersuaded to act differently. He may refuse to accept the fact that he *should* do something as a justification for doing or wanting to do it; i.e. he may attempt to acknowledge the truth of the statement without accepting its motivational content . . . This explains why a successful attack must be directed against volitional rather than cognitive scepticism.[21]

Consider first what Nagel meant, if and insofar as he was using 'motivational content' in its justificatory sense. Nagel would be claiming that, even if we convinced the sceptic that he should do something, he might not accept that this fact was a justification for doing it. Though someone might hold such a view, it would be a form of cognitive rather than 'volitional' scepticism. Nor would this view be worth considering, since it is obviously incoherent. If we *should* do something, that is a justification for doing it. Anyone who denied that fact would not know what 'should' means.

Consider next what Nagel meant, if and insofar as he was using 'motivational content' in its explanatory sense. Nagel's point might have been that there are people who, though believing certain moral statements to be true, do not accept that these beliefs could, by themselves, motivate them. There are indeed such people. But their view is not a form of moral scepticism. These people combine moral cognitivism with the Humean theory of motivation. Such a view was held, for example, by David Ross. And, of those who hold such views, some might never doubt, or fail to do, their duty. These people's moral beliefs would always motivate them. Their mistake would be only to regard such motivation as requiring an independent desire to act on these beliefs.

[21] *PA* 143–4.

Nagel's point may instead have been that there are some cognitivists who, though believing that they should do something, are *not* motivated by that belief. As before, though there might be such people, they do not seem to be moral sceptics. What are they doubting? Such people might accept both of Nagel's claims about the motivational content of moral beliefs. When they believe that they should do something, they might accept that this fact was a justification for doing it. And, unlike Ross, they might agree that, in believing that they should act, they are in a state that *could* by itself motivate them. 'Could' does not mean 'does'. Moral beliefs, as Nagel wrote, 'can sometimes fail to prompt action or desire'. These people might say, 'My belief, regrettably, is one such case.'

Volitional scepticism, it may be objected, need not involve *doubting* anything. Nagel's point may have simply been that moral beliefs sometimes fail to motivate. In that case, however, Nagel's wording was misleading. When these people's moral beliefs fail to motivate them, they are not, as Nagel claims, refusing to be persuaded that certain acts would be justified.

It may next be said that, in making these remarks, I am missing Nagel's point. Such people *cannot* have been persuaded that these acts would be justified. If they really *believed* that they should do something, they could not fail to be motivated to do this thing. As we have seen, however, Nagel rejects this view. Moreover, if this view were true, that would undermine Nagel's conclusion. On this view, by defeating cognitive scepticism, we would defeat volitional scepticism. To motivate people to act morally, it *would* be enough to persuade them that there are certain moral truths.

It seems then that, in this passage, Nagel might have been making any of these claims:

(7) There could be people who did not understand that, if they ought to do something, that justifies their doing it.

(8) There are people who, though believing that there are moral truths, accept the Humean theory of motivation.

(9) Moral beliefs sometimes fail to motivate.

These claims are all true. But they are not, as Nagel seemed to think, claims that his arguments support. The most important claim—(9)—is something that his arguments assume.

Nagel's arguments do support several significant conclusions. One example is his rejection of the Humean theory. But I believe that, because he conflated normativity with motivation, and justification with persuasion, Nagel sometimes mis-stated, or misunderstood, his conclusions. Thus, in the passage we have been considering, Nagel seems to be claiming something other than (7) to (9).

5

Consider next Nagel's account of practical reasoning. Nagel wrote:

> a judgment that a certain action or desire is justified has motivational content. To accept a reason for doing something is to accept a reason for *doing* it, not merely for *believing* that one should do it.[22]

As Nagel's second sentence claims, in believing that we have some reason for acting, we are believing that we have a reason for *acting*. But this is not some further 'motivational content' that, when combined with this belief, makes it a practical judgment. This is the content of this belief. *What* we believe is that we have this reason for acting.

Nagel seems to be intending, here, to reject a different view. His remarks suggest that, according to some people, when we believe that we have some reason for acting, we are not believing that we have this reason for acting. Our belief is only that we have a reason for *believing* that we have this reason for acting. There is, however, no such view. It is impossible to think that, in having some belief, we are not having this belief. Nor is it possible to think that, in having some belief, we are believing only that we have a reason for having it. If we have some belief, we have this belief.

Nagel was intending, I assume, to reject some other view. There are two possibilities. On the same page, Nagel wrote:

> the crucial point is that a practical reason is a reason to do or want something, as a theoretical reason is a reason to conclude or believe something . . . To hold, as Hume did, that the only proper rational criticism of action is a criticism of the beliefs associated with it is to hold that practical reason does not exist. If we acknowledge the existence of reasons for action we must hold not merely that they justify us in believing certain special propositions about action, but rather that they justify the action itself . . .

On Hume's view, all reasoning is theoretical. Since reasoning is concerned with truth, there are no *practical* reasons: reasons for caring or for acting. Unlike beliefs, desires and acts cannot be either true or false; so they cannot be supported by, or contrary to, reason.

[22] *PA* 63–4.

Nagel rightly rejected this view, for which Hume gave no argument. And when Nagel insisted that reasons for acting are reasons for *acting*, Hume may have been his only target. But that would not explain his claim that, in accepting that we have some reason for acting, we are not accepting merely that we have a reason for *believing* that we have this reason for acting. Since Hume ignored reasons for acting, he expressed no view about what is involved in accepting that we have such reasons.

Nagel's target seems here to be, not Hume's view that all reasoning is theoretical, but an overly theoretical view about practical reasoning. His claim may be that, when we engage in practical reasoning, it is not enough to reach conclusions about what we should do. Such reasoning should also lead to *decisions*, and to *acts*.

Many other writers make a similar but stronger claim. According to them, practical reasoning is not concerned with beliefs, or truths. That is how Korsgaard, for example, criticizes rational intuitionism, or what I am calling practical realism. Intuitionists, Korsgaard writes, 'do not believe in practical reasoning, properly speaking. They believe there is a branch of theoretical reason that is specifically concerned with morals.' According to them, when we ask 'practical normative questions . . . there is something . . . that we are trying to find out . . . our relation to reasons is one of seeing that they are there or knowing truths about them'. This view, Korsgaard claims, is deeply mistaken. As Kant saw, practical reasoning is wholly distinct from theoretical reasoning. There are no such independent normative truths. We *create* reasons, and morality consists, not in *truths*, but in *imperatives*.

We shall return to Korsgaard's view. Surprisingly, in his first book, Nagel sometimes made similar claims. Thus he wrote:

> I suspect . . . that it is really an unrecognized assumption of internalism that underlies Moore's 'refutation' of naturalism. The evaluative factor which is always left out by any naturalistic description of the object of ethical assessment is in fact the relevant inclination or attitude. But Moore did not realize this, and consequently [held a view] in which a peculiar non-natural quality served to flesh out the content of ethical claims.[23]

These remarks suggest that, in judging some act to be good or right, we are not claiming that this act has some normative property, but expressing an inclination or attitude.

This suggestion must have been a slip, since Nagel was not an emotivist, or non-cognitivist. He believed that there are moral truths. But, in his claims about 'motivational content', he came close to abandoning that belief.

[23] *PA* 8.

Thus, after mentioning Moore's view that words like 'good' and 'right' refer to irreducibly normative properties, Nagel wrote that, on this view,

> it can only be regarded as a mysterious fact that people care whether what they do is right or wrong . . . Such views are, it seems to me, unacceptable on their surface, for they permit someone who has acknowledged that he should do something, and seen *why* it is the case that he should do it, to ask whether he has any reason for doing it.

Several other writers make such claims. For example, when discussing Moore's alleged normative truths, Nowell-Smith wrote:

> No doubt it is all very interesting. If I happen to have a thirst for knowledge, I shall read on . . . Learning about 'values' or 'duties' might well be as exciting as learning about spiral nebulae or waterspouts. But what if I am not interested? Why should I *do* anything about these newly-revealed objects? Some things, I have now learnt, are right and others wrong; but why should I do what is right, and eschew what is wrong?[24]

When words are 'used in the ordinary way', Nowell-Smith goes on to say, such questions are absurd. But they 'would not be absurd if moral words were used in the way that intuitionists suppose'. In 'ordinary life there is no gap between "this is the right thing for me to do" and "I ought to do this" '. But if 'X is right' were taken to mean that X had the 'non-natural property' of being right, we *could* deny that we ought to do what is right.

There is an obvious reply. If these acts had the non-natural property of being the right thing to do, they would have the non-natural property of being what we ought to do. Nowell-Smith's suggested questions would still be absurd. Nowell-Smith's remarks are intended to show that the intuitionists' alleged moral truths could not be normative. But his argument amounts to the claim that, even if we knew that some act was right, or was what we ought to do, we could still deny that this act was right, or was what we ought to do. That is not so.

Nowell-Smith could still have said, 'But what if we are not interested? What if we don't care about what we ought to do?' That reply, however, is no objection to the intuitionists' view. It confuses normativity with motivating force. Even if we don't care, we should.

Consider next a remark of Hare's about the 'alleged moral properties which', on the intuitionist view, 'actions are supposed to have'.[25] If 'it

[24] Nowell-Smith, *Ethics*, 61.

[25] R. M. Hare, *Moral Thinking* (Oxford: Oxford University Press, 1981), 217.

just is the case that . . . the acts open to a person have the moral property of wrongness, one of their many descriptive properties, why should he be troubled by that?' Hare's remark assumes that there could not be normative truths, since any truth would be merely 'descriptive', and could not provide reasons. On Hare's view, even if it were true that this person had reason to be troubled, he would have no reason to be troubled. As before, that is not so.

Williams similarly writes:

> this critic deeply wants this *ought* to stick to the agent . . . This is the right place for the standard emotivist or prescriptivist argument, that even where 'It ought to be that p' has the particular form, 'It ought to be that A does X', if it just tells one a fact about the universe, one needs some further explanation of why *A* should take any notice of that particular fact.[26]

Suppose that the normative facts were, not only that A ought do X, but also that A ought to take notice of that fact. And suppose we knew why these facts obtained. Perhaps A ought to do X because he promised to do so, and A ought to take notice of this fact because we all have reason to support the practices that make cooperation possible. If these were the normative facts, as this emotivist argument allows us to suppose, we wouldn't need a further explanation of why A ought to take notice of them. That would be one of the facts that we had already explained.

Return now to Nagel's rejection of Moore's view. If some acts had the 'non-natural' property of being right, it would be mysterious, Nagel wrote, that people cared about that fact. And, like Hare and Williams, Nagel suggested that, even if we knew that we *should* do something, we could deny that we had any reason for doing it.

There is one way to make sense of this second claim. Nagel might have been appealing to Internalism about reasons. His point might have been that, if some act had this alleged non-natural property of being right, that would not be a fact about this act's relation to our own motivation. According to Internalists, such a fact could not provide a reason for acting.

This seems unlikely, though, to have been what Nagel meant. He rejected this form of Internalism. Nor was he contrasting moral and non-moral uses of the word 'should'. He seems to have meant that, if some act had

[26] Bernard Williams, 'Ought and Obligation', *Moral Luck*, 122. (I have expanded some abbreviations.)

the non-natural property of being what we should do, we could still ask whether we should do it.

As before, that suggestion makes no sense. Nor was it really Nagel's view. But, because he conflated normativity and motivation, Nagel slipped into endorsing this emotivist argument.

Consider next Nagel's claim that practical reasoning should lead us to decisions and to acts. We can here distinguish three views:

(A) Practical reasoning does not lead to beliefs. It leads to practical judgments, such as 'I should do that'; and such judgments are not beliefs. The words 'I should' express some decision, or attitude.

(B) Practical reasoning leads to beliefs, such as 'I should do that'. But, to be practically rational, it is not enough to reach such conclusions. When we believe that we should do something, we should decide to do it, and we should act on that decision.

(C) To be fully practical rational, it is enough to reach true or justified beliefs about what we should do.

Non-cognitivists accept (A). So, in a way, do certain Kantians—who may include Kant. Despite his remarks about Moore, Nagel's view was, and remains, (B). When he wrote his first book, Nagel seemed to think that some philosophers accept (C). That may be why he insisted that, in judging that we have some reason for acting, we are judging that we have a reason for acting, not merely for believing that we should act. But I can think of no one who accepts (C).

Consider one more passage. Practical judgments, Nagel wrote, do not consist

> merely in the observation that certain features of one's situation fall into categories called 'reasons.' . . . [They] are not merely classificatory: they are judgments about what to do; they have practical consequences. If they were merely classificatory then a conclusion about what one *should* do would by itself have no bearing on a conclusion about what *to do*. The latter would have to be derived from the former, if at all, only with the aid of a further principle, about the reasonableness of doing what one should do.[27]

Practical judgments would be merely classificatory, Nagel assumed, if they did not have motivational content. Suppose first that he was using 'motivational' in his justificatory sense. His point would then be that, if the

[27] *PA* 109.

claim that we *should* do something did not imply a justification for doing it, it would not have normative force. Though true, that is too obvious to be all that, in this passage, Nagel meant.

Suppose next that Nagel had in mind the explanatory sense. His point, in this passage, would then be this. If our judgment that we should do something could not by itself motivate us, such a judgment would not be relevant to a decision about what to do. In other words, if the Humean theory of motivation were correct, practical judgments would not provide reasons for acting. If this were Nagel's point, he would again be conflating normativity and motivating force.

This reading seems to fit the start of this passage. Practical judgments would be 'merely classificatory', Nagel says, if they were merely beliefs about what we should do. In having such beliefs, we would merely be observing that certain features of our situation fell into categories called 'reasons'. That wording suggests that there could not be normative truths. Nagel's claim seems to be that, if our belief that we had some reason could not by itself motivate us, this belief's content would be only that certain natural features of our situation can be correctly called 'reasons'. On such a view, now very widely held, there are natural facts about the world, including facts about our motivation. But there are no other, irreducibly normative facts, or truths.

On Nagel's later view, when we believe that certain natural facts give us reasons for caring or for acting, we are not believing that these facts can be called 'reasons'. These beliefs are *normative*. We are believing that we *should* care, or *should* act. And such beliefs might be, in a strong sense, true. As Nagel wrote:

> If I have a severe headache, the headache seems to me not merely unpleasant, but a bad thing. Not only do I dislike it, but I think I have a reason to try to get rid of it. It is barely conceivable that this might be an illusion, but if the idea of a bad thing makes sense at all, it need not be an illusion . . .[28]

At the start of his first book, Nagel claimed that, to rescue ethics, we must regard it 'as a branch of psychology'. Rational requirements must be grounded in motivational claims. But, as Nagel later claimed, this widely held belief is a deep mistake.[29] Unless we distinguish between reasons and motivating states, we cannot claim that, as the young Nagel wrote, 'to

[28] Thomas Nagel, *The View from Nowhere* (Oxford: Oxford University Press, 1986), 145.

[29] Thomas Nagel, *The Last Word* (Oxford: Oxford University Press, 1997), ch. 6.

accept a reason for doing something is to accept a *reason* for doing it'. The young Nagel also wrote:

> in so far as rational requirements, practical or theoretical, represent conditions on belief and action, such necessity as may attach to them is not logical but natural or psychological.[30]

Though such necessity is not logical, it is not natural or psychological either. This necessity is normative. As Nagel later claimed, it is reason that has the last word.

6

We can end by considering Korsgaard's view. Of reductive accounts of normativity, Korsgaard gives the fullest; and she also takes seriously the kind of non-reductive practical realism that, following Nagel, I am trying to defend.

To introduce Korsgaard's view, it will help to reconsider David Falk's. According to Falk reasons for acting are not normative: they are facts belief in which might cause us to act. Normativity, Falk assumes, belongs most clearly to imperatives. A normative utterance, he writes, 'is one like "Keep off the grass" '. Since such utterances are not statements, they could not be either true or false.

Harder to classify, on Falk's view, are claims that use the word 'ought'. Falk suggests that, while an order like 'Keep off!' is purely normative, a claim like 'You ought to keep off' is partly normative and partly descriptive. Though this claim tells you to keep off, it also implies that you have a reason for doing that. As Falk writes of another such claim—'You ought to go now'—this claim 'needs support from "your bus is leaving" . . . or any other natural feature of the situation which may count as a reason'. Since such statements are backed by reasons, they seek to persuade by rational means. They do not merely *goad*: they *guide*.

Such statements, Falk remarks, are in one way puzzling. The claim 'You ought . . .' does not itself *give* you a reason. Since that is so, Falk writes, such a claim

> seems a logically redundant part of this machinery. One persuades by rational methods when one gives reasons, reports those features of the situation likely to count in favour of a doing . . . What else but another reason could add persuasive force to the reasons already given? But

[30] *PA* 22.

> 'you ought to' is said after everything to count as a reason has been enumerated. It seems persuasive, and like adducing a reason, and yet is not. It seems both to belong to persuasion by rational methods, and not to be part of it.[31]

Falk here plausibly assumes that, if we ought rationally to act in some way, this fact is not a reason for doing so. It is the fact *that* some fact gives us such a reason, and one that is not outweighed by other reasons. In the same way, something's being good is not a reason for choosing it; it is the fact that this thing has features that provide such reasons. But these points do not make *ought* and *good* logically redundant parts of practical reasoning. Falk comes close to seeing this when, in this passage, he forgets his definition of the concept of a reason. We give reasons for acting in some way, Falk writes, when we report 'those features of a situation likely to count in favour' of this act. In claiming that these features *count in favour*, we do not mean that, if the agent knew about these features, that might cause him to act. We mean that these features *support* his acting, and thereby support the conclusion that he *should* act normatively.

Falk continues:

> 'persuading by giving someone a reason' is an ambiguous notion. It may mean 'by stating a fact calculated to act as a reason'; and also 'by stating such a fact and stating that, if considered, it will act as such a reason' . . . Prescriptive speech of the guiding type reaches a new level . . . when it turns from purely stating persuasive facts to announcing the claim *that* they constitute reasons, 'good reasons', 'valid' reasons, etc. . . .

Falk's use of the phrase 'good reasons' may seem to be normative. But that is not so. Facts are good reasons, in Falk's sense, if belief in these facts would have persuasive or motivating force. When we use 'ought' to imply the presence of a reason—which is what Falk calls the *motivation sense* of 'ought'—we are making a psychological prediction, which can be proved either true or false. As Falk writes:

> I have defended the view that reasons are forces . . . If reasons are choice-guiding because they are forces, then a circumstance that holds a reason for one need not hold a reason . . . for everyone. What . . . can qualify as a choice-determining consideration is in essence an empirical matter.

As he writes elsewhere, when we say 'You ought to go', we want our claim to be 'put to the test . . . we desire the hearer to have the benefit of *experiencing* what we claim'.

[31] W. D. Falk, *Ought, Reasons, and Morality* (Ithaca, NY: Cornell University Press, 1986).

We can now see why, on Falk's view, the concepts *ought* and *good* are logically redundant. They have no distinctive sense, or conceptual role. Falk ignores the possibility that, in making a claim like 'You ought to go', we might be stating a normative truth. When such claims are true, he assumes, their truth consists in a motivational prediction. Insofar as such claims are normative, they are like the imperative 'Go!' Though such claims can be both normative and true, their normativity is not part of what makes them true. 'You ought to go' means, roughly, 'If you knew the truth, you would want to go, so: Go!'

Falk would have rejected this assessment of his view. While he believed that, in some of its uses, the word 'ought' merely expresses an imperative, that is not true, he claims, of the motivational or psychologically predictive use of 'ought'. Rational people, Falk writes, 'are interested in other people's emotive noises' only insofar as these present 'an objectively valid recommendation for them'. And this motivational ought has, he claims, such objective recommending force.

Falk then considers an objection like mine. Critics may say, he writes, 'that "I ought" is different from "I would want if I first stopped to think". The one has a normative and coercive connotation which the other has not.' Falk replies that, when we use 'ought' in this motivational sense, our claim may not only be about what we *would* want. It may be about what we would *have* to want. Such a use of 'ought', Falk then writes, meets Kant's criterion of normativity. According to Kant, when we say that we 'ought' to do something, we mean that 'we have, contrary to our inclinations, not only a rational but a *rationally necessary* impulse or "will" ' to do this thing.[32]

This reference to rational necessity again looks promisingly normative. But that promise is not fulfilled. On Falk's account, an impulse is *rational* if it is one that 'a person would have if he both acquainted himself with the facts and tested his reactions to them'. Such an impulse is *necessary* if it would be unalterable 'by any repetition of these mental operations'. There is here no practical reasoning. To find out what is rationally necessary in Falk's sense, we merely review the relevant facts and *test our reactions*. Falk continues:

> And this is meant by a 'dictate of reason': an impulse or will to action evoked by 'reason' and . . . one which derives a special forcibleness from [the fact that] no further testing by 'reason' would change or dislodge it . . . A conclusive reason would be one [that is] unavoidably stronger than all opposing motives.

[32] Ibid.

> [People are] under obligations when . . . they have, contrary to their inclinations, a specially compelling or deterring motive for doing or not doing them.

Both reasons and obligations are here reduced to motivating states, or empirical facts about such states. As Falk notes, 'what are here called . . . obligations would in one sense be facts of nature in their ordinary empirical meaning'.

Normativity, on Falk's view, is provided by the gap between our actual motives and the motives that we would have if we reviewed the facts. When someone claims that he 'ought' to do something, in Falk's motivational sense, what this person means is that, though his 'impulse or desire' to do this thing may not now be 'sufficiently strong, dispositionally he was under an effective and overriding compulsion to do it'.

Falk's motivational 'ought' is not, in my sense, normative. There is nothing normative in the compulsiveness or inescapability of our desires. That can be partly shown by considering what Falk's view implies. We have seen that, on Brandt's view, for our desires to be rational, it is enough that we be incurably insane. Similar remarks apply to Falk's view. Suppose that, when I reflect on the facts, I find myself irresistibly impelled, against my other inclinations, to act in some crazy way, such as eating light-bulbs, or leaping over a precipice. On Falk's view, I would then be rationally obliged to act in these ways.

When Falk discusses morality, he notes that we are drawn to a pair of potentially conflicting views. We assume both that, as internalists claim, moral beliefs necessarily motivate, and that that, as externalists claim, morality applies to everyone, whatever their motivational states. These assumptions, Falk writes, produce a paradox. We are inclined to believe that

> our doing what we ought to do needs a 'justification' additional to that which we express by saying that we morally ought to do it. We can ask, 'is there . . . any real need for my doing it?'

But it can also seem absurd

> that moral conduct should require more than one kind of justification: that having first convinced someone that regardless of cost to himself he was morally bound to do some act we should then be called upon to convince him as well that he had some . . . sufficiently strong reason for doing this act. 'You have made me realize that I ought, now convince me that I really need to' seems a spurious request, inviting the retort 'If you were really convinced of the first you would not seriously doubt the second.'

This appearance of paradox, Falk then argues, comes from our failure to distinguish between two senses of 'ought': the motivational or reason-giving sense, and the sense that is used by those whom I call Moral Externalists. When we draw that distinction, we shall see that there are 'some people who can maintain . . . as a plain matter of fact that, though admittedly they are morally bound to do some act . . . there is no real need or sufficient reason for them to do it'. In making such a claim, 'what they mean is that there is no thought about this act which has the power to cause them to do it'. Since these people are not motivated, it is a 'plain matter of fact' that they need not do their duty.[33]

Falk then suggests that, since the morally externalist sense of 'ought' breaks the link between morality and reasons for acting, it should be abandoned. The moral sense of 'ought' should be 'identified' with the motivational sense. In this way, Falk writes, 'the connection of duty with sufficient motivation becomes logically necessary'.

This proposal also has unwelcome implications. According to Moral Internalists morality may not apply to those who lack moral motivation. That is why Harman, for example, claims that Hitler may not have acted wrongly. Falk's proposal is more extreme. Suppose that Hitler's strongest desires would have survived reflection on the facts. On Falk's proposal, fulfilling these desires would then have been Hitler's duty.

We can now turn to Korsgaard's view, which partly overlaps with Falk's. I cannot do justice to this view, whose complexity and scope make it unusually hard both to summarize and classify. Korsgaard combines Kantian, Humean, and existentialist ideas in unexpected, platitude-denying ways. My concern will be only with Korsgaard's account of normativity, and with her objections to practical realism.

Korsgaard asks, 'what, if anything, we really ought to do', and 'what *justifies* the claim that morality makes on us'. She calls this *the normative question*.

Realists, Korsgaard claims, cannot answer this question. Suppose, she writes, that

> you are being asked to face death rather than do a certain action. You ask the normative question: you want to know whether this terrible claim on you is justified. Is it really true that this is what you *must* do? The realist's answer to this question is simply 'Yes'. That is, *all* he can say is that it is *true* that this is what you ought to do.[34]

[33] W. D. Falk, *Ought, Reasons, and Morality* (Ithaca, NY: Cornell University Press, 1986).

[34] Korsgaard, *Sources*, 38.

In this and similar passages, Korsgaard's objections seem to be one or more of the following:

(A) Realists discuss the wrong question.
(B) Realists cannot convince us that some answer to our question is really true.
(C) Even if our question had some true answer, that would not solve our problem.
(D) Ours is not a question to which some truth could be the answer.

These objections, I shall argue, fail. If Korsgaard's question could not be answered by some truth, it cannot be normative. When there are answers to normative questions, these answers must be truths of the kind that realists describe. And, if we cannot convince some people that there are such truths, that is no objection to realism.

Different writers, Korsgaard says, ask her question in different ways, since they differ in what they regard as the *normatively loaded* word. Thus, for Prichard, this word is *obligation*, for Moore, it is *good*, and for Nagel, it is *reason*. Korsgaard therefore gives her question several formulations. In the passage just quoted, Korsgaard's doubter asks

Q1: Is it really true that this is what I must do?

Realists cannot help us, Korsgaard says, because their answer to this question is simply 'Yes'. Realists do not support their answer. As she writes: 'if someone falls into doubt about whether obligations really exist, it doesn't help to say "ah, but indeed they do. They are real things." Just now he doesn't see it, and therein lies his problem'.[35]

On the most straightforward reading, Q1 means

Q2: Is it really true that this act is morally required?

But, if this were Korsgaard's question, realism might provide the answer. It might be really true that this act is morally required. If that were true, it would be no objection to realism that Korsgaard's doubter doesn't see this truth.

Korsgaard's question is, however, different. Thus she writes:

> the realist . . . can go back and review the reasons why the action is right. . . . But this answer appears to be off the mark. It addresses someone who has fallen into doubt about whether the action is really required by morality, not someone who has fallen into doubt about whether moral requirements are really normative.

[35] Korsgaard, *Sources*, 38.

Korsgaard's doubter isn't asking whether he is really morally required to face death. He believes that this action *is* required. He is asking whether this requirement is really normative.

Korsgaard's question might be

> Q3: *Should* I do what I am morally required to do?

Korsgaard writes, for example, 'Should we allow ourselves to be moved by the motives which morality provides?' For this to be a good question, its sense of 'should' cannot be moral. But it might be prudential. If morality were good for us, Korsgaard claims, that might answer the normative question. She also claims that, for some moral requirement to be worth dying for, violating this requirement must be worse than death. These remarks suggest that her question is

> Q4: If I did what is morally required, would that be good for me?

A similar question would be whether, in doing what is morally required, we would be acting on motives that were good for us. As Korsgaard writes:

> We can then raise the normative question: all things considered, do we have reason to accept the claims of our moral nature, or should we reject them? The question is not 'are these claims true?' as it is for the realist. The reasons sought here are practical reasons: the idea is to show that morality is good for us.

If Korsgaard's question were Q4, realism could, if true, provide the answer. Realists could appeal to truths about what is good or bad for us. They might not be able, in some cases, to defend the answer Yes. Perhaps, as Sidgwick argued, acting morally could be bad for us. But that would be no objection to realism.

Korsgaard's question is not, however, whether morality is good for us. As she would say, even if some act would be best for us, and were thus prudentially required, we could still ask whether that requirement was really *normative*.

A better suggestion is

> Q5: If I did what is morally required, would I be acting on motives that I am glad to have?

When we understand the motives that morality provides, we should ask, Korsgaard writes, whether we *endorse* these motives. And she ties normativity to reflective endorsement. But Q5 is too weak to be the normative question. Return to Korsgaard's first formulation:

> Q1: Is it really true that this is what I *must* do?

As Korsgaard writes elsewhere, she is asking what we *have* to do. Normativity, in its clearest form, involves a *requirement*.

Korsgaard's question is not, we have seen, about either moral or prudential requirements. But it might be

> Q6: Is this act *rationally* required? Is it what I have most reason, or overwhelming reason, to do?

This interpretation seems the best. Realism fails, Korsgaard argues, because it fails to understand the difference between theoretical and practical reasoning. According to realists, when some act is rationally required, that requirement is a normative fact, which holds 'independently of the agent's will'. On such a view, Korsgaard claims, we could ask

> Q7: *Why* should I do what I am rationally required to do?

And to this question, Korsgaard argues, realists would have no answer. Rational requirements, if understood in a realist way, would not have normative force. We might have no reason to do what, according to realists, reason required.

7

Korsgaard's strongest critique of realism comes in her discussion of instrumental reasons.[36] According to

> *the instrumental principle*, reason requires us to take the means to our ends.

On internalist, desire-based theories, this is the central principle of practical reasoning. On value-based theories, for the instrumental principle to apply, our aims must be rational, or worth achieving. But, when we have such aims, our reasons to pursue these aims give us derivative reasons to take the necessary means. So, on both kinds of theory, some form of the instrumental principle is uncontroversial.

Before giving her own, Kantian account of this principle, Korsgaard criticizes two others. One is the empiricist account, given by writers such as Hume and Falk. This account, Korsgaard writes, 'explains how instrumental reasons can motivate us, but at the price of making it impossible to see how they could function as requirements or guides'.[37] That objection, I have claimed, is justified.

Korsgaard also rejects the realist account, given by writers such as Sidgwick and the later Nagel. This account, Korsgaard writes, 'allows instrumental

[36] In The Normativity of Instrumental Reason', henceforth *NIR*, in Garrett Cullity and Berys Gaut (eds.), *Ethics and Practical Reason* (Oxford: Oxford University Press, 1997).

[37] *NIR* 219.

reasons to function as guides, but at the price of making it impossible for us to see any special reason why we should be motivated to follow these guides'. Realists 'cannot provide a coherent account of rationality'. According to them:

> rationality is a matter of conforming the will to standards of reason that exist independently of the will, as a set of truths about what there is reason to do . . . The difficulty with this account . . . exists right on its surface, for the account invites the question why it is rational to conform to those reasons, and seems to leave us in need of a reason to be rational.[38]

If realists were asked why it is rational to respond to reasons, they could answer: 'That is what being rational *is*. We are rational if we want and do what we have most reason to want and do.'

Korsgaard considers this reply. She writes:

> There is one way in which the realist strategy might seem to work. We may simply *define* a rational agent as one who responds in the appropriate way to reasons, whatever they are, and we may then give realist accounts of all practical reasons.

This reply, Korsgaard objects, would make realism trivial.

Realism would indeed be trivial if it were made true by a stipulative definition. Suppose we asked a different question: whether it is always rational to do our duty. In considering that question, it would be no help to define 'rational' to mean 'doing our duty'. Since that is not what 'rational' actually means, our proposed redefinition could not answer our question.

Consider next the question whether it is always right to do our duty. We might claim: 'Yes. That is what moral rightness *is*.' This claim is analytic, since it is implied by the meaning of the words 'right' and 'duty'. And, since we have not redefined these words, our claim answers this question. It is of course trivial to claim that it must be right to do our duty. But that claim is trivial only because it is so obviously true.

The same applies to the realist claim that to be rational is to respond to reasons. When realists make that claim, they are not appealing to a stipulative redefinition. Given the meaning of 'rational' and 'reason', their claim is another analytic truth. This claim is also trivial, because so obviously true. But that does not make realism trivial. According to realists, there are non-trivial truths about what we have reason to care about, and do.

Return now to Korsgaard's claim that, if rationality were a matter of responding to such truths, that would 'leave us in need of a reason to be

[38] *NIR* 240.

rational'. This claim is also far from trivial. According to realists, we are rational if we want, and do, what we have most reason to want and to do. Korsgaard's suggestion therefore is that, if realism were true, we might have no reason to want, and to do, what we had most reason to want and do.

For this suggestion to be coherent, Korsgaard must be using 'reason' in two senses. She cannot mean that, if we have a normative reason to do something, we might have no such reason to do this thing. But her point might be this. According to realists, the fact that we had this normative reason would be a truth that was independent of our will. If that were so, Korsgaard may mean, we might still need a *motivating* reason to do this thing. We might not be motivated to do what we believed that we had this reason to do. If realists could not exclude this possibility, that might seem to count against their view.

Other passages support this reading. Thus Korsgaard writes:

> realism about reasons . . . may be criticized on the grounds that it fails to meet the internalism requirement. . . . On a realist interpretation, astonishingly enough, even instrumental reasons fail to meet this requirement. For all we can see, an agent may be indifferent to the fact that an action's instrumentality to her end constitutes a reason for her to act.[39]

Korsgaard's objection seems here to be that, if it were an independent truth that we had reason to do whatever would achieve our ends, we might recognize that truth but fail to be motivated to do these things.

For this to be an objection to realism, Korsgaard would have to be appealing to Belief Internalism. She could then say that, if beliefs about reasons were beliefs about such independent truths, we could not explain how these beliefs necessarily motivate us. As we have seen, however, Korsgaard rejects Belief Internalism. She refers to 'the strange idea that an acknowledged reason could never fail to motivate'.[40]

Similarly, when she discusses morality, Korsgaard writes:

> If someone finds the bare fact that something is his duty does not move him to action, and asks what possible motive he has for doing it, it does not help to tell him that the fact that it is his duty just is the motive. That fact isn't motivating him just now, and therein lies his problem.[41]

Korsgaard here clearly rejects Moral Belief Internalism.

[39] *NIR* 242.

[40] 'Scepticism about Practical Reason', in *Creating the Kingdom of Ends*, henceforth *CKE* (Cambridge: Cambridge University Press, 1996), 331.

[41] *Sources*, 38.

In this second passage, Korsgaard might be appealing to Internalism about reasons. According to Deliberative Internalism, we cannot have a normative reason to do something if, though having deliberated on all the relevant facts, we are not at all motivated to do this thing. Korsgaard's imagined doubter *is* deliberating on the facts, including the fact that he has a certain duty. When he doubts that this duty is really normative, his point may be this: since he is not motivated to do his duty, the fact that he has this duty does not give him any reason for acting. If she were a Deliberative Internalist, Korsgaard would agree.

If this were Korsgaard's point, she would be rejecting some forms of realism. Some realists are Externalist Moral Rationalists, who believe that an act's rightness is always a reason for doing it. Deliberative Internalists reject this view. Other realists, however, also reject this view. Some believe that, though there are some external reasons, these are given only by facts about our own well-being. And, what is more important here, some realists are themselves Internalists, of a non-reductive kind. As that implies, if Korsgaard were appealing to Internalism about reasons, that would not explain why, according to her, realists cannot explain even the simplest instrumental reasons. Her objections to realism must be different.

One of her objections is the following. According to the realists she is considering, it is an independent normative truth that we have reason to do what is needed to achieve our aims. But realists have not explained how our awareness of this truth motivates us. When she discusses moral realism, Korsgaard often makes such claims. The eighteenth-century realists, she writes,

> did not explain *how* reason provides moral motivation. They simply asserted that it does. For Samuel Clarke, for instance, it is a fact about certain actions that they are 'fit to be done'. It is a self-evident truth built into the nature of things, in the same way that mathematical truths are built into the nature of things (whatever that way is). But people do not regulate their actions, love, hate, live, kill, and die for mathematical truths. So Clarke's account can leave us completely mystified as to why people are prepared to do these things for moral truths.[42]

Realists might reply as follows. We do not act upon mathematical truths, except in a purely instrumental way. But, when we believe that we ought rationally to accept the conclusion of some piece of mathematical or logical reasoning, it is not a mystery how that belief may lead us to accept that

[42] Ibid. 12.

conclusion. Similarly, when we believe that we ought, either rationally or morally, to act in some way, it is not a mystery how these beliefs may lead us to act.

This reply, Korsgaard might say, overlooks the difference between theoretical and practical reasoning. Since mathematics is concerned with truth, it is not mysterious how mathematical reasoning can affect our beliefs. Practical reasoning, in contrast, is not about what we should *believe*, but about what we should *do*. Realists, Korsgaard thinks, misunderstand this difference. They mistakenly regard ethics as another branch of theoretical reasoning, whose aim is knowledge. They assume that, when we ask 'practical normative questions . . . there is something . . . that we are trying to find out'.[43] On their view, 'our relation to reasons is one of seeing that they are there or knowing truths about them'. Realism fails, Korsgaard claims, because no knowledge of such truths could answer normative questions. In her words:

> Suppose it is just a fact, independently of a person's own will, that an action's tendency to promote one of her ends constitutes a reason for doing it. Why must she care about *that* fact?[44]

In asking why this person *must* care, Korsgaard might again be asking an explanatory question. She might mean: 'Why must it be true that this person cares? If it is such an independent fact that this person has this reason for acting, how can it be necessarily true that this person cares about this fact?' But, as we have seen, Korsgaard denies that beliefs about reasons necessarily motivate us.

In asking why this person must care, Korsgaard may instead be asking a justificatory question. She may mean, 'If it is such an independent fact that this person has a reason to do what will achieve her ends, why is it rationally required that she care about this fact?' Realists might answer: 'If this person's ends are rational, because she has reasons to have them, she has these same reasons to care whether her acts will achieve these ends.'

Korsgaard might now revise her question. She might say:

> Suppose it is just a fact, independently of this person's will, that she is rationally required to care whether she will achieve her ends. Why must she care about *that* fact?

Realists would answer by appealing to another normative fact. They might claim: 'If we are rationally required to care about something, we are

[43] *Sources*, 44. [44] *NIR* 241.

rationally required to care whether we care about this thing.' Korsgaard, however, might reply:

> If there is such a rational requirement, why are we rationally required to care about *that*?

Realism, Korsgaard claims, faces an infinite regress. In her words, if the instrumental principle 'is to provide the needed connection between the rational agent and the independent facts about reasons, it cannot in turn be based on independent facts'. In reply to Korsgaard's questions, all that realists can do is to appeal to another such fact, or truth. But, if that truth is also independent of our will, it cannot, Korsgaard claims, have normative force. Such truths cannot answer the normative question.

This objection, I shall argue, fails. But we should first consider Korsgaard's proposed alternative to realism. If Korsgaard were right, what *could* answer the normative question?

There are at least two other possibilities. This question might be answered by a truth that is *dependent* on our will, because it is *about* our will. Or this question might be answered, not by a truth at all, but by our will.

In some contexts, it would be important to distinguish these possible answers to Korsgaard's question. The first is a form of normative naturalism; the second a form of non-cognitivism. But, for our purposes here, it will be enough to consider what these answers have in common: their appeal to our will.

8

Modern thought about normativity, Korsgaard suggests, went through four stages. Such thought began, in the seventeenth century, with *voluntarism*, or an appeal to the will. According to Hobbes, Locke, and others, normativity consists in, or is created by, some law or command, issuing from the will of some external power, such as a sovereign or God. Realists like Clarke and Price replied that, if we ought to obey such laws or commands, this must be an independent moral truth. In Korsgaard's third stage, realism was rejected as both metaphysically incredible and incapable of answering the normative question. Sentimentalists, like Hutcheson and Hume, appealed instead to our attitudes and second order desires, or to reflective endorsement. This view, Korsgaard argues, though an advance on realism, cannot fully explain normativity. In her fourth, Kantian stage, an appeal to *rational autonomy* finally answers the normative question.

Korsgaard's use of the word 'rational' can make her view look like the realism that she rejects. Thus she writes of our being 'guided by what reason presents as necessary'. But she calls that only a 'preliminary formulation'; and she goes on to argue that 'a rational agent is guided by herself, that is, that being governed by reason amounts to being self-governed'.

On this Kantian view, Korsgaard claims, it turns out that

> voluntarism is true after all. The source of obligation is a legislator. The realist objection—that we need to explain why we must obey that legislator—has been answered, for this is a legislator whose authority is beyond question and does not need to be established. It is the authority of your own mind and will . . . It is not the bare fact that it would be a good idea to perform a certain action that obligates us to perform it . . . it is the fact that we *command ourselves* to do what we find it would be a good idea to do.[45]
>
> The reflective structure of human consciousness requires that you identify yourself with some law or principle that will govern your choices. It requires you to be a law to yourself. And that is the source of normativity.[46]

These are not rhetorical claims. Korsgaard means what she says. On her view, there are no independent truths about reasons which should guide our decisions and our acts. Like normativity, reasons are created by our own will.

Korsgaard sees the implications of this view. As a result, her concept of a reason is very different from the one that realists use. Return, for example, to a passage that I have discussed before. Korsgaard writes:

> According to internalists, if someone knows or accepts a moral judgment then she must have a motive for acting on it. The motive is part of the content of the judgment: the reason why the action is right is a reason for doing it. According to externalists: this is not necessarily so: there could be a case in which I understand both that and why it is right for me to do something, and yet have no motive for doing it. Since most of us believe that an action's being right is a reason for doing it, internalism seems more plausible.[47]

When I first read this passage, I found it baffling. For this passage to make sense, I assumed, Korsgaard must be using the words 'motive' and 'reason' to mean the same. When she says that, according to externalists, we might have 'no motive' for doing what we knew to be right, she must mean that we might not be motivated to act in this way. This use of 'motive' must refer

[45] *Sources*, 104–5.

[46] Ibid. 104.

[47] *CKE* 43.

to a psychological state. But when she says that, according to internalists, an action's being right is a reason for doing it, she must be using 'reason' to mean 'normative reason'. Since these uses of 'motive' and 'reason' cannot mean the same, I could not imagine what, when she wrote this passage, Korsgaard was intending to claim.

I overlooked the obvious way in which this passage would make sense. Korsgaard may believe that, though the words 'motive' and 'reason' do not always mean the same, what they refer to is the same. If that is so, though the concept of a normative reason is not the concept of a psychological state, normative reasons *are* psychological states. They are states of our will, or states that our will creates.

Here is one simple argument for this view. We might claim:

> Normative reasons, when we act upon them, are *motivating* reasons, or the reasons why we acted as we did.
>
> Motivating reasons are psychological states.

Therefore

> Normative reasons are psychological states.

This argument, however, wrongly conflates two views about motivating reasons. On what we can call the *non-psychological view*, our motivating reasons are what we believe, or what we want, when these beliefs and desires explain our decisions and our acts. In the cases that are most relevant here, our motivating reason is what we believe to be our normative reason. In such cases, when our belief is true, the same fact is both our normative and our motivating reason. For example, suppose we know that

> (A) by telling some lie, we would save someone's life.

If we tell this lie, and are later asked why, we would say, 'Because it saved someone's life'. On this view, the fact reported in (A) is both a normative reason for doing what we did, and our motivating reason for doing it. On the *psychological view*, motivating reasons are not what we believe, or what we want, but the psychological states of having these beliefs or desires. Thus, in this example, our motivating reason was not (A) itself, but our belief in (A). (If they held this view, Humeans would add that this belief was only part of our motivating reason, since, for beliefs to motivate, they must be combined with desires.)

In the argument just sketched, the first premise assumes the non-psychological view, but the second assumes the psychological view. Since these are different views, the argument is invalid. It cannot show that, when we have normative reasons to act in some way, these reasons are motivating states.

Though Korsgaard does not appeal to this argument, she seems, I have said, to accept its conclusion. Consider, for example, her account of how her internalist view differs from that of externalists like Ross. Suppose that you act rightly, for some moral reason. Korsgaard writes that, according to these externalists,

> (1) 'The reason why the act is right and the motive you have for doing it are separate items',[48]

whereas, on her internalist view,

> (2) 'the reason why the act is right is the reason, and the motive, for doing it'.

It seems clear that (2) means

> (3) the reason why your act was right was both a normative reason for doing what you did, and your motivating reason for doing it.

Korsgaard's claim is that, while externalists like Ross distinguish between normative and motivating reasons, internalists like her reject that distinction.

There is one obvious way to explain this claim. Korsgaard might mean that, while externalists like Ross accept the psychological view of motivating reasons, internalists like her accept the non-psychological view. Suppose that, as in our example, you tell a lie because you believe that

> (A) this act would save someone's life.

If Ross accepted the psychological view, he might have claimed: 'The reason why your act was right was the fact that, as you believed, it saved someone's life. Your motivating reason was not (A) itself but your believing (A).' If Korsgaard accepted the non-psychological view, she might claim: 'On the contrary, the fact reported in (A) was not only the reason why your act was right, and a normative reason for doing what you did. This fact was also your motivating reason for doing it.'

This cannot, however, be what Korsgaard means. If she were thinking of the distinction between these two views, she would have known that Ross did not accept the psychological view, and that nothing in externalism supports that view. Similarly, many internalists do accept that view, as internalism allows them to do.

There is a better way to explain Korsgaard's claim. First, like these other internalists, she may accept the psychological view. She may regard motivating reasons as motivating states, such as beliefs, or desires, or

[48] This quotation continues, 'although it is nevertheless the case that the motive for doing it is "because it is right" '.

states that involve the agent's will. Korsgaard may also hold another, more important view. As I have said, she may believe that *normative* reasons are motivating states. Her point may then be this. According to externalists like Ross,

(4) the reason why your act is right is not the same as the psychological state that motivates you to do it,

whereas, on her internalist view,

(5) the reason why your act is right *is* the state that motivates you to do it.

Other passages support this reading. Thus Korsgaard writes:

> Ross in effect separates the justifying reason—the fact that the action is right—from the motivating reason—the desire to do what is right . . . .

Korsgaard then criticizes Ross's view. This suggests that, on her view, we should *not* separate the fact that some act is right from the agent's being motivated to do it. Such a claim would be too loosely worded, since the *fact* that some act is right cannot be a motivating state. Facts and states are in different categories. But, as before, Korsgaard's point might be that the *reason* why the act is right is a motivating state.

Korsgaard's view, so described, may seem obviously false. Would not Korsgaard agree that, in my example, the reason why your act was right was the fact that it saved someone's life? And, if this reason was a fact, then, as I have just implied, it too cannot be a motivating state.

This objection, Korsgaard might say, mis-states this moral reason. Suppose that, though your act did indeed save someone's life, you believed falsely that it would kill that person. Your act would then have been wrong. So the reason why your act was right was *not* the fact that it saved someone's life. It was your belief in this fact. And that belief *was* a motivating state.

This reply shows the need for another distinction. I have suggested that, on Korsgaard's view,

(5) the reason why your act is right is the state that motivates you to do it.

But this claim is ambiguous. When applied to our example, (5) might be making a pair of claims:

(6) The reason why your act was right was your belief that it would save someone's life.

(7) This belief was the state that motivated you to act.

If this were Korsgaard's view, however, she would not be disagreeing with externalists like Ross. Ross could have accepted both these claims.

For (5) to describe a view that Ross would have rejected, it must have a different sense. On the view just described, even though your belief was a motivating state, that is not what made your act right. In the sense that I intend, (5) means

> (8) The reason why some act is right—or what makes it right—is the agent's being in a certain motivating state.

Though (8) is suggested by some of the claims by Korsgaard quoted above, those claims may also have been too loosely worded. Like Ross, Korsgaard would reject (8). Thus she would agree that, in our example, the reason why your act was right was your belief that it would save someone's life. (8), however, points us towards what I believe to be Korsgaard's view. She could agree with Ross about the reasons why certain acts are morally right. She and Ross disagree at another, deeper level.

Writers differ, Korsgaard says, in what they regard as the *normatively loaded* words, or concepts. For Ross, these are such words as 'right' and 'morally required'. For certain other normative realists, they are such words as 'reason' and 'rationally required'. But, for Korsgaard, these words are merely classificatory. When these words are correctly applied, they can be used to state truths about what is morally or rationally required. But such truths do not, in themselves, have normative force. The normatively loaded words are, for Korsgaard, 'obligatory', 'binding', and one use of 'necessary'.

Return to Korsgaard's imagined doubter who is morally required to face death. This person does not doubt that this act is morally required. He is asking whether this requirement is really normative. Is it really true that he *must* face death? The answer to this question, Korsgaard claims, cannot be provided by some truth that is independent of this person's will. It must be provided either by a fact about his will, or by his will. Though Korsgaard would reject (8), she would, I believe, accept

> (9) The reason why some act is *normatively necessary* is the agent's being, through an act of will, in a certain motivating state.

Such a state is partly passive. For a law to be normative, Korsgaard writes: 'It must get its grip or hold on me.' 'To be obliged to the performance of an action is to believe that it is a right action and to find in that fact a kind of motivational necessity.'[49] But *we* are the source of such necessity. As Korsgaard also writes, 'Nothing except my own will can make a law normative *for me*.'

Korsgaard's account of normativity, as she often claims, differs deeply from a realist account. This difference, as I have said, is sometimes veiled by her

[49] My italics.

use of certain words. Thus she writes that she uses 'the term "normativity" to refer to the ways in which reasons direct, guide, or obligate us to act', or 'to what we might call their authoritative force'. Realists would accept all of that. Similarly, Korsgaard claims that the normativity of morality consists in

> its power to bind, or justify, and its power to motivate, or excite.

If Korsgaard were using 'justify' in its ordinary sense, this use of 'normativity' would differ only verbally from a realist's use. While Korsgaard would be taking normativity to have two elements—justifying and motivating force—realists use 'normativity' more narrowly, so that it refers only to justifying force.

This disagreement, however, is more than verbal. Like the young Nagel, Korsgaard often uses 'justify' to mean 'persuade' or 'motivate'. When we do moral philosophy, she writes, we are asking 'what *justifies* the claims that morality makes on us', or 'whether we are justified in according this kind of importance to morality'. But she then writes:

> A moral sceptic is not someone who thinks that there are no such things as moral concepts, or that our use of moral concepts cannot be explained, or even that their practical and psychological effects cannot be explained. Of course these things can be explained somehow. Morality is a real force in human life, and everything real can be explained. The moral sceptic is someone who thinks that the explanation of moral concepts will be one that does not support the claims that morality makes on us. He thinks that once we see what is really behind morality, we won't care about it any more.[50]

For Korsgaard, as for Nagel, moral sceptics are not people who doubt the *truth* of moral claims. Korsgaard does not even say whether, according to her sceptic, moral concepts can be truly applied. And, when her sceptic doubts that we can *support* morality's claims on us, thereby *justifying* these claims, what he doubts is whether, when we understand these claims, or what lies behind them, we shall care about morality. We justify morality's claims, in Korsgaard's sense, if we get people to care about these claims, thereby motivating them.

It matters greatly whether we can support morality in Korsgaard's sense. Suppose that, unless Korsgaard's doubter does what is morally required, several other people will die. Those other people's lives would then depend on whether Korsgaard's doubter can be motivated by morality's claims—or whether, in Korsgaard's phrase, morality is normative *for him*. But this

[50] *Sources*, 13–14.

sense of 'normative', like Korsgaard's sense of 'justify', does not even partly overlap with the sense that realists employ. Normativity, on their view, neither includes nor requires motivating force.

Consider next another passage. Internalism, Korsgaard writes,

> captures one element in our sense that moral judgments have *normative* force: they are *motivating*. But some philosophers believe that internalism, if correct, would also impose a restriction on moral reasons. If moral reasons are to motivate, they must spring from an agent's personal desires and commitments. This is unappealing, for unless the desires and commitments that motivate moral conduct are universal and inescapable, it cannot be required of everyone. And this leaves out the other element of our sense that moral judgments have normative force: they are *binding*.[51]

For moral judgments to be binding, Korsgaard here implies, they must be *universal* and *inescapable*. That suggests the familiar claim that, whatever our desires or commitments, moral judgments apply to all of us. But that is not what Korsgaard means. Korsgaard's doubter does not deny that he is morally required to face death. He is asking whether this requirement is really *binding*, or whether it *obligates* him. And Korsgaard does not use those words in their moral sense. She writes:

> 'obligation' refers to . . . the *requiredness* of an action, to its normative pull.
>
> An obligatory action is one that is binding—one that it is necessary to do.

When Korsgaard calls obligatory actions *required* or *necessary*, she does not mean that they are morally required, or morally necessary. Nor does she mean that they are rationally required, or necessary. That is why she claims that realism, even if true, could not answer the normative question. Suppose that, as realists believe, there are irreducibly normative facts, or truths, which hold independently of our will. And suppose that, as one such fact, we are morally and rationally required to act in some way. Korsgaard would say, 'Why must we care about *that* fact?'

If we are obliged or bound, in Korsgaard's sense, that is a fact about our own wills. As she writes:

> The primary deliberative force of saying 'I am obliged to do this' is . . . 'my judgment that it is right impels me to do this.'

Though Korsgaard claims that normativity has two elements, the power to motivate and to bind, she does not regard these as two separate elements.

[51] *CKE* 43.

Normativity, on her view, is one kind of motivating force: it is what she calls the 'motivational necessity' of normative beliefs.

As before, the disagreement here is deep. According to realists, normativity consists in truths about reasons, or about what is morally or rationally required. On Korsgaard's view, no such truths could be in themselves normative. When such truths are normative, it is we who make them so, either by an act of will, or by finding that our will is already irresistibly engaged.

9

Which of these is the better view?

There are here three questions:

Q1: How should we understand the normative concepts that we actually use?

Q2: Could there be concepts that were, as realists claim, irreducibly normative?

Q3: If there were such concepts, could they be truly applied?

The first question is, in a way, the least important. We might start by asking what the word 'normative' means. But this word has many uses; and both Korsgaard and the realists are entitled to theirs. It is more fruitful to ask how we should understand such words as 'should', 'right', and 'reason'. When we ask whether we should do something, or have a reason to do it, are we asking a question about our own motivation? Are we asking whether we *will* this act, or whether we find ourselves impelled to do it?

The answer, I believe, is No. But that answer would not refute Korsgaard's view. She could claim to be describing, not what we do mean, but what we should mean. Her view, she might say, gives the right account of what practical reasoning really involves. In the same way, even if Korsgaard describes what we do mean, that would not refute realism. Nor would it refute realism if most of us use such words in some other non-realist sense, such as those described by non-Korsgaardian naturalists, or by non-cognitivists. Even if that is true, it might be possible to use these words in a different, irreducibly normative way.

We should ask whether that is possible. *Could* words like 'should' and 'reason' have the sense that realists take them to have? Is practical realism intelligible? And, if the answer here is Yes, we can turn from meaning to truth. Do these concepts apply to reality?

There are grounds for answering No to both these questions. Irreducibly normative concepts could not, I have said, be explained in other terms. Such concepts, it is often claimed, could not be learnt or understood. Nor, it is claimed, could there be irreducibly normative properties or truths. Normativity, as realists understand it, is a mere dream.

Before I turn to these claims, I shall consider Korsgaard's own distinctive objection to practical realism. Korsgaard claims that realism, even if true, would be irrelevant. Normative questions must be answered, not by truths about reasons, or about moral and rational requirements, but by truths about ourselves. And these are truths that we create, by acts of will.

Korsgaard's objection is, I believe, mistaken. Perhaps there are no irreducibly normative truths. If that is so, Korsgaard's account of normativity may be the best that we could hope to defend. But realism, if true, would be the better view.

In defending this claim, I shall continue to use the word 'normative' in what I shall call the realist sense. But the disagreement here is not about what this word means. It is about what practical reasoning, at its best, either does or could involve.

Korsgaard's account of normativity is, as she would agree, reductive. It is not as bleak as that of most naturalists or non-cognitivists. Most naturalists appeal merely to certain facts about our own motivation, or about the effects of our acts. Most non-cognitivists appeal merely to certain attitudes, or mental acts, such as the acceptance of some imperative. Korsgaard's view makes both these appeals, but she carries them to a deeper level.

Korsgaard shows that, despite their other differences, there are striking similarities between the views of Hume and Kant, and, among more recent writers, Sartre, Hare, Williams, Brandt, and Gibbard. These writers all reject realism, and they all place normativity, in Korsgaard's words, not 'in the metaphysical properties of actions' but 'in the motivational properties of people'. Similarly, according to all these writers, nothing is in itself good or bad. Just as 'moral properties are the projections of human dispositions', 'our relation to values is one of creation and construction'.

Of this family of views, Korsgaard's Kantian version may be the least reductive. Some of the strengths of her view I shall barely mention here. One example is her appeal to what she calls *practical identity*. On her view, it is not merely reasons and values that, by our acts of will, we create.

We even create ourselves, as free and rational agents. If it were not for these acts of will, we would not exist as agents, but would be only places where events occur, or bodies that were governed by conflicting instincts and desires. In these self-creating acts of will, we give ourselves laws, or endorse normative principles. The 'function' of these principles, Korsgaard claims, 'is to bring integrity and therefore unity—and therefore, really, existence—to the acting self.' Though we are the source of normativity, if we impose this category on reality in the kind of way in which, on Kant's view, we impose the categories of space and time, that might be claimed to go some way towards fulfilling the realist's dream.

While this view offers more than most other forms of naturalism or non-cognitivism, Korsgaard's account of normativity is still, I believe, bleak. And her objection to realism does not, I believe, succeed.

Consider first Korsgaard's claim that, to answer the normative question, we must appeal to the motivational necessity of normative beliefs.

After claiming that realism 'seems to leave us in need of a reason to be rational', Korsgaard continues:

> To put the point less tendentiously, we must still explain why the person finds it *necessary* to act on those facts, or what it is about her that makes them normative *for her*. We must explain how these reasons get a grip on the agent.

Normativity, so understood, is a kind of unavoidable and irresistible motivation. Korsgaard's doubter asks whether he really *must* face death. And Korsgaard says that, according to some writers, the word 'right' is 'normatively loaded', so that we should not call some act right unless we are 'sure that we really *have* to do it'.

Korsgaard's account of such necessity partly overlaps with Falk's. According to Falk, when we ask 'Must I do that?', we can best be taken to be asking whether there is any belief 'sufficiently compelling to make' us do it. Rational necessity is the presence of a motive that is both 'an effective and overriding compulsion', and a compulsion that no further reflection would dislodge. We are rationally compelled to act in some way when it is true that, if we reflected on the facts, we would be irresistibly and unchangeably moved to do so.

Falk's view, I have claimed, abandons normativity. An irresistible impulse is not a normative reason. Nor can such an impulse be made normative by its ability to survive reflection on the facts. Moreover, since Falk appeals only to the strength of the agent's motives, his proposed equation of morality with such 'rational necessity' yields incredible conclusions. Thus it could imply that it was Hitler's duty to act as he did.

Korsgaard's view differs from Falk's in ways that she claims avoid such objections. On her view, for our strongest impulse to give us reasons for acting, we must reflectively endorse that impulse. Thus she writes:

> given the strength of the moral instinct, you [might] find yourself overwhelmed with the urge to do what morality demands even though you think that the *reason* for doing it is inadequate . . . Then you might be moved by the instinct even though you don't upon reflection endorse its claims. In that case the . . . theory would still explain your action. But it would not *justify* it from your own point of view. This is clear from the fact that you would wish that you didn't have this instinct.

And she writes that, according to Kant,

> the test of reflective endorsement is the test used by actual moral agents to establish the normativity of all their particular motives and inclinations. So the reflective endorsement test is not merely a way of justifying morality. It is morality itself.

Hitler's strongest motives would be likely to have passed this test. Korsgaard adds, however, 'I am not saying that reflective endorsement—I mean the bare fact of reflective endorsement—is enough to make an action right'.

I shall not consider here what else, on Korsgaard's view, would be enough to make some action right. My question is only about her claim that, unlike the realist's appeal to normative truths, her appeal to motivational necessity answers the normative question.

In assessing Korsgaard's claim, we should distinguish three kinds of practical necessity: or three senses in which, in practical reasoning, we might conclude that we *must* act in some way.

Consider first a claim like

(A) If you want to catch your train, you must leave now.

This use of 'must' expresses what we can call *instrumental* necessity. As several writers argue, such claims are not normative, since they merely report the causally necessary means to the achievement of some aim. (A) means, roughly, 'Given the distance to the station, catching your train would be impossible unless you leave now.' This non-normative use of 'must' is irrelevant here.

Consider next

(B) Since the building is on fire, you must jump into the canal.
(C) Since those children are your responsibility, you must rescue them.

These uses of 'must' are fully normative, since they claim that the acts in question are rationally or morally required. This gives 'must' what we can call its *requirement sense*.

An act is rationally required if it is not merely what we have most reason to do, but is what, as Williams writes, 'the weight of reasons overwhelmingly supports'. The same is true of some moral uses of 'must'. But moral requirements can also come, not from an overwhelming weight of moral reasons, but in a more direct way. Thus it might be morally necessary never to violate some constraint.

The word 'must' can also have what Williams calls the *incapacity sense*. We must do something, in this sense, if we could not possibly act differently. In the cases that are most relevant here, such incapacity is not physical, since this different act would be within our powers. Our inability depends instead on facts about our motivation, and may be the result of deliberation. It is in this sense that, for example, we may be unable to shoot some innocent person. When we ask what we ought to do, it becomes clear to us that there is something that we must do, because we couldn't act differently even if we tried, or because we couldn't even try. And what makes us incapable of acting differently might be our beliefs about what, in the requirement sense, we must do. Thus we might find it impossible to shoot this person because that would violate what we regard as an absolute constraint. Such cases involve what Williams calls *moral incapacity*.[52]

This incapacity sense of 'must', even when it takes this moral form, is quite different from the requirement sense. Williams notes a simple proof that these necessities are different. Suppose we claim that we must keep some promise, because that is morally required, or morally necessary. If we fail to keep our promise, because we give in to some temptation, we are not forced to withdraw our earlier claim. We can still believe that what we failed to do was indeed morally necessary. Things are different with the use of 'must' which states an incapacity. Suppose we say: 'I must keep my promise, since I couldn't possibly let her down'. If we fail to keep this promise, because we give in to some temptation, we must withdraw our earlier claim. 'I had to do it' implies 'I did it'. If we *did* act differently, we can't still claim 'I *couldn't* have acted differently'.

This point also shows that, as Williams argues, the incapacity sense of 'must' is not normative. Whether we have a reason to do something, or ought to do it, or are required to do it, cannot depend on whether we actually do it. In contrast, whether we must do something in the incapacity sense,

[52] 'Moral Incapacity', in *Making Sense of Humanity*. See also 'Practical Necessity', in *Moral Luck*.

because we couldn't act differently, does depend on whether, when given the opportunity, we do this thing.

Return now to Korsgaard's claim that, if there were normative facts that were independent of our will, realists could not explain why we find it 'necessary to act on those facts'. In trying to explain this necessity, Korsgaard might appeal instead to facts that are *not* independent of our will, because they are *about* our will. And her appeal might be to psychological necessity, or to the use of 'must' that reports an incapacity. But, if this were Korsgaard's view, it would not provide a better account of normativity. As we have just seen, this kind of necessity is not normative.

This point is easy to miss, since such psychological necessity may have both normative origins and normative implications. It may be our normative beliefs that make us incapable of acting differently. And, if we could not act differently even if we tried, or if we couldn't even try, that may undermine the claim that we ought to act differently. But psychological necessity, though it may have normative significance, is not normative necessity. That is most obvious in those cases in which such necessity is not produced by normative beliefs. If kleptomaniacs could not act differently, that doesn't make their stealing morally or rationally necessary.

When psychological necessity is produced by moral beliefs, there is a further complication. Consider some conscientious SS officer, whose oath to Hitler makes him incapable of disobedience. When this officer obeys some order to slaughter civilians, what he does is, in one sense, very wrong. But, according to Aquinas and others, it would also be wrong for this officer to do what he believed to be wrong. On this view, when it is psychologically necessary that we act on our moral beliefs, that may also be, even if our beliefs are mistaken, morally necessary. In such cases, whatever we did would be wrong. But even though these two necessities did in this way coincide, that would not make psychological necessity in any sense normative. It is this officer's moral belief that, according to Aquinas, would make it wrong for him to disobey his order. That belief would make this act wrong even if he were psychologically capable of such disobedience. And, if he did not have that belief, such incapacity would not have made it wrong.

Return now to Korsgaard's doubter, who asks whether he really must face death. This doubter could be using 'must' in either of these ways. He might be asking whether facing death is either morally or rationally necessary. These questions are normative; and, if they have answers, realists could give the answers. Korsgaard could not claim that, even if her doubter knew that facing death was, in those senses, necessary, he could still ask whether that was true. Korsgaard's doubter may instead be asking whether he is capable of acting differently. But, as Williams shows, that question is not normative.

Though Korsgaard sometimes appeals to psychological necessity, she would agree, I believe, that such necessity is not normative. As she writes elsewhere, 'This answer does not have the structure of reason-giving: it is a way of saying "I can't help it".'

10

There is a powerful objection, Korsgaard claims, to any realist view. Realists face an infinite regress, from which they cannot escape. That is why realism, even if true, could not answer the normative question.

'Justification', Korsgaard writes, 'like explanation, seems to give rise to an infinite regress; for any reason offered, we can always ask why.' We can indeed go on asking 'Why?' And, when we are asking for an explanation, the question 'Why?' sometimes has no answer. Most explanations must, in the end, appeal to some brute fact. But that does not, as some suggest, undermine these explanations. It shows only that not everything can be explained.

When we ask for a justification, things are different. Justifications can end with some irreducibly normative truth. And such truths are *not* brute facts. The most important normative truths could not have been false. If we ask why these truths are true, we can sometimes give no further answer. But, since these truths are not brute facts, they can provide full, or complete, justifications.

Korsgaard would reject these claims. On her view, even if there were such normative truths, they could not provide justifications. But, like several other writers, Korsgaard does not take seriously the possibility that there may be such truths. When she describes the justificatory regress, she ignores the answers that realists would give. She writes, for example:

> I ask to know why you are doing some ordinary thing, and you give me your proximate reason, your immediate end. I then ask why you want that, and most likely you mention some larger end or project.
>
> I can press on, demanding your reason at every step, until we reach the moment when you are out of answers. You have shown that your action is calculated to assist you in achieving what you think is desirable on the whole, what you have determined that you want most.[53]

[53] *CKE* 163–4.

Korsgaard here assumes that, in judging something to be most desirable, we are judging that we want it most. If we had that conception of desirability, Korsgaard would be right to claim that we would soon run out of answers. We would soon reach some desire for which we could give no further desire-based justification. Realists can appeal instead to a value-based conception. Our aims are desirable, realists can claim, when these aims have features that give us reason to have them, or to want to achieve them.

Korsgaard continues:

> The reasons that you have given can be cast in the form of maxims derived from imperatives. From a string of hypothetical imperatives, technical and pragmatic, you have derived a maxim to which we can give the abbreviated formulation:
>
> > 'I will do this action, in order to get what I desire'.
>
> According to Kant, this maxim only determines your will if you have adopted another maxim that makes it your end to get what you desire. This maxim is:
>
> > 'I will make it my end to have the things that I desire'.
>
> Now suppose that I want to know why you have adopted this maxim. Why should you try to satisfy your desires?

That is a good question, which rightly challenges desire-based theories. But, if you were a practical realist, you need not appeal to your desires. You could appeal to claims about what we have reason to want, and do. Your maxim might be:

> I will make it my end to achieve whatever I have most reason to try to achieve, because these are the ends that are most worth achieving.

Korsgaard's question would then become:

> Why should you try to achieve what you have most reason to try to achieve?

Such a question has no force. If we know that some aim is what we have most reason to try to achieve, we could not ask whether we have reason to try to achieve this aim.

Korsgaard continues:

> We are here confronted with a deep problem of a familiar kind. If you can give a reason, you have derived it from some more fundamental maxim, and I can ask why you have adopted that one. If you cannot, it looks as if your principle was randomly selected. Obviously, to put an end to a regress like this, we need a principle about which it is impossible, unnecessary, or incoherent to ask why a free person would have chosen it.

As before, Korsgaard ignores the realist's view. Any reason, she assumes, must be derived from some maxim, or principle, which we have *adopted*. To end the justificatory regress, we must find some principle about which we need not or cannot ask why we have *chosen* it. According to realists, however, we can appeal to normative truths about what we have reason to want, and do. And, if there are such truths, they are not principles that we adopt or choose. We *believe* truths.

We could of course be asked why we believe these truths. We might answer: 'Because they are true'. We might then be asked why, if some normative claim is true, that gives us a reason to believe this claim. But that is not a question about practical reasons, or the justification of our desires and acts. So, if there are such normative truths, they would end Korsgaard's justificatory regress.

There is another kind of question that Korsgaard might ask. Suppose, for example, that we are trying to relieve our own or someone else's suffering. Korsgaard asks why we are trying to achieve this aim, and we appeal to the truth that suffering is bad, or is a state that we have reason to try to relieve. Rather than asking why we believe this truth, Korsgaard might ask why it is true. Why is suffering, in this sense, bad?

When realists discuss this question, as Korsgaard notes, they have not found much to say. The badness of suffering, most realists would claim, is a fundamental truth, which neither has nor needs any further explanation. Korsgaard's answer to this question is more original. But we are not now asking why suffering is bad. We are asking whether, if there are truths of the kind to which realists appeal, these could answer normative questions, and end the justificatory regress. And, as before, the answer is Yes. If suffering really is bad, or is a state that we have reason to prevent and relieve, that justifies our wanting and our trying to achieve this aim. We could still ask why it is true that suffering is bad, or what, if anything, makes that true. But that is a theoretical or philosophical question. Though it is a question about practical reasons, it is not a practical question. In asking why suffering is bad, we are not asking what we have reason to want, or to do. So, as before, practical realists do not face a damaging infinite regress. Suppose we know that, as realists claim, we have reason to want, and to try, to relieve suffering. We might be asked, 'Why do you want to relieve suffering?' But, since 'Why?' asks for a reason, we can answer this question. We have this aim because we are rational, and we have a reason to have it. As Korsgaard says, we could always be asked further questions. Someone might say, 'If you have a reason to have this aim, why is that a reason for having it?' But that is even easier to answer. Any truth is true. If we have a reason, we have a reason.

In trying to answer the normative question, Korsgaard writes, we are engaged in what Kant called 'the search for the unconditioned'. We are looking

> for something which will bring the reiteration of 'but why must I do that?' to an end. The unconditional answer must be one that makes it impossible, unnecessary, or incoherent to ask why again . . .
>
> The realist move is to bring this regress to an end by *fiat*: he declares that some things are intrinsically normative. . . .

It isn't realists who end this regress by *fiat*. Unlike Korsgaard, realists do not believe that we can make something normative by willing that to be so. Nor do realists merely *declare* that some truth is normative. They believe that, as Korsgaard writes, when we ask normative questions 'there is something . . . that we are trying to find out'. On their view, these questions can have true answers, and these truths are normative, not because we declare them to be so, but because they are truths about reasons, or about what we are rationally or morally required to do.

On Korsgaard's view, even if there were such truths, they could not answer normative questions. To end the justificatory regress, we must appeal to motivational necessity, and to our own will. That, I have argued, is not so. Motivational necessities are not reasons, nor are they normative. And the regress could only be ended in the way that Korsgaard rejects. If we knew that we must do something, and why we must do it, we could not then ask, 'But why must we do it?'

As Korsgaard rightly claims, practical reasoning should not end with beliefs. To be fully practically rational, we must respond to reasons in our desires and acts. But it is the content of certain beliefs that provide the answers to practical questions. Normativity is not created by our will. What is normative are certain truths about what we have reason to will, or ought rationally to will.

Index