OXFORD STUDIES IN Agency and Responsibility

Volume 1

OXFORD

OXFORD STUDIES IN AGENCY AND RESPONSIBILITY VOLUME 1

This page intentionally left blank

Oxford Studies in Agency and Responsibility Volume 1

Edited by

DAVID SHOEMAKER

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

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

First Edition published in 2013
Impression: 1

British Library Cataloguing in Publication Data
Data available

ISBN 978–0–19–969485–3 (Hbk)
ISBN 978–0–19–969486–0 (Pbk)

Printed in Great Britain by
CPI Group (UK) Ltd, Croydon, CR0 4YY

Contents

This page intentionally left blank

Notes on Contributors

Matt Bedke is Assistant Professor of Philosophy, University of British Columbia

Michael E. Bratman is U. G. and Abbie Birch Durfee Professor in the School of Humanities and Sciences, and Professor of Philosophy, Stanford University

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

Sarah Buss is Associate Professor of Philosophy, University of Michigan

Zac Cogley is Assistant Professor of Philosophy, Northern Michigan University

Oisín Deery is PhD candidate in Philosophy, University of British Columbia

Luca Ferrero is Associate Professor of Philosophy, University of Wisconsin, Milwaukee

Daniel Jacobson is Professor of Philosophy, University of Michigan

Heidi L. Maibom is Associate Professor of Philosophy, Carleton University

Michael McKenna is Professor and Keith Lehrer Chair, Department of Philosophy and Center for the Philosophy of Freedom, University of Arizona

Dana K. Nelkin is Professor of Philosophy, University of California, San Diego

Shaun Nichols is Professor of Philosophy, University of Arizona

Paul Russell is Professor of Philosophy, University of British Columbia

David Shoemaker is Associate Professor, Department of Philosophy, Murphy Institute, Tulane University

Tamler Sommers is Associate Professor, Philosophy Department and Honors College, University of Houston

Matthew Talbert is Assistant Professor, Department of Philosophy, West Virginia University

This page intentionally left blank

Introduction

David Shoemaker

It is my pleasure to introduce the inaugural volume of *Oxford Studies in Agency and Responsibility*, a new entry in the *Oxford Studies* series. The papers in this volume are excellent illustrations of the important cutting-edge work being done in this field.

"Agency and Responsibility" is a new label for an old set of unified, yet highly diverse, concerns. It involves investigation into such familiar questions as these:

- What does it mean to be an agent?
- How (if at all) does the nature of personhood and personal identity across time bear on questions of agency?
- What is the nature of moral responsibility? Of criminal responsibility? What is the relation between moral and criminal responsibility (if any)? Is there a relevant conception of responsibility generally?
- What is the relation between responsibility (moral, criminal, or general) and the metaphysical issues of determinism and free will?
- What do various psychological disorders (e.g. autism, psychopathy, mental retardation, insanity, and dementia) tell us about agency and responsibility?
- How do moral agents develop? How does this developmental story bear on questions about the nature of moral judgment and responsibility?
- What is the will, willpower, and weakness (or strength) of will? What role can psychology play in the investigation of these issues generally?
- What do the results from neuroscience imply (if anything) for our questions about agency and responsibility?
- What is the nature of political agency (e.g. citizenship), and how does it differ, if at all, from moral agency?
- What is the nature of autonomy and how is it related to agency and responsibility?

Work on these questions, while more or less having an uneasy home base in the world of moral philosophy, draws from a diverse range of cross-disciplinary sources, including moral psychology, psychology proper (experimental, developmental, etc.), philosophy of psychology, communicative disorders, philosophy of law, legal theory, metaphysics, neuroscience, neuroethics, political philosophy, and more. But it is nevertheless unified by its focus on who we are as deliberators and (inter)actors, embodied practical agents negotiating (sometimes unsuccessfully) a world of moral and legal norms. It is an attempt to understand both what it takes to be moral and prudential beings—agents who reason about and act (or fail to act) on various practical norms—as well as what it takes to be responsible beings—agents who are responded to in distinctive ways in virtue of their reasoning about and acting on such norms. We are engaged, then, to borrow a phrase from Harry Frankfurt, in the project of "philosophical anthropology" (Frankfurt 1999, p. ix). This project thus doesn't fit squarely into any of the standard subdivisions of moral philosophy—applied ethics, normative ethics, and metaethics—because in important respects it is presupposed by, or provides the background conditions for, all of them.

The new papers collected herein were drawn from presentations at the first New Orleans Workshop on Agency and Responsibility (NOWAR) in November 2011, and they all grapple with one or more of our questions. The first half of the papers are roughly about "Agency" and the second half are roughly about "Responsibility," although given the interrelated nature of the topics, several papers take up questions on both sides of the agency/ responsibility map, so the division is somewhat artificial.

Sarah Buss's paper, "The Possibility of Action as the Impossibility of Certain Forms of Self-Alienation," kicked off the workshop and provides our jump start as well. Buss is interested in the general conditions of rational agency, and she reveals some serious problems with both the standard instrumentalist account and the two alternative, constitutivist accounts. She begins with a simple truth: to act for reasons includes both taking various features of our circumstances as reasons to do certain things and being motivated to do certain things by various considerations. But the relation between these justifying and motivating senses of acting for reasons is extremely complicated, and ultimately quite puzzling. As part of her approach, she presents a stratagem: if I am to govern myself, I must render my motivations authoritative with respect to self-determination, so that my actions are genuinely attributable to *me*, the ruling self. But if my ruling self consists in reason, and my to-be-controlled motivational part consists in my indifferent-to-reason desires, then how could it be sensible to make demands of desire *based on reasons*? One kind of purported solution restricts reason to a merely instrumental role in aiding our motivational parts' execution, but this will not work insofar as we can question the desirability of satisfying any of our desires. The other kind of purported solution attempts to find a way to give reason a greater role in determining the goals we pursue, so that it is already, in a way, built into the motivational part. But this kind of solution is problematic, argues Buss, insofar as its advocates do not build *enough* into the motivational part. We are left with a real challenge to agency theorists: how might we understand how the goal-directed behavior of creatures like us is transformed into behavior *we cause* in pursuing those goals? Buss urges that this challenge cannot be met absent metaethical vexation about the nature of reasons.

Michael Bratman's essay, "The Fecundity of Planning Agency," may be construed as an attempt to answer this challenge (although that is not how it is explicitly framed). As agents, we are capable of (1) acting—and organizing our actions—across time; (2) joint action (or shared intentionality); and (3) self-governance. A theory of rational agency will gain in plausibility to the extent that it can provide a unified account of these three remarkable capacities. Bratman's conjecture, fleshed out in insightful ways here, is that these three capacities are grounded in our core capacity for planning agency. The issue engaging most directly with Buss is that of self-governance. Bratman attempts to avoid worrisome homuncular models of the self by locating it (or at least locating its representative voice) in the agent's plan-like commitments to what matters (such that they carry weight in our deliberative thought). As Bratman puts it, "[W]hen these plan states guide, the agent governs." And such commitments obviously build on the agent's planning agency. The question, then, is to what extent locating the agent (or the agent's voice) in such commitments can avoid the worries articulated by Buss.

Luca Ferrero is also interested in an important aspect of attributability, of what is involved in an attitude's being *mine*. In particular, Ferrero is interested in one sort of attitude—intention—which is commonly thought to take as its proper object only the intender's own action. Call this the "own action" condition. Ferrero's aim in "Can I Only Intend My Own Actions? Intentions and the Own Action Condition," is to show that the "own action" condition is false. The relevant aspects of agency deemed necessary to intention may actually be met without restricting the *objects* of the agent's intention to her own actions. He arrives at this conclusion by first considering "aimings," a simpler kind of practical attitude, and he attempts to show that the own action condition isn't necessary to aimings, and to the extent that intentions are not relevantly different in key aspects, the own action condition isn't necessary to them either. This conclusion bears significance for several skirmishes in the field of agency, including how to properly characterize the relation between intentions and actions, the fate of the causal action theory, and the nature of joint, or shared, intention and agency. This last, of course, is one of the three agential capacities ostensibly unified by Bratman's more basic theory of planning agency.

In "Regret, Agency, and Error," Daniel Jacobson focuses on a different psychological aspect of the attributability of actions to agents, namely, what Bernard Williams called "agent-regret." For Williams, regret is the general emotional response one has to the thought, "How much better if it had been otherwise," and agent-regret is this sort of response with respect to one's own actions and their consequences, i.e. a response to the thought, "How much better if *I* had done otherwise." This construal is representative of the view Jacobson calls a cognitivist theory of the emotions, and one of the aims of his rich paper is to cast doubt on this theory. He does so, in part, by pointing out ways in which a cognitivist theory like Williams's (and Stocker's) conflates distinct sentiments (e.g. regret and guilt) under the same constitutive thought and also yields some allegedly distinct specific emotions that have neither coherent expression in action nor any clear connection to motivation. (And for the important distinction between emotions and sentiments, see the paper.) Another commitment of the cognitivist theory is that certain emotions are *irrational* in virtue of the falsity of their constitutive thought. This is thought to be illustrated by Williams's famous lorry driver case, in which a faultless driver kills a child. He should feel *something* bad, we think, and Williams posits agent-regret as the emotion it is rational for him to feel, given that he can think, "How much better if I had done otherwise" about "his" action, despite its being a complete accident. In response to this stance, Jacobson argues persuasively that the regret of the lorry driver may well be irrational, given that he did in fact do nothing wrong. Nevertheless, his regret may still be *admirable*: we would expect such a response in the virtuous. We might, then, view the rationality of emotions in terms of their fittingness to the properties of their objects (and not responses to evaluative judgments), and then preserve our general sentimental categories of regret and guilt without proposing special emotions in order to account for Williams's special cases.

The thought, "Ahh, to have been able to do otherwise!" will be resonant to those familiar with the long-standing dispute over the relation between free will and determinism. Taking up a lot of room in this dispute has been the wrangling over the Principle of Alternate Possibilities, or PAP, which states that one is morally responsible for some action A only if one could have done other than A (Frankfurt, 1988, 1). While many theorists continue to think PAP is true, they strenuously disagree over the interpretation of the nature of such an ability and, relatedly, what reason there may be to think we have it. In taking up both issues, many libertarians (those who believe that we are free agents in a way that is incompatible with determinism) think phenomenological considerations favor their view. In particular, they say, we experience ourselves from the inside as having the ability to do otherwise when engaged in moral deliberation and/or action, an ability that is incompatible with determinism and so provides some evidence for our incompatibilist freedom. Compatibilists, in return, may grant that we have some such experience, but deny that it implicates any sort of incompatibilist ability; rather, they say, the experience we have of being able to do otherwise is simply of being able to do otherwise *if the circumstances had been different*. But if this is all my experience is of, then it merely implicates an ability that is perfectly compatible with determinism.

Oisín Deery, Matt Bedke, and Shaun Nichols take this to be an empirical dispute and so ripe for the methods of experimental psychology. In their paper "Phenomenal Abilities: Incompatibilism and the Experience of Agency," they discuss a series of experiments they performed investigating how people actually regard what is going on in their experiences of moral deliberation and decision-making. As it turns out, they found remarkably consistent results: people across the board overwhelmingly (a) view themselves as having the ability to do otherwise in a variety of practical (including moral) circumstances, and (b) regard this ability explicitly in incompatibilist terms. When testing various compatibilist reinterpretations, the authors found that they just weren't what subjects meant by their experienced ability to do otherwise, an experience that remained firmly incompatibilist. This paper is a model of experimental philosophy, for it identifies an empirical matter on which a certain aspect of a long-standing philosophical dispute hangs and subjects it to scientific scrutiny. It then discusses the findings in distinctly (and insightful) philosophical terms. If we accept the authors' findings, we must move on from wrangling over the interpretation of our phenomenal ability to do otherwise to the role the phenomenology ostensibly plays in support of libertarianism: is it really *evidence* for the view?

Alternatively, perhaps the ability to do otherwise *isn't* necessary for freedom at all. This is the course taken by many compatibilists who have become convinced by the so-called "Frankfurt cases," scenarios in which a counterfactual intervener stands at the ready to ensure that some agent decides as he (the intervener) wants her to (and so removes her ability to do/decide otherwise) but never does a thing, as she decides all on her own to do the hoped-for action. But if PAP is false, what *does* the sort of control implicit in freedom and responsibility require? One popular move has been to view free agents as those who are reasons-responsive. The problem with doing so is that it looks as if agents in Frankfurt cases actually *aren't* reasons-responsive, as they wouldn't have been sensitive to the reasons not to perform the hoped-for action. In responding to this worry, Fischer and Ravizza (1998) have argued that their favored notion of guidance control over some action consists in its issuing from the agent's own reasons-responsive *mechanism*, something which can be sensitive to reasons to do otherwise in Frankfurt cases, even if the agent herself isn't.

Michael McKenna, in his intricate paper, "Reasons-Responsiveness, Agents and Mechanisms," traces this dialectic with great care before discussing some serious problems with a mechanism-based approach to the nature of compatibilist control, in particular the problem of individuating the specific mechanisms operative in cases of free action. This problem, McKenna argues, is really symptomatic of a deeper, structural problem with mechanistic theories generally: if (as seems plausible) we allow that various sub-mechanisms clearly interact with one another to issue in different types of action, we seem driven to attribute actions to "the mechanism" that can somehow include all such simultaneously-operating sub-mechanisms, but then this mechanism looks an awful lot like a full-fledged *agent* once more. So if we are forced back to an agent-based reasons-responsive view, are we also back to accepting PAP (given that it seems these types of views give the wrong answer in Frankfurt cases)? No, argues McKenna: it is possible for an agent to be reasons-responsive in the Frankfurt case without being able to *react* to reasons to do otherwise when given sufficient reason to do so. Developing and defending this intriguing proposal takes up the remainder of his essay.

Turn, then, from a focus on agency to a focus on responsibility. In recent years, most discussions of moral responsibility jump off from P. F. Strawson's deeply influential "Freedom and Resentment" (Strawson 1962), and ours is no different. In "Responsibility, Naturalism and 'the Morality System'," Paul Russell offers a thoughtful discussion of Strawsonian naturalism about moral responsibility, as part of his defense against construing both Strawson and the general notion of moral responsibility too narrowly, constrained by and within "the moral system" and so defined by concepts like obligation and wrongness. One worry about this construal is that it has us dealing with positive and negative actions in an asymmetrical fashion, for it views our responsibility-assessments-via-reactive-emotions as roused only by violations of norms, but this is a one-eyed view of our rich emotional and interpersonal lives. Another worry is that this approach is too locally biased, a responsibility conception steeped exclusively in modern, Western, Christian culture, implausibly marking us off conceptually from the ancient Greeks, say, and other shame-based cultures, whose activities we must now regard as non-responsible, as alien to our own sort of ethical lives. But this has the cost of removing any apparatus enabling us to criticize and *engage* with those cultures, which we seem quite able to do. The third and most serious worry, however, is that construing responsibility so narrowly, as so culturally located, generates the possibility that our responsibility practices need an external justification, a possibility that opens the door again to deep skepticism about responsibility generally. Russell, in the remainder of his essay, tries to close this door by offering a middle-path naturalism about our ethical reactive attitudes—motivated by Bernard Williams's critique of "the morality system"—one that charts a path between Strawson's own overly wide approach (given his inclusion of too many emotional reactions as ostensibly relevant to moral responsibility, e.g. love and friendly feeling) and R. Jay Wallace's overly narrow one (which distorts our understanding of human ethical life and looks conceptually imperialistic).

One of the many innovations of Strawson's approach was to view being morally responsible as somehow a function of being *regarded* as morally responsible, where to be regarded as such is just to be the target of various sorts of reactive emotions (although there are significantly different ways to interpret this view; see Brink and Nelkin's paper for a realist—in contrast to a response-dependent—understanding of Strawsonian responsibility). Strawson divided the sorts of pleas that get one off the hook as a target of such reactions into two groups, what Watson (1987) has helpfully called "excusing" and "exempting" pleas. Excusing pleas—"It was an accident"; "I didn't know"—are those that get the agent off the hook for a particular action but continue to leave her vulnerable to reactive emotions generally. Exempting pleas—"She's only a child"; "He's a hopeless schizophrenic"—remove the offender from the community of morally responsible agents altogether. What Strawsonian theorists have searched for, then, are ways of theoretically unifying these categories. In other words, is there any unified grounding for the inappropriateness of resentment for accidents, ignorance, insanity, and the like? Some have thought there is theoretical unity given the sanctioning aspect of (negative) reactive attitudes, where the unity conditions are grounded in the *fairness* of such sanctions. Others have sought the appropriateness conditions in the expressive nature of the reactive emotions (such that some emotions are unfitting in virtue of their failure to meet the felicity conditions of certain forms of communication). Others have found unifying ground by focusing on how the reactive emotions target certain qualities of will.

If Zac Cogley, in "The Three-Fold Significance of the Blaming Emotions," is right, this search for a unified treatment may be quixotic. Cogley first draws from recent psychological research to articulate three distinct functions of the (negative) reactive emotions: appraisal, communication, and sanction. Second, Cogley shows that each of these three functions imports its own distinct set of appropriateness conditions: while a blaming emotion may meet the conditions rendering it an accurate or fitting appraisal of responsibility, say, it may not meet the conditions rendering it a fair or deserved sanction. This point leads to a third: different (unifying) theories of responsibility are typically motivated by a focus on different functions of the blaming emotions. This fact of course could explain their disagreements about the nature of moral responsibility, but it may also help explain the long-standing dispute between compatibilists and incompatibilists: if compatibilists focus on appraisal and/or communicative functions of blame, whereas incompatibilists focus on the sanctioning functions (and each has different appropriateness conditions), it is no wonder they wind up with radically different metaphysical views. Cogley's sort of analysis provides a tantalizing glimpse of progress in this well-entrenched debate.

Turn, then, from the general significance of blame to its specific appropriateness conditions. As already mentioned, one of the assumed excusing conditions is ignorance. This assumption is widely held across all three sorts of theories of moral responsibility. In particular, the thought is that if one is *morally* ignorant—non-culpably ignorant of moral principles rendering what one is doing is morally wrong—then it would be inappropriate to blame one, regardless of the differential functions of the blaming emotions. Matt Talbert, in "Unwitting Wrongdoers and the Role of Moral Disagreement in Blame," argues strenuously against this assumption. For Talbert, to blame someone is to have certain attitudes responding to features of the blamed agent's actions/attitudes that are objectionable *from the perspective of the blamer*, and often the features we find blameworthy are judgments that our interests don't count (or don't count sufficiently) to the blamed agent. But these features may well be present in those who act from non-culpable ignorance, and so unwitting wrong-doers may well be blameworthy. What Talbert's move does is undercut a recent kind of skepticism about moral responsibility premised on the assumption about the excusing status of moral ignorance. It also reveals the importance and role of moral disagreement: what actually matters in our moral disputes is our judgment about the significance of the various interests and needs being affected by the offending agent, and not whether this agent knows (or it was reasonable for her to know) the moral principles rendering her actions wrong.

Tamler Sommers, in "Partial Desert," also focuses on features external to blamed agents that may nevertheless ground blame. While he admits that facts about the agent's knowledge and ability to control her actions specify a range of appropriate responses, what Sommers adds to the mix is the thought that the *victim's* responses are relevant to narrowing down reactions within that desert range. To make the view more intuitively appealing, Sommers draws on the moral luck literature in describing two cases. In both, a drunk driver kills a girl who happens to run out in the road, then keeps on driving. In the first variation, the state catches him and puts him to death by firing squad (it's Utah). In the second, he's tracked down by the father of the girl, who shoots him. In both cases, the driver dies, but only in the second might we think he *deserved* his fate. This reaction, Sommers argues, reveals our commitment to there being facts external to the punished agent and what he knew or did that are relevant to assessments of desert. In particular, it matters whether the offended parties carry out (or at least contribute to a determination of) the punishment. His defense of this proposal carries him into interesting territory, including an important exploration of whether the proportionality required for just legal sentencing and moral sanctioning must be ordinal as well as cardinal, i.e. an exploration of whether like cases really must be treated alike.

While Sommers exclusively focuses on *moral* desert and responsibility, some have thought that we can achieve real insight into features of moral responsibility by thinking about *criminal* responsibility. There are, for instance, several explicitly articulated aspects of criminal responsibility—excuse, justification, *mens rea*, fault, and liability—that seem to play an analogous and important role in moral responsibility too (for doubts about this ostensible relation, however, see Shoemaker Forthcoming). Susan Wolf, for example, has famously appealed to the legal notion of *insanity* to ground a normative competence condition for moral responsibility. She discusses the case of JoJo, the son of an evil dictator, who soaks up his daddy's values and, when grown, treats the peasants in the same horrible ways as dad. Wolf's thought is that, insofar as JoJo isn't viewed as morally responsible (a dubious finding, however; see Faraci and Shoemaker 2010), it is because his deep self (his reflecting, valuing self) isn't normatively sane, i.e. he is unable to recognize the correct moral values. Insofar as previous compatibilist theories hadn't taken this fundamental orientation to the good as necessary to moral responsibility, they were incomplete.

Heidi Maibom, in "Values, Sanity, and Responsibility," sees the Wolfian position as dilemmatic: either this understanding of insanity is supposed to capture the legal notion or it is not. If it is supposed to do so, then it overreaches, insofar as legal insanity typically incorporates only those suffering from delusions and hallucinations, and not those who merely have different values than ours; indeed, the legally insane usually completely *share* our values. If, on the other hand, the Wolfian view advances a different understanding of moral insanity, then it is wrong-headed, as it would require us to cease regarding as responsible those we clearly do, namely, those who are merely intransigent in holding different values than ours. As long as there is an available inferential route for them from their own values to beliefs about wrongness we can share, they are eligible for responsibility assessments for failing to do so. If we didn't allow for this possibility, then we would have to limit our responsibility assessments only to those who acted wrongly-while-fully-aware-what-they-were-doing-was-wrong, but this is to implausibly restrict our responsibility community.

In the final selection of the volume, "Fairness and the Architecture of Responsibility," David Brink and Dana Nelkin further explore and draw from the relation between criminal and moral responsibility to provide a theory of the framework of general responsibility, a conception that is unified, on their view, by its commitment to the target's having a fair opportunity to avoid wrong-doing. On Brink's and Nelkin's picture, drawing from both criminal and moral responsibility enables us to avoid some unfortunate elisions in each realm. For example, moral theorists often overlook the importance of situational factors in determinations of responsibility, something which legal theorists tout. Even when agents are normatively competent, there may be features of the situation—e.g. involving coercion or duress—that undermine their ability to avoid wrong-doing. But there are also features emphasized by moral theorists that legal theorists sometimes overlook or wrongly discount, for example, the importance of volitional impairments (e.g. paralyzing fear or depression) as providing excuses from responsibility. On their overall model, cognitive and volitional elements are of equal importance in enabling people to have the fair opportunity to avoid the sort of wrong-doing to which we typically respond with moral blame and, in the criminal realm, legal sanctions. And because the various elements in their architectonic are scalar, there is a coherent notion of *partial* responsibility ultimately implied by their view, and they offer interesting suggestions for how this notion may be applied and dealt with in the criminal realm.

In each case above, I have only skimmed the surface of these rich essays. But what I have said should be sufficient to reveal both the variety of topics falling under the rubric of "agency and responsibility" as well as their interesting and variegated relations to one another. While most writing on responsibility focuses solely on the negative, blaming, reactions to the deeds of others, my reaction to the contributions of the authors herein is nothing but philosophical and personal gratitude.[1]

[1] My thanks to all of the authors, both for their contributions to this volume and for their comments on an earlier draft of these introductory remarks. I am also deeply grateful to Steve Sheffrin and the Murphy Institute at Tulane University for financing the first workshop on agency and responsibility in November 2011, and to Nathan Stout, Jason Tyndal, and Frankie Worrell for compiling the index. Finally, thanks to Peter Momtchiloff at Oxford University Press for his early and enthusiastic support of this new series.

REFERENCES

Faraci, David, and Shoemaker, David. (2010). "Insanity, Deep Selves, and Moral Responsibility: The Case of JoJo." *Review of Philosophy & Psychology* 1: 319–32.

Fischer, John Martin, and Ravizza, Mark. (1998). *Responsibility and Control.* (Cambridge: Cambridge University Press).

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

—— (1999). *Necessity, Volition, and Love.* (Cambridge: Cambridge University Press).

Shoemaker, David. Forthcoming. "On Criminal and Moral Responsibility." In Mark Timmons, ed., *Oxford Studies in Normative Ethics, Volume 3* (Oxford: Oxford University Press).

Strawson, P. F. (1962). "Freedom and Resentment." *Proceedings of the British Academy* 48: 1–25.

Watson, Gary. (1987). "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In Ferdinand Schoeman, ed., *Responsibility, Character, and the Emotions: New Essays in Moral Psychology* (Cambridge: Cambridge University Press, 1987): 256–86.

1

The Possibility of Action as the Impossibility of Certain Forms of Self-Alienation

Sarah Buss

I begin with an extremely simple observation. Whenever we do things for reasons, we do what we do because we regard certain features of our circumstances as reasons to behave in certain ways and not others, and because we are motivated to behave in certain ways and not others. But what is the relation between these two elements of acting for a reason? What is the relation between our capacity to respond to reasons and our practical impulses? In trying to answer this question, we discover that what seems so very simple is really quite complicated.

The complications that interest me here arise from the fact that agents who do things for reasons ("rational agents") can reflect on their own psychological states, and in so doing, relate to these states as features of their circumstances. This means that they can relate to their own impulses to do things (their inclinations, desires, and whatever other conditions move them to initiate change) as features of their circumstances. And this means that they can be alienated from themselves in a way that threatens their capacity to act. An instance of behavior constitutes an agent's action only if the agent herself is its direct cause. But if it is possible for an agent to dissociate herself from, and even repudiate, her own practical impulses, then it is possible for these impulses to cause her to behave in certain ways without this behavior being attributable to her.

Given this possibility, an account of what is required in order for rational beings to act must differ significantly from an account of what is required in order for nonrational beings to act. Precisely because the latter sort of beings cannot be alienated from their own psychic states in the way I have just indicated, we do not need to tell a story about why behavior caused by various elements of their psyche qualifies as their actions. But if a being can be self-alienated in the relevant way, then the fact that she engages in instrumentally rational, goal-directed behavior does not suffice to ensure that the goals she pursues are her own. The fact that she behaves in a way that *would* qualify as an action *if* she *were not* a rational being is compatible with the fact that she is alienated from her behavior; and so, it is compatible with the fact that her behavior does not qualify as her action; and so, an adequate account of what is required in order for her to act will have to be an account of the conditions under which she is not alienated from herself in the relevant way.[1]

In the first half of this paper, I will press the case against the assumption that rational beings succeed in acting as long as they engage in instrumentally

[1] The point is not, as Jennifer Hornsby puts it, that we are under "pressure to add an extra ingredient in order to reveal [a rational agent] as a more or less reasonable, conscious being" (Hornsby 2004: 185). The point, rather, is that we are under pressure to identify the extra ingredient that is present when a rational being does not dissociate herself from her motives. It is perhaps worth noting at the outset some of the ways in which my critique of the standard picture of action is at once similar to and different from the sort of challenges that target what Hornsby calls "the naturalistic assumptions" at the heart of this picture. Like Hornsby, I will be arguing that there is a form of self-alienation which is a threat to even less-than-"robust" agency. More particularly, I share the concern that insofar as one relates to oneself as a feature of the world, one is a bystander to one's own motives, and so one is not the direct cause of anything that is directly caused by these motives; i.e. their causal efficacy is not an instance of one's own agency. Whereas, however, Hornsby and others who challenge so-called "causal" accounts of action think it is a mistake to characterize bodily actions as events caused by agents, I do not challenge this assumption. (This assumption, it is important to stress, is *not* the assumption that "an action . . . is something on the physical side of a supposed mental/physical divide and called a bodily movement."(180) Rather, the claim is that *bodily* actions are bodily movements *insofar as they have* a certain *non*physical status—which includes, but (as the requirement of instrumental rationality makes clear) is not exhausted by, their status as the effects of mental causes. Note that, Hornsby's suggestion to the contrary notwithstanding, the same story applies to actions that consist of an agent's intentionally *refraining* from moving her body in various ways.)

In addition to these respects in which my criticism resembles and differs from the typical challenges to causal theories of action, there are others. Whereas Hornsby thinks that the problems with the standard conception reveal the misguidedness of any attempt to account for action in terms of the "mental states" of the agent, I think that we can offer a diagnosis of what has gone wrong without ruling out the possibility of such a reduction. In other words, it seems to me that taking an "external perspective" on oneself can "pose a genuine threat to our agency," even if our actions can be "unproblematically accommodated in the naturalistic explanatory order" (181). (This is true, even if, as I suggest, to do something for a reason, one must take oneself to be responsive to reasons, and even if it is a mistake to offer a "naturalistic" account of reasons.) Finally, as I hope to make clear, one can claim that "X's desiring something and believing something [can be] translated into talk of items with causal potential" without claiming that "X's having a reason is . . . a matter of the existence of a pair of states" (180). Indeed, as far as I can tell, very few advocates of the standard conception endorse the latter claim.

rational, goal-directed behavior. I will challenge this assumption by calling attention to the problem posed by the capacity for self-alienation. In so doing, I will side with those who insist that an adequate account of action (not just "robust," or "full-blooded," or "self-governing" action, but whatever it is that suffices to qualify a person's behavior as voluntary) will be an account of what is distinct about someone who "identifies with," or "endorses," or "shares" the goals that her desires (and other psychic conditions) move her to pursue.

Whatever form such an account will take, it must explain how (i) a person's assumptions (conscious or unconscious) about what she has reason to do and (ii) her nonrational impulses (whatever inclinations, desires, etc. are neither constitutive of her rationality nor the product of her reasoning) can be sufficiently unified, or integrated, to constitute two aspects of a single point of view on her behavior. Unless an account of action accomplishes this task, it will not explain why a rational being's behavior can be attributed to *her*, the one who reflects on her impulses and wonders whether she has reason to endorse them. In the paper's second half, I will consider two proposals regarding how this integration might be affected. Both accounts fail, I will argue, and for the same reason: they assume that an agent's reason and her nonrational practical impulses play entirely different roles in the production of action; and they thus lack the resources necessary for explaining how these elements interact to constitute an agent. When the advocates of these accounts refer to "the agent's" exercise of authority over *herself*, or "the agent's" understanding of what she *herself* is inclined to do, they help themselves to a self-reflexive relation they cannot make sense of in their own terms.

Rational action can be an agent's way of "constituting herself," I will argue, only if the relationship between a rational being's reason and her nonrational desires to bring about various changes is a partly internal relationship—only if, that is, these two parts of the soul are already partially integrated. In order for desires to figure among the constituents of a "self" that is committed to responding to reasons, desiring must render the desirer vulnerable to the commitments and demands she makes in her rational capacity. *Qua* subject of desires, she cannot be indifferent to her normative assessments, even if her desires cannot be attributed to these assessments.[2]

[2] This is a point Aristotle makes in the *Nicomachean Ethics*: "In the soul too there is something contrary to the rational principle, resisting and opposing it. . . . Now even this seems to have a share in a rational principle, as we said; at any rate in the continent man it obeys the rational principle—and presumably in the temperate and brave man it is still more obedient; for in him it speaks on all matters, with the same voice as the rational principle. Therefore the irrational element also appears to be twofold. For the vegetative element in no way shares in a rational principle, but the appetitive, and in general the desiring element in a sense shares in it, in so far as it listens to and obeys it; this is the

It follows, I will argue, that rational beings cannot engage in goal-directed behavior without assuming that they are responding to mind-independent reasons to behave in this way. They must make this assumption because the self-integration necessary for them to initiate their behavior requires them to regard themselves as responding to the same sort of thing—reasons—in both their capacity as rational beings and their capacity as beings with substantive goals. If a person's nonrational impulses do not render her open to the demands of her reason, then insofar as she is the subject of these demands, these impulses are just so many features of her circumstances. As such, they pose no more threat to her integrity than anything else she must "take into account" in deciding what to do. Accordingly, she has no need to integrate them into herself in order to act.

THE STANDARD ACCOUNT: ACTION AS INSTRUMENTALLY RATIONAL GOAL-DIRECTED BEHAVIOR

Most people who think about what is involved in doing something voluntarily assume that an adequate account of action will apply equally to (i) those, like squirrels, to whom it is a mistake to attribute any opinions about what they have reason to do and (ii) healthy, mature human beings who form such opinions quite frequently, even though they rarely do so self-consciously. According to this widespread assumption, when it comes to plain old acting, a rational being need not satisfy any conditions other than those that suffice for the action of a nonrational being.

But this assumption is false. For it overlooks the fact that the capacity to reason is the capacity to dissociate oneself from the very sort of psychic forces whose causal role would yield an action if they were attributable to a nonrational being. Whereas having a behavior-determining impulse to bury a nut is all that it takes for a squirrel to have the goal of burying a nut, a rational being need not share the goal of her desire at a given moment in time, no matter how powerful this desire may be, no matter how nicely it

sense in which we speak of 'taking account' of one's father or one's friends, not that in which we speak of 'accounting' for a mathematical property. That the irrational element is in some sense persuaded by a rational principle is indicated also by the giving of advice and by all reproof and exhortation. And if this element also must be said to have a rational principle, that which has a rational principle (as well as that which has not) will be twofold, one subdivision having it in the strict sense and in itself, and the other having a tendency to obey as one does one's father." (Aristotle, *Nicomachean Ethics*, Bk. 1, ch. 12, 1102b25–1103a3) may cohere with her other desires and commitments, and no matter how confident she may be that some considerations do count in favor of the goal-directed behavior. Any conception of agency that suggests otherwise is thus inadequate. It is inadequate because it implies that a form of self-alienation is compatible with agency, even though in relating to one's behavior in this way, one would be no less passive than one is when one suffers a blow.

So, at any rate, I hope to show. I will argue, first, that it is not possible to act if one dissociates oneself from the motivating force of whatever psychic condition directly causes one's behavior. I will then point out what this implies: in order for a rational being to act, it is not enough that she engages in instrumentally rational, goal-directed behavior.

Let me begin by calling attention to a metaphor. Most philosophers seem to agree that the paradigm case of an agent who is not responsible for what she does is someone who is, as Harry Frankfurt puts it in characterizing the "unwilling" addict, a "passive bystander" to her own motives (Frankfurt 1988). But the metaphor of the "passive bystander" reveals a problem: if there is anything that being an agent is *not*, being an agent is *not* being a "passive bystander" to the direct mental causes of one's behavior. Insofar as one is a passive bystander to an event, this event is not one's own doing. It is not one's own action.

Of course, it is possible for me to reflect on my actions even as I am engaged in them: here I am, typing away at my keyboard, even as I am observing the movements of my own fingers. But, as this example makes clear, the point of the metaphor is not simply that we can take a detached stance toward our own actions, but that we can be *opposed* to what we ourselves do, even as we are doing it for a reason. Indeed, there is more to it than this. Even if we very much wished that we had better alternatives, and even if we keenly appreciated what compelling reasons we have to act otherwise, we would not necessarily be opposed to our behavior in the relevant sense. The point of the metaphor is that—with eyes wide open, as they say—we can prefer to do one thing *under the circumstances*, even as we are intentionally doing another. We can do this, moreover, in cases where the action is not merely "reflexive"—cases, that is, in which what we do is not a spontaneous response to a sudden change in our circumstances.[3]

Since, as Frankfurt reminds us, it would be possible to satisfy the stipulated conditions only if it were possible for the forces that move us to render us passive bystanders to their motivating power, and since insofar

[3] For an example of a reflexive movement, think of a case in which one quickly moves one's hand in front of one's face when a baseball suddenly appears in close proximity to one's head.

as someone is a passive bystander to an event, she is not the agent of this event, the metaphor reveals that no such self-alienated intentional action is conceptually possible. The behavior of someone who was alienated from her motives in this way would not be caused by *her*—the one who is committed to responding to reasons. It would not be attributable to her even if, as is surely the case, she took herself to have *some* reason to behave this way. As we are to imagine the case, her assumption about what it makes sense for her to do cannot account for the force of the desire that moves her to do it. Accordingly—and this is the point of the metaphor—the force that "moves her" is a force that prevents her from moving herself. And so—and this is what the metaphor reveals—this force prevents her from acting.

The same point can be made in a way that anticipates the divided soul models I will discuss in the paper's second half. *Qua* rational being, Sarah's goal is to do what she takes herself to have sufficient reason to do; *qua* nonrational being, her goal is to do otherwise (though not, of course, under that description!). But no one can wittingly commit herself at one and the same time to realizing two incompatible goals. We can have conflicting desires, wishes, hopes, fantasies, etc. But we cannot regard something as our goal while wittingly committing ourselves to behaving in a way that will ensure that we do not achieve it. This is as impossible as walking off in two different directions at once.[4]

This image calls attention to the fact that the conceptual limitation on the possibility of self-alienation can be expressed in the functionalist terms popular among many theorists of action. As Jay Wallace notes, future-directed intentions can play their role in long-term planning only if they are in harmony with the agent's normative judgments. This is, roughly, because if I *believe* that it would be bad for me to do X in two weeks, then I *hope* I will *not* do X in two weeks. But if I hope I will not do X in two weeks, then I am open to being moved by any considerations *against* doing X in two weeks. But then I am not in the mental state definitive of the future-directed intention to do X. I am not in the state whose function it is to lead me *not* to be moved by the considerations (or at least *most* of the considerations) that count against my doing X (Wallace 2001: 8).

Like most other philosophers, Wallace thinks that a similar argument cannot be constructed for tying present-directed intentions to normative judgments (Wallace 2001: 8). But I don't see why. Why should the addition of a longer time interval make a difference? If I believe that I have overriding reason *not* to do X right now, then I am in a state whose

[4] For a discussion of the principle of instrumental rationality in which I argue that it is not possible to wittingly will to do what is incompatible with achieving one's end *under this description*, see Buss, "Norms of Rationality and the Superficial Unity of the Mind," Unpublished Manuscript.

function it is to lead me to hope that I will not do X right now; and so I am in a state whose function it is to lead me to be moved by the considerations *against* doing X right now—whichever such considerations I am right now aware of. Of course, I can form two incompatible intentions when I am unaware of doing so. But, again, just as it is impossible for me to walk in two different directions at once, so too, it is impossible for me to knowingly commit myself to walking in two directions at once. This means that I cannot knowingly commit myself to—at one and the same time—being moved by the considerations against doing X right now and not being so moved. And this means that I cannot commit myself to contravening my own contemporaneous normative judgment.

More carefully, I cannot wittingly commit myself to contravening my own contemporaneous normative judgment *insofar as I am committed to responding to reasons.* Under special circumstances I might appraise the desirability of a given course of action out of idle curiosity, with no intention of treating my own verdict as a guide to my action. In the ordinary case, however, I am committed to being so guided. So were I to intend to do X without altering my conviction that I have overriding reason *not* to do X, I would wittingly commit myself to doing what I am committed to *not* doing. And so, again, I would be making a commitment I cannot myself credit as such.

Note that in making this point, I am not endorsing the controversial thesis that "the will is (just) reason in its practical capacity." Though in the usual case, a person forms a judgment concerning what she has reason to do in response to her commitment to being guided by such judgments, it hardly follows that all her commitments are determined by her normative judgments. To the contrary, a person's action is always one of several possible actions that would have been compatible with her all-things-considered verdict concerning what she has sufficient reason to do. (Think, for example, of the countless possible variations on the theme of "going up stairs.") Every action thus involves an "exercise of will" that cannot be reduced to the practical employment of reason. In committing oneself to complying with a given normative verdict *in-this-way-rather-than-that*, one is doing something that cannot be fully explained by the normative verdict itself.[5] My point is simply that it is not possible to will to act *contrary* to one's contemporaneous normative judgment. Though the conceptual limits of willing are not determined by the conceptual limits of our *normative* judgments, when we impose a normative constraint on our will, we impose a *conceptual* constraint as well.[6]

[5] Buridan's ass choices are a special instance of this phenomenon.

[6] For an account of weakness of will that does justice to this relationship between practical reason and the will, see Buss 1997.

Interestingly, support for this thesis can be found in the testimony of people who are widely agreed to be the paradigm examples of agents overpowered by inclinations they wish they could resist: those suffering from obsessive-compulsive disorder (OCD). Consider, for example, the anguished testimony of a man who could not overcome his powerful compulsion to return, over and over, to the place where he thought he might have hit someone with his car:

> The pain is a terrible guilt, that I have committed an unthinkable, negligent act. At one level, I know this is ridiculous, but there's a terrible pain in my stomach telling me something quite different.... The anxious pain says to me, "*You Really Did Hit Someone.*" The attack is now in full control. Reality no longer has meaning. My sensory system is distorted. I have to get rid of the pain. Checking out this fantasy is the only way I know how.
>
> I start ruminating, "Maybe I did hit someone and didn't realize it.... Oh my God! I might have killed somebody! I have to go back and check." Checking is the only way to calm the anxiety. It brings me closer to truth somehow. I can't live with the thought that I actually may have killed someon—I have to check it out.
>
> My fantasies run wild. I desperately hope the jury will be merciful. I'm particularly concerned about whether my parents will be understanding. After all, I'm now a criminal. I must control the anxiety by checking it out. Did it really happen?
>
> I think to myself, "Rush to check it out. Get rid of the hurt by checking it out."
>
> I'm now close to hysteria because I honestly believe someone is lying in the brush bleeding to death. Yes,.. the pain makes me believe this. (Rapoport 1991: 24)

In short, even though there is an obvious sense in which someone responding to the characteristic obsessions and compulsions of OCD wishes that she did not behave as she did (and even though there is usually "a part of [her] psyche [that] keeps telling [her] that this checking behavior is ridiculous" (Rapoport 1991: 26)), there is an equally obvious sense in which, at the moment of action, she is convinced that she has no better alternative, under the circumstances. She believes that nothing is more important than behaving as she does; and her understanding of why it is important incorporates both the irrational "fantasies" and the very real pain that accompanies the fantasies. Most importantly for our purposes here, her experience as of being overpowered by her compulsion is not the experience of being passive in relation to her behavior. To be sure, she is passive in relation to the pain and anxiety. But her response to her agony is the response of someone who is absolutely (and not unreasonably, given the intensity of the pain, and given that nothing else is likely to stop it) convinced that she really has no viable alternative. Her compulsion dominates her by determining *her* point of view. Her problem is precisely that she can*not* dismiss her fantasies as alien to her own beliefs and concerns.[7]

[7] For a development of this account of how "compulsions" prevent agents from being accountable for their actions, see Buss 2012.

To see that reports of what it is like to be in the grip of compulsions call our attention to a conceptual limitation on self-alienation, it may help for us to focus our attention on how things change when we cross the border from the voluntary to the involuntary. Imagine, then, that someone is hanging from a high ledge. She really, really, really does not want to fall to the ground. Nonetheless, at some point, she lets go. What has happened? It could be that the growing pain and discomfort of resisting the pull of gravity, together with the dawning conviction that no one is going to show up to help, led her to give up, or give in—to decide (however unselfconsciously) to stop trying to resist. Alternatively, it could be that she never wavered in her commitment to hold on, but that eventually her finger muscles simply lost their strength. According to the first hypothesis, letting go was an intentional action. According to the second hypothesis, letting go was involuntary behavior.

Is there a third possible scenario? One that appeals to an irresistible desire to let go? So as to distinguish the "pull" of such a desire from the "pull" of gravity, let us consider a second case. In this case, someone really, really, really does not want to satisfy her powerful desire to take drugs. Nonetheless, at some point, she goes off in search of a dealer, purchases the desired substance, and very deliberately injects it into her veins. What has happened? The obvious hypothesis is that she gave up, or gave in: the pain and discomfort of resisting the pull of her desire was simply too great relative to the benefit. So it seemed to her, at any rate; and so she decided (however unselfconsciously) to stop resisting.

But isn't it possible that she never wavered in her commitment to resist the pull of the desire? No, it is not. If never wavering in her commitment to resisting were compatible with the fact that she was moved to satisfy the desire, then it would have to be possible for the desire, like gravity, to so tax her body that her limbs would no longer be capable of obeying her commands.[8] But if this were to occur, then the resulting bodily movements would be—like the relaxation of her fingers in the second version of the previous story—involuntary. Of course, the desire to take the drugs would itself play a key role in causing the bodily movements. But though, unlike gravity, a desire is a goal-directed state, according to the conditions we are here imagining, it is the *force* of the desire that explains why she does not do what she takes herself to have sufficient

[8] The image of "the will as a muscle" is misleading, at best. "Will power" is not *analogous* to the power we manifest when we hold on for dear life; it is not the power of "the will." Rather, it is the *very sort* of power we manifest when we hang on for dear life. The difference simply has to do with what is being resisted, and—as I will soon have occasion to stress—with what form it is possible for defeat to take.

reason to do. So, according to the stipulated conditions, her desire explains her behavior because it plays the same role that gravity plays in the earlier story—the role of a force the agent is trying to resist, a force that is external to her will, a force that is opposed to her agency. If, then, we are to imagine that she loses whatever physical power is necessary to resist this force, without thereby ceasing to be committed to resisting it, we must conclude that the resulting behavior is—like the behavior directly caused by muscle fatigue—involuntary.

It seems, then, that an addict cannot intentionally take drugs without giving up her effort to resist her desire to take them. More generally, our desires cannot directly overcome our efforts of resistance in the manner of an irresistible physical force; we cannot relate to our desires as if they were irresistible physical forces.[9] In order to play this role, desires would have to overwhelm our agency, even while—in their capacity as desires—they moved us to pursue a goal. To put the same point the other way around: in order for desires to play the role of gravity, they would have to be, at one and the same time, the sort of causes that prevent our behavior from being involuntary and the sort of causes that overpower our agency.[10]

This conclusion represents a challenge to any conception of agency according to which doing something for a reason is compatible with being alienated from one's behavior in the way I have here been discussing. This means that it is a challenge to the standard conception of agency. Again, the point is not that satisfying the conditions spelled out in the standard conception never suffices for acting. (For all I am claiming, a squirrel acts as long as it satisfies these conditions.) But nor is the point simply that there is a more "robust" form of action which requires the satisfaction of additional conditions. The point is that agents who are capable of being alienated from their own desires must satisfy additional conditions if they are to qualify as acting; and this is because they must satisfy whatever conditions ensure that they are not alienated from the psychic conditions that move them.

[9] This is a point that Gary Watson stresses in his more recent work. (See Watson 2004) There he rejects the model of unfree action I am criticizing here. For an alternative account of the distinction between self- and other-determined action that is key to assessments of moral responsibility, see Buss 2012.

[10] Note what this implies about cases in which our demands on ourselves are not met with obedience: the defiance necessarily takes the form of forcing a reappraisal of our options—even if (as in the case of akrasia) we are unpleasantly aware that this reappraisal is a mere rationalization.

In order to drive home the critique of the standard conception already implicit in the preceding discussion, let me quote at some length from Michael Smith's recent attempt to clarify this conception by explaining what distinguishes intentional actions from behavior that is caused by an internal wayward causal chain. According to Smith (who is here appealing to work by Christopher Peacocke), if a person is to act intentionally, then "the movement of [his] body mustn't just be caused by a suitable desire and belief pair, but . . . it must also be the case that, abstracting away from the causal role played by factors external to the agent's psychology, if the agent had had a range of desires and beliefs that differed ever-so-slightly in their contents from those he actually has, he would still have acted so as to realize the contents of these desires, given these beliefs" (Smith 2012: 399). "What [this] differential sensitivity requirement guarantees," Smith explains, "is that when an agent acts intentionally, he doesn't just try to realize the desires he actually has, given the means-end beliefs he actually has, but that he would have tried to realize his desires, given his means-end beliefs, in a range of nearby possible worlds in which he had desires and means-end beliefs that differ ever-so-slightly in their contents. . . . [The] . . . requirement thus suggests . . . that, for someone to act intentionally at all, she must not just possess but also exercise a certain capacity to be instrumentally rational." (Smith 2012: 398). With a nod to Hempel, Smith continues:

> There are three distinct psychological states that play a causal role whenever an agent acts . . . [C]ausal roles are played . . . by the agent's desire that the world be a certain way, and his belief that the thing done is a way of making the world that way, [and] by . . . the agent's possession and exercise of the capacity to be instrumentally rational [so that] he [can] put his desire and belief together in the way in which they need to be put together if they are to cause an action. . . . [I]f the agent doesn't do what she does because of her possession and exercise of her capacity to be instrumentally rational, then her desires and beliefs with ever-so-slightly different contents will not stand in the modally rich pattern of connections that they must stand in if they are to satisfy the differential sensitivity requirement. . . . (Smith 2012: 399)

Smith concludes that the standard story can do justice to the fact that "those who act 'intervene' between their desires and their beliefs and their bodily movements" (Smith 2012: 399). Accordingly, the story can make sense of the fact that agents are the cause of their actions. If, however, the preceding observations are correct, then Smith is mistaken: a goal pursued in an instrumentally rational way is not necessarily the agent's own goal. More carefully, if these two goals cannot come apart, this is because whenever someone does something intentionally, she satisfies whatever additional condition ensures that she is not alienated from her own goal-directed states.

To reinforce the point that Smith's account leaves out something important I want now to offer two little thought experiments:

[Thought Experiment 1] Imagine an artifact, built so that it can achieve two different goals. Consider, for example, a rather special clock. It can either tell the time or indicate what the time will be fifteen minutes later. It is, moreover, so exquisitely (and diabolically!) constructed that it can substitute the one goal for the other; and when it does so, the mechanism relevant to satisfying the alternative goal will be engaged. (In other words, nothing about the causal chain initiated by the one goal-seeking mechanism prevents the alternative causal mechanism from being engaged when the clock's goal changes. In other words, our clock is "differentially sensitive" to the change in its goals.) Having (let us imagine) stipulated that the clock's goal-seeking behavior is not attributable to wayward causal chains, we can now ask: Have we imagined an agent?

Perhaps we are inclined to answer "no" because our clock lacks a "point of view." So, let us give this already magnificent clock a point of view. Now, we could be rather stingy about this, and merely give it an awareness of the goal "built into" it by the manufacturer, and of what it is doing—and even what it must do—in order to achieve this goal. Would this be enough to make it an agent? More importantly for our purposes, would it be enough to make it the case that when its hands do what they do "in order to tell the time," it—the clock—is acting?

Rather than stopping to answer this question, I want to add something to our clock's point of view. I want to give it an interest in what it is doing, a concern about which goals it has reason to pursue. With this important addition, let us now suppose that it has concluded that it doesn't much like the goal imposed on it by the manufacturer. Perhaps it can see that there is quite a lot to be said for telling and foretelling the time. But, at least now and then, it believes it would be nice to use its hands for other purposes. It would be nice, for example, if its hands would enact a little choreography it has thought up during the many boring hours it has spent diligently marking time. Indeed, it believes that the reasons in favor of behaving this way now and then outweigh whatever reasons it has to devote itself wholeheartedly to the cause of accurately representing (or predicting) the time. Yet, let us also imagine, it cannot translate this belief into reality; it lacks the power to put its all-things-considered judgment into action. Should we insist that it nonetheless has the power to act? Or should we concede, instead, that when its hands move in order to indicate that the time is now five past midnight, and when this happens despite its conviction that it has overriding reason to do otherwise, these movements no more qualify as its actions than do the hand movements of an ordinary clock?

Of course, in the case as described, the clock's opinions about "its" behavior are so causally isolated from this behavior as to disqualify the behavior from being its own. But this is just the point. Nothing of significance would change, moreover, if many of its opinions did have an effect on its behavior. As long as its judgment regarding the all-things-considered desirability of the two-handed choreography was inefficacious, the regular movements of (what we can now safely call) *its* hands would not be *its doings.* They would not be its actions.

This situation resembles a case in which a person's body drags her around, or does other things, "against her will." Apparently, something like this does sometimes happen. When a person suffers from anarchic or alien hand syndrome, for example, one of her hands may undo the buttons of her shirt, without *her* having any intention of doing this.[11] So as to contrast this behavior with a mere spasm, we could refer to it as an "action." Since, however, the goal being pursued is not a goal that *anyone* "has in mind," it is not anyone's *intentional* action. And in any case—and this is what matters to me here—it is not an action attributable to the person herself. A person is passive in relation to what her anarchic fingers are doing to the buttons of her shirt—no less passive than she would be if these fingers were on somebody else's hand.[12]

[Thought Experiment 2] Consider a second goal-seeker. Call him A. A has no capacity whatsoever to reason. He has no capacity to ask himself whether anything is worth doing. Nonetheless, he has dispositions

[11] People suffering from anarchic or alien hand syndrome "lose the 'sense of agency' associated with the purposeful movement of the limb while retaining a sense of 'ownership' of the limb. They feel that they have no control over the movements of the 'alien' hand, but that, instead, the hand has the capability of acting . . . independent of their voluntary control. The hand effectively has a 'will of its own.' " "The alien hand is directed toward goals of which the patient is not consciously aware." ("Alien Hand Syndrome," *Wikipedia* <http://en.wikipedia.org/wiki/Alien_hand_syndrome>)

[12] "Sufferers of alien hand will often personify the rogue limb, for example believing it to be 'possessed' by some intelligent or alien spirit or an entity that they may name or identify." ("Alien Hand Syndrome," *Wikipedia* <http://en.wikipedia.org/wiki/Alien_hand_syndrome>)

to pursue various goals. And he also has the disposition to be guided by B's theoretical reasoning about what it would take in order to satisfy these goals. When B is aware of a goal A has, she engages her reason in the task of understanding what it would take to achieve this goal. And when she reaches a conclusion on this point, A—who also has a special capacity to read her mind!—adjusts his goal-pursuing accordingly.

By stipulation, this story contains two distinct individuals: one (A) disposed to pursue goals in an instrumentally rational way, despite having no opinion whatsoever about the desirability of doing so; the other (B) disposed to employ reason to determine which events are likely to give rise to which other events. Is A an agent? In contemplating this question, we can consider some others: Would it make a difference if B deliberately "fed" A the results of her reasoning—because (motivated, perhaps, by the interest in understanding things) she was curious to see what he would do? Would it matter if we implanted B inside A—so that they now shared a single body? Perhaps this would suffice to render *B* an agent—if we also gave her the higher-order goal of achieving A's goals. Note, however, that in this case, B's goal-seeking behavior would constitute an action only because B would endorse this behavior; the fact that it was instrumentally rational goal-seeking behavior would not suffice.

This point is even more obvious if we stipulate that B strongly repudiates the goals pursued by the body she now shares with A. Under these circumstances, who is the agent of the resulting behavior? Should we not concede that no one is, and that, accordingly, the behavior does not qualify as an action, despite satisfying Smith's conditions?

We cannot solve the problem revealed by these examples by simply introducing more complex causal feedback mechanisms. As long as, despite these greater complexities, the explanation thereby provided does not rule out the possibility that the behavior is unresponsive to the contemporaneous normative judgments of the rational being whose action it would otherwise unproblematically be, it will not qualify as an action. I thus conclude that we must reject the standard conception of agency.

Before turning to consider an alternative conception, I want to warn against reading too much into this conclusion. I am not claiming that acting always requires reviewing the considerations pro and con. I am not claiming that when we do things for reasons, we generally have a vivid, determinate conception of why what we are doing is really worth doing. I am sure that neither of these claims is true. Though, as Talbot Brewer eloquently reminds us, we can, and often do, make a point of pursuing certain activities with the aim of gaining an ever-greater appreciation of their value (Brewer 2009), for the most part, our actions involve relatively unreflective attempts to continue the activities we have already initiated, often without ever having given the matter much thought. As Hilary Bok points out in discussing the Milgram experiments, our capacity to act without regarding every change in our circumstances as a new occasion for deliberation and choice makes it possible for us to become accomplices to evil without endorsing any evil ends: if we don't pay attention—and, again, we often don't—continuing on the path we have set ourselves can lead us to a place we did not know we were headed (Bok 1996). Less dramatically, philosophers of action are surely right to stress the extent to which we are inclined to give ourselves permission to act "from habit"; and David Velleman and Peter Railton (among others) are right, too, to note that even instrumental reasoning is often superfluous: our habits of response include dispositions to adjust our means to our immediate ends without engaging in any more "reasoning" than occurs when the average backyard squirrel scampers from branch to branch (Velleman 2009 and Railton 2011).

We can concede each of these points, however, while insisting that even the most ho-hum instances of agency are incompatible with the sort of self-alienation I have here been exploring. Insofar as (however unconsciously) we—we reasoning beings—take a stance on what we are doing (insofar as we take ourselves to have sufficient *reason* for what we are doing), we *endorse* what we are doing—not in the sense that we think especially well of ourselves, or are especially pleased with the options we face—but in the sense that we—the ones who bother forming an opinion about whether what we are doing makes sense—accept the goals of our behavior as our own. An adequate account of rational agency will thus be an account of this self-relation, and of the normative judgment that is inseparable from it.[13]

AN ALTERNATIVE ACCOUNT: ACTING AS A WAY OF INTEGRATING CONTINGENT GOAL-DIRECTED MENTAL STATES WITH THE CAPACITY TO REASON

I am not alone in drawing this conclusion. Other philosophers have called attention to the fact that rational beings can intelligibly ask themselves

[13] There is a vast philosophical literature devoted to making sense of the fact that rational agents "identify with" their motives. Some philosophers appeal to higher-order desires. (For the classic defense of this approach, see Frankfurt 1971 and Frankfurt 1977.) Others appeal to evaluative judgments (for the classic defense of this approach, see Watson 1975. Others appeal to hierarchies and intentions, or other "plan-like" states. (See Bratman 2002.) By the end of this paper, I hope it will be clear why my sympathies lie with Watson's approach, even though I do not endorse his account.

whether they have sufficient reason to do what they are contingently inclined, committed, or otherwise disposed to do. Indeed, they have stressed that we can intelligibly ask whether we have sufficient reason to do what we *would be* contingently inclined to do, *if* we were aware of all the nonnormative, nonevaluative facts that are "relevant" to the satisfaction of our noninstrumental desires. Unless, these philosophers insist, our motives reflect our answer to this question, they lack *our* endorsement. And if our motives lack our endorsement, then *we* are not the cause of *their* effects; their effects could not be our *actions*—even if they qualified as instances of goal-directed behavior.[14]

To acknowledge this possibility is to acknowledge that reason is not simply "a slave of the passions." This is compatible with acknowledging that no one can reason without falling under the causal influence of nonrational motivating states. And it is compatible with acknowledging that, as a matter of contingent fact, some (most) rational agents believe they have reason to satisfy their (strongest, highest-order, most coherent) desires. The point is simply that rational agents can intelligibly ask themselves whether they have sufficient reason to satisfy any or all of their contemporaneous desires. And this means that they can be divided from themselves in a way that would not be possible if their reason were simply a slave of whatever psychic forces move them to pursue certain substantive goals.

How do self-reflective rational beings manage to achieve the necessary integration of their reason and their practical impulses if the former is not a slave of the latter? In the second half of this paper I want to consider two answers that have recently been offered by philosophers who take the problem of self-alienation seriously. According to the first account, we cannot make sense of rational action without replacing the alleged power relation between master and slave with an authority relation: a rational being constitutes herself when and only when her practical impulses obey the orders of her reason. According to the second account, the self-constituting relationship between these two parts of the soul is more like a partnership between two agents with two different, but ultimately compatible, agendas: a rational being constitutes herself when and only when her nonrational motivating states accommodate themselves to her defining interest as a reasoner. On both accounts, the integration of the two parts is accomplished when an intention is formed. On both accounts, there is

[14] It is important to distinguish this point from the claim that agents *must* reflect on ("foreground") *their desires.* The point is that they *can* reflect on—and call into question—what these desires represent as to-be-done. (For a defense of the thesis that desires are not typically among the things agents take into account, See Pettit and Smith 1990.

thus an important sense in which until there is an intention, there is no agent.[15]

In suggesting that determining one's action is the necessary means of constituting oneself, the alternative accounts share an assumption with the standard accounts they reject: they contrast the goals we have insofar as we are rational beings (e.g. avoiding incoherence, discovering causal relations,. . . .) with the substantive goals we have insofar as we are moved to initiate change. According to this division of labor, the roles played by each part of the soul do not give them a common interest; the person herself does not have a single interest insofar as she is thus divided. More carefully, the two parts do not have a common interest if, as the alternative accounts also assume, a person's reason is not limited, in its practical role, to figuring out how to satisfy her desires. I will try to show that this is precisely why we cannot accept the alternative accounts: they fail to explain how a soul thus divided can assemble itself into a whole.

In making my case, I will also be exploring the relationship between (i) the claim that acting involves unifying one's practical and rational parts and (ii) the claim that the conditions necessary for agency provide the criteria for what we have reason to do. Many philosophers have offered compelling objections to the second claim. Before I press my own objection, however, I want to stress how little one needs to assume in order for it to seem as though there is no plausible alternative. One need merely embrace cognitivism about practical reasons, while conceding both that (i) we can intelligibly wonder whether we have reason to satisfy any of our desires, even when we understand the likely effects of so doing, and that (ii) there is no mind-independent substantive criterion for what counts as a reason for what. Given (i), it seems that we must locate the criterion for practical reasons in a commitment or goal that does not depend on which (contingent) substantive desires a prospective agent happens to have. But given (ii), it seems that the criterion must be grounded in one or more of the prospective agent's commitments or goals. Is there, then, any alternative to acknowledging that what someone has reason to do is determined by whatever goal is inseparable from her identity as a prospective agent? And isn't this just the goal of not being a passive bystander to one's behavior? Isn't it the goal of "constituting oneself" as an agent?

In short, from rather plausible premises, it is reasonable to conclude that the commitment to being responsive to reasons is a commitment to being governed by formal principles (whatever these may be), compliance with

[15] Korsgaard is quite explicit about this: "I am going to argue that in the relevant sense there is no *you* prior to your choices and actions, because your identity is in a quite literal way *constituted* by your choices and actions (Korsgaard 2009: 19).

which ensures that the ends associated with our practical impulses are truly our own. According to the first of the two accounts that endorse this conception of practical reasons, this means that the criteria for what counts as a reason for action are the necessary conditions of *self-determined* behavior. According to the second account, the criteria are the necessary conditions of *nonobservational knowledge of one's behavior*. If, as I will argue, we must reject these accounts, then we must reconsider the premises that appear to support them. We must reconsider the assumption that action is possible if agents endorse the conception of practical reasons these premises express.

Version 1: Constituting Oneself as a Self-determining Cause

Consider the story often repeated by philosophers who note that acting makes special demands on beings with the capacity to wonder what they have reason to do. According to this story, the capacity to reason is (among other things) the capacity to question the desirability of being moved by whatever impulses we happen to have. We rational beings thus need to discover a *reason* to do what we find ourselves *inclined* to do. Unless we discover such a reason, our point of view as rational beings will be insufficiently unified with the point of view constituted by our dispositions to initiate change. Under these circumstances, the effects of our desires will not be attributable to *us*—the ones who are searching for reasons. *Their* motivating force will not be *our* motivating force.

According to this story, whenever we do things for a reason, we have managed to transform a mere collection of mental states, in relation to whose motivating force we rational beings would otherwise be mere passive bystanders, into a unified rational agent—the sort of rational being we are when *we* cause our behavior. How do we manage to do this? According to one compelling and influential version of the story, rational beings incorporate their various impulses into a unified whole by ensuring that these impulses obey the laws that are constitutive of their rationality, where these include laws that determine which ends they have reason to pursue. They employ their reason, not only to determine *how* to satisfy their noninstrumental desires, but also to determine *whether* to do so. They put their reason to practical use in order to "govern" their nonrational dispositions to pursue certain ends. As Christine Korsgaard puts it, "the reflective structure of self-consciousness inevitably places us in a relation of authority over ourselves, and . . . we are as a consequence accountable to ourselves. . . . I act under my own authority as lawgiver, and I am accountable to myself if I do not. So my reasons—and indeed, practical reasons in general—are grounded in the authority the human mind necessarily has over itself" (Korsgaard 2007: 10–11).

But why are the laws of our reason (whatever these are) authoritatively binding on any motives that are independent of these laws, and independent of the commitment to discovering reasons for action that is essential to "the reflective structure of self-consciousness"? How could *we* possibly regard these laws as authoritatively binding insofar as *we* are constituted by inclinations that are independent of our need to discover reasons for acting? I want to argue that this is not possible unless the authority of these directives is grounded in considerations that are relevant from the point of view of the motives themselves. Otherwise, there is no intrapersonal authority relation after all, and so we cannot act. Unless our normative verdicts have an authority for us that is not simply grounded in our identification with our reason, we are in no position to expect any desires that are independent of our reason to obey the demands that are grounded in the verdicts we reach when we reason about the desirability of satisfying these desires; and so, we (the ones who make these demands on ourselves) are in no position to ensure that the motivating force of *our desires* can be attributed to *us*.

In pressing this charge, let me begin by noting that there do seem to be occasions on which we exercise authority over ourselves: we find ourselves tempted to do something we think we have overriding reason *not* to do, and we order ourselves not to do it. ("Sarah, do NOT reach your hand toward that cookie!" "Sarah, KEEP WORKING ON THAT PAPER FOR THE CONFERENCE ON AGENCY AND RESPONSIBILITY!") Surely, however, this is not the way things usually work—at least not for most of us. Exercising authority over oneself is something one does when one's rational agency is under threat. It is not what occurs in the paradigm instances of rational agency. When things are going well, our reasons for action do not include the fact that we have ordered ourselves to act this way, nor the fact that we issue this order in our capacity as rational beings.

This is what Bernard Williams gets right when he says that giving oneself permission to save one's wife would generally involve "one thought too many" (Williams 1981: 18). Most of us rarely need to defer to an exercise of authority in order to act. And because we have no need to *defer* to authority, we have no need to *assert* authority either.

In this respect, we are no different from the purely rational beings to whom Kant attributes a "divine," or "holy," will. (Kant 1956: 81) Because such beings are wholly identified with the verdicts of their own reason (because their point of view just is the point of view constituted by these verdicts), they are not sufficiently divided from themselves to impose demands on themselves; they are not subject to any imperatives; no obligations can figure among their reasons for action. So, too, even if we less-than-divine rational beings can be divided from ourselves in a way that enables us to give and take the same orders, most of us are usually sufficiently identified with the verdicts of our own reason—i.e., these verdicts are usually sufficiently motivating—to make such assertions of authority superfluous. For us, too, the thought that our reason permits or requires a certain course of action is usually one thought too many.

This is, I think, an important objection to the conception of rational agency as the product of an intrapersonal authority relation. I want, however, to focus on a deeper problem with what we can call the "authority-based" conception. To put it bluntly, on this conception of rational agency, insofar as we are desiring beings, we could not care less about whether the motivating force of these desires can be attributed to an agent. And so we could not care less about the fact that in defying the demands of our reason, we would be sabotaging our agency. And so, we have no motive to obey these demands. And so, we cannot obey them.

Of course, in describing the possibility of "sabotaging our agency," I have been forced to speak about what *we* would *do*, and so I have been forced to imagine a situation in which we are sufficiently unified to perform various agent-sabotaging acts. But this merely brings home the point that it is not possible to make sense of action if we posit the division of labor that is essential to the authority-based account. If, as the authority-based conception assumes, the laws of our reason and the ends of our desires stand in no internal relation to each other, then there is no possible way for *us* to exercise authority over *our* desiring selves.

The argument for this claim went by very quickly. I want to rehearse it again more slowly, beginning with an observation by the person who has done more than any other philosopher working today to develop and defend the authority-based account. "If I am to govern myself," Korsgaard tells us, "there must 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" (Korsgaard 2008: 60). According to this division of labor, the govern*ing* self is the agent insofar as she is identified with her reason, and more specifically, with whatever formal principles of rationality impose constraints on which desires she has reason to satisfy. Accordingly, *t*he govern*ed* self—the self on the receiving end of reason's "laws"—must be the agent insofar as she is identified with motives that are not grounded in this commitment.[16] And here is where things get tricky. As we have learned

[16] Note that these motives could nonetheless be responsive to reasons. Indeed, this could well be true of even our deepest instinctual desires.

from the literature on authority, in order for one person to exercise authority over another, the latter must be capable of recognizing the demands of the former as distinct (preemptive) reasons for action. This means that in order for my rational self to exercise authority over my nonrational self, the latter must be capable of regarding the demands of the former as (preemptive) reasons. And this means that it must be possible for *me* (= that aspect of my identity which is constituted by a concatenation of motives not grounded in my reason) to regard the directives given me by *myself* (= that aspect of my identity which is constituted by the verdicts and demands of my reason) as distinct reasons to do certain things and not others—distinct reasons to behave as directed. And *this* means that the nonrational motives themselves—the very motives whose independence from my reason prompts me to give orders to myself—must be such as to enable me (= the one constituted by these motives) to appreciate the normative force of my demands (= the demands I make as the representative of reason).

How is this possible? According to Korsgaard, "the acting self concedes to the thinking self its right to government. And the thinking self, in turn, tries to govern as well as it can. [This is] a relation which we have to ourselves. And it is a relation not of mere power but rather of *authority*." (Korsgaard 1996: 104) But what is it about the thinking self that gives it the "right" to expect obedience from the acting self? What justifies its claim to authority? The point cannot simply be (the relatively trivial point) that *insofar as I am identified with my thinking self* (and its commitments), I cannot call its authority into question. We need to know why that aspect of me which is *not* identified with my thinking self is nonetheless bound to treat the demands of this self as reasons to respond as demanded.

If, *qua* rational being, I demand compliance/obedience from myself *qua* nonrational being, if I base this demand on nothing more than my assumption (which I cannot fail to make insofar as I employ my reason) that the laws of my reason are to-be-obeyed, and if I hear this demand from a point of view from which I am entirely indifferent to the verdicts of my reason and its guiding principles, then my attempt to assert authority over myself fails. But this means that insofar as "exercising authority over myself" is not merely a manner of speaking—insofar as it is not merely what I do whenever I reason about which of my desires it really makes most sense for me to satisfy—it is not possible for me to exercise authority over myself under the conditions stipulated by the account here under consideration. Asserting authority over myself under these conditions would be a fraud of the very sort Williams thinks he sees in every attempt to get someone to respond to "external reasons" to do something (Williams 1981: 101–13). It would be a fraud precisely because on the view we are here considering, the alleged authority of reason cannot be justified without appealing to considerations (e.g. what is (allegedly) necessary in order to act) whose normative force is invisible from the point of view of the mental states over which the authority is asserted.

Of course, as my reference to our "manner of speaking" acknowledges, there is a perfectly unproblematic sense in which I cannot fail to regard the laws of my own reason as authoritative: I cannot intelligibly challenge the authority of the laws of my own reason even as I am reasoning about what to do. But the reference here to "the laws of *my own* reason" is not a reference to some *distinct aspect* of *myself* that asserts authority over *me* as I reason. When I reason, the "laws" of my thought are not among the things I take into account—nor is the fact that they govern my thinking, nor the fact that I impose them on myself. Though there is an obvious sense in which I treat these laws as authoritative, and an obvious sense in which, in so doing, I am imposing certain demands on myself, the fact that I am making these demands is not an additional—trumping—reason for me to act in the ways spelled out in the demands. And so, my relation to myself is not a relation of practical authority; it is not a relation in which I defer to, or obey, myself.

Insofar as my desires reflect my normative judgments, I have no need to unify myself in order for my goal-directed behavior to qualify as my action. My acting self and my thinking self are one and the same.[17] If, on the other hand, my desiring self is completely distinct from my thinking self, then my thinking self has no choice but to regard the relevant desire as a feature of its circumstances—one more fact to be "taken into account." Accordingly, under these circumstances, too, my thinking self must be able to generate its own motives; and so, it must already be my acting self. In order, then, for my actions to result from the integration of two parts of my soul—in order for acting to involve "self-constitution"—the two distinct parts cannot be entirely independent. As a subject of desires which do not themselves reflect my commitment to being governed by my reason, I must nonetheless be capable of recognizing the verdicts of my reasoning as authoritatively binding.

[17] Can my thinking self be alienated from itself? Not in the sense that matters here. No sooner do I call my contemporaneous normative judgment into question, than it ceases to be my judgment. No sooner do I disavow a given assessment of the considerations for and against a given action than I endorse a different assessment—if only one that can be expressed in terms of the disavowal. This having been said, it does seem possible to be alienated from one's normative judgments in a different way: one can call their legitimacy into question, not on normative, but on metaphysical, grounds. This is the sort of self-alienation familiar from discussions in metaethics.

This conclusion has an important implication: "practical reasons [cannot be] grounded in the authority the mind has over itself." For consider. How could I (*qua* subject of reason-independent desires) come to regard the verdicts of my reasoning as authoritatively binding unless I (*qua* subject of reason-independent desires) could take there to be some *reason* to regard myself (*qua* reasoner) as having authority over myself (*qua* subject of desires)? And how could I (*qua* subject of reason-independent desires) have a reason to regard myself (*qua* reasoner) as having this authority unless I had practical reasons for doing so that are *not* grounded in the authority my mind has over itself? These questions lead me to conclude that Korsgaard's account of agency does not "answer" "the realist objection—that we need to explain why we must obey [a] legislator." It does not identify "a legislator whose authority is beyond question and does not need to be established." It does not show us how to make sense of "the authority of [our] own mind and will" (Korsgaard 1996: 104).

There are two respects in which the authority of "our own mind and will" is, for us, "beyond question." First, we are in *no position* to question this authority insofar as we are identified with our own mind and will. But under these circumstances, we lack the distance from ourselves to stand in an authority relation to ourselves. Second, even when we do exercise authority over ourselves, this is possible only because, however strongly we may be inclined to defy the orders of our own mind and will, it is not possible for us—the ones thus ordered—to reject the order as an abuse of authority, a brute assertion of power. This means that, even in our nonrational capacity, we must acknowledge the force of whatever reasons there are to regard the order as authoritative.

How are we to make sense of this double-minded stance? How is it possible for a self to be just divided enough and just unified enough to exercise authority over itself?[18] Joseph Raz's work on *inter*personal authority (Raz 1985) suggests one possible answer: even insofar as we apprehend our options from a point of view that is not grounded in our reason, we acknowledge the *epistemic* authority of the judgments that are so

[18] It is interesting to compare my answer to this question—and, more generally, my critique of Korsgaard's account of action as self-constitution—with Tamar Schapiro's attempt to make sense of the Kantian thesis that when someone does something intentionally, she "incorporates" her desire ("inclination") into her action-guiding principle. Schapiro notes that insofar as I have an inclination that is not grounded in my reason, "part of me sees the world through [the] eyes of [this inclination] and responds to the world [accordingly], and part of me is aware of itself as not being the source of this way of seeing and responding" (Schapiro 2011: 164). So far, so good. My point, however, is that if an inclination really is internal to *my* point of view, then even though "it has no capacity to demand reason or justification for its way of seeing and responding to the world" (156), I cannot be indifferent to such demands insofar as I am the subject of this desire.

grounded.[19] I am not sure whether this is the *only* way things could work. I want, however, to explain what I have in mind in suggesting that satisfying this condition would suffice to make an intrapersonal authority relation possible. In so doing, I hope to reinforce the two related lessons I have just drawn from my discussion of Korsgaard: (1) a genuine intrapersonal authority relation presupposes that there are facts about what we have reason to do that are not grounded in our commitment to acting; and (2) a genuine intrapersonal authority relation requires that in our capacity as the one who is governed, we are not so alienated from our governing self that we cannot appreciate any reason for regarding the demands of the governing self as authoritatively binding.

Suppose, then, that I have reached the conclusion that I have overriding reason NOT to eat that second piece of pie. And suppose that I would not have bothered to employ my reason on this occasion if I had not assumed that its verdicts would have a greater *epistemic* authority with respect to what is really to-be-done than any desire prompted by the close proximity of the pie. Alas, I might nonetheless persist in desiring to eat that second piece of pie, and this desire might be stronger than any others. If to be in this state is to represent the pie as to-be-eaten, then when I am in this state, I am opposed to being guided by the verdict of my reason(ing). Insofar, however, as I represent the pie in this way, I cannot dismiss as irrelevant the belief that it is *not* to-be-eaten-by-me. And so, I cannot be indifferent to which of these representations has greater epistemic authority. If, moreover, I believe that the verdicts of my reason have greater epistemic authority, then even as I am disposed to act contrary to these verdicts, I am also disposed to acknowledge the authority of the demand to act otherwise. In short, precisely because, qua pie-desirer, I am *not* wholly identified with my reason, I am in a position to regard the epistemic authority of my own normative judgments as a distinct reason to grant practical authority to whatever demands reflect these judgments.

Under the stipulated circumstances, my desire to eat the pie puts me at odds with myself; it moves me to defy the verdicts of my own reason. To this extent, *I* regard the directives that are grounded in these verdicts as external constraints—even as *I* am the one who is imposing them. At the same time, however, since I also regard the directives as coming from a source that has epistemic authority with respect to the very point at issue between me and myself (since—rightly or wrongly—even as I desire the pie, I regard the contemporaneous verdicts of my reason as more likely to

[19] I agree with those who reject Raz's appeal to epistemic authority to account for relations of practical authority *between* agents. My suggestion here is that *intra*personal authority relations are importantly different.

track the reasons I really have than are any pie-related desires that are *not* grounded in my reason[20]), I assume that *I* (the person I am insofar as I employ my reason to figure out what to do) am justified in asserting *practical* authority over *myself* (the person I am insofar as I am not identified with my reason). And so I regard the directives as authoritatively binding. I regard these orders as authoritatively binding because I regard them as reflecting my own superior epistemic position with respect to the very issue my desire represents as being at stake when it represents the piece of pie as really, really, really to-be-eaten-by-me.

Version 2: Constituting Oneself as Someone Whose Behavior One does not Find Perplexing

The preceding sketch suggests one way in which a rational being's relation to her own motives could be just external enough to allow for a relation of authority, but not so external as to reduce all assertions of authority over herself to futile attempts to bully herself. I leave for another day the task of considering how this sketch might be filled in, and what other conceptions of the relation of self to self might also strike the right balance. For my purposes, the most important points to stress are these. If someone is as divided from herself as the self-reflective pre-agent appears to be on Korsgaard's account, then there is nothing that this deformed being can possibly do to "pull herself together in order to act." Under these circumstances, if her "thinking self" is not already capable of generating motives for action—perhaps in response to certain (nonrational) motivating states that were not so generated, and so cannot be attributed to her (the thinker)—then it is not possible for her to act. But if for less-than-holy wills, the possibility of rational agency is inseparable from the possibility of intrapersonal authority, then agents with less-than-holy wills (agents whose desires are not necessarily in harmony with the verdicts of their reason) cannot act unless they assume that there are reasons for action whose normative status is independent of both the commitments constitutive of their rationality and whatever other commitments they happen to have.

[20] It is important to stress that I could be mistaken about this: sometimes one's reason leads one astray. The important point here is that *insofar as* I am the subject of a certain normative judgment, I take myself to have reason to rely on it. If I did not believe that the judgment was reliable, then I would see things differently.

I suspect that any account of rational agency is bound to come up short if it fails to do justice to this last requirement. I want now to lend support to this hunch by considering a second account of the conditions under which rational agents avoid self-alienation. This is the account developed with great ingenuity by David Velleman. According to Velleman, "reflection on the phenomena of action reveals that being the subject of causally related attitudes and movements does not amount to participation of the sort appropriate to an agent" (Velleman 1992: 463). More particularly, if desires and means-end beliefs "cause an intention, and intention causes bodily movements," and if there is nothing more to the causal history of our behavior than this, then "nobody—that is, no person—*does* anything; . . . the person serves merely as the arena for [the psychological and physiological events that take place inside]: he takes no active part" (Velleman 1992: 461).

Like Korsgaard, Velleman assumes that in order for rational beings to take an "active part" in the production of their actions, they must determine which of their desires they have reason to satisfy, and the "drive" to be guided by this normative verdict must play a decisive causal role in their behavior. In order to act, rational beings must do what they do because they are moved to do it in their capacity as the ones who demand a reason for satisfying their desires; they must be motivated by whatever motive "drives" their "critical reflection on, and endorsement or rejection of, the potential determinants of [their] behavior" (Velleman 1992: 477).

On Velleman's account, too, there is a tight connection between (i) the need to locate a criterion of practical reasons in the necessary conditions of agency and (ii) a picture of agency that divides the pre-agent in two. But, according to Velleman, for agents who can question the desirability of satisfying their desires, acting is not the result of giving oneself a reason to satisfy a given desire by asserting practical authority over one's practical impulses. Rather, it involves using one's reason to make predictions that elicit the cooperation of that aspect of oneself which is constituted by one's nonrational, practical impulses.

Since on this account the reasoning that yields the predictions is grounded in the desire to understand things in a relatively straightforward way, the division within the self that one overcomes whenever one cooperates with oneself in this way is not the division between one's reason and one's desires. It is the division between a desire one cannot fail to have insofar as one is rational and one's other (contingent) desires. According to Velleman, insofar as a person is identified with her reason, she has one guiding aim: to understand things. And insofar as she is disposed to behave in certain ways in order to bring about certain states of affairs, she is not responsive to reasons.

Before we consider how, according to Velleman, these two "parts" interact to produce action, it is important to see just how far apart their job descriptions ensure that they are. To this end, imagine for a moment that you occupy a point of view from which nothing you see can strike you as the way it ought or ought not to be, or as in any way good or bad. Imagine that you are curious to discover the way things are, but in no way "invested" in what you discover. Imagine, in other words, that it makes no difference to you *what* you come to understand. You are indifferent, for example, to whether you understand that human beings lack the intelligence to slow global warming or whether you understand that slowing global warming is well within their intellectual and emotional capacities. Now, imagine that your response to the world is very different. You find yourself drawn to make various propositions true, without, however, having any capacity to form any opinion about these behavioral dispositions. Finally, imagine that when you occupy the first point of view, one of the aspects of the world that comes into view is the dispositions that constitute the second point of view, as well as their actual and likely effects. Is there anything about these two ways of relating to reality that could enable them to relate *to each other* so as to generate behavior that qualifies as *your* acting for a reason? (Having explored this question, I will consider the possibility that no cooperative relation is necessary—since, in effect, the desire to understand our own behavior can generate new (rational, "motivated") desires to perform certain actions and not others. On this alternative account, action does not involve self-integration.)

By stipulation, none of the contingent, nonrational behavioral dispositions our self-reflective rational being discovers in herself could care less about whether it plays a role in producing an action. None of these dispositions could care less about whether the behavior to which it gives rise is intelligible to anyone. Accordingly, insofar as someone is identified with these dispositions, *she* does not care about whether her behavior is intelligible. This seems to imply that, on the account we are here considering, if we are ever to overcome the divide between us and ourselves, both the aim of overcoming this divide and the capacity to overcome it must somehow be built into our interest in understanding things. But how could this be? Why would (how could) a rational being, construed as one for whom all considerations have normative significance only insofar as they are relevant to understanding things care about whether the world to be understood is a world in which at least some practical impulses are responsive to this theoretical aim?

Of course, since we do, in fact, act, we do, in fact, understand what we are doing in a way that is different from the way we understand the actions of others.[21] So there must be something that accounts for this understanding, and its distinctive features. But this puts the cart before the horse. We need an account of how, on the view we are here considering, we come to act.

Again, we might wonder how the desire for understanding could suffice to pull the trick off. After all, we want to understand *other people's* behavior too; and to achieve this aim, we certainly do not need to understand their behavior as we would understand it if it were *our* action. So, again, why should we think that we have this aim insofar as we aim to understand *our own* behavior? Why does the desire to understand things give us any more reason to object to being surprised by our own behavior than we have reason to object to being surprised by the behavior of others? Wouldn't such behavior simply provide us with different material to understand?

Velleman's answer to this question is, first, that insofar as an event is at odds with one's expectations, one does not really understand why it occurred, and second, that when the motivations are one's own, one has a means of avoiding such surprises, which one does not have when one's predictions concern the behavior of others. As he puts it in the Precis of *The Possibility of Practical Reason*,

> a normal person is aware . . . of being identical with an especially salient member of the objective order—identical, that is, with the creature walking in his shoes sleeping in his bed, eating his meals, and so on. That creature is certainly of great interest to a person, and its doings consequently become the object of the person's intellectual motives. But the person's awareness of being identical with that creature opens up an obvious shortcut to the cognitive goal. The subject can know what that creature is doing simply by doing what he conceives of the creature as doing. . . . Practical knowledge thus supplants theoretical knowledge, as a more secure route to the same cognitive goal. (Velleman 2005: 229)

But what grounds this identity, and the awareness of it? It would seem that, on Velleman's account, the only ground I could have for making a special proprietary claim to the behavior of a particular "member of the objective order" is that I have a special mode of access to the psychic forces that cause this behavior, namely, I am the subject of the qualitative experiences to which these forces directly give rise. Since my position in relation to these forces is, on this account, an exclusively receptive one, it is hard to see why I would think that *their* behavioral effects are *my* movements. And so, it is hard to see why I would want them to be intelligible to me in the way they would be if they were my own.

[21] For the classic discussion of our special epistemic relation to our own actions, See Anscombe 1957. See also Moran 2001.

But perhaps this line of questioning misses the point. After all, Velleman's comment suggests that what matters is not that we will better understand our behavior if we understand it in the way that we do when we understand it as our action. What matters is that we are in a privileged position to ensure that our behavior will not surprise us. According to Velleman, our beliefs about what we will do are based on what we know about our behavioral dispositions. Given our desire to understand our behavior, these beliefs will generally be responsive to what we are motivated to do; and for the same reason, they will generally be sufficiently motivating for the resulting behavior to make them true. When this happens, it is *we* who are the behavior's determining cause. We are the cause of our behavior because the desire that causes it would not do so if we thought that this made no sense in light of everything else we believe and desire. Because the motivating force of the desire is constrained by our desire for self-understanding, the aim *it* moves us to seek is *our* aim too.

But something still seems missing from this account. In particular: why would the aspect of myself that is indifferent to reasons be moved to "accommodate" the "naturally inquisitive" aspect of myself "by enacting ideas of what it would be intelligible for [me] to do"? (Velleman 2009: 18) The problem identified here is evident when we take a careful look at Velleman's most recent attempt to explain his conception of action as the product of a cooperative relationship between a rational being (*qua* rational) and certain reason-independent behavioral dispositions. In particular, his use of pronouns reveals that, on his account, it is not possible for this relationship to qualify as the reflexive self-relation that only genuine unity/identity will support. Here is a representative passage:

> The cognizable object [i.e. **the person** of whom *someone* (= *the inquirer*) is aware in being aware of **himself** and **his** dispositions] is disposed to instantiate what the inquirer already understands: **it** is so disposed because **it** consists in *the inquirer himself,* with *his* drive toward self-understanding. *The inquirer* learns that *he* can make sense of **himself** by [**his**] making sense to *himself*— that is, by [**his**] doing what makes sense to *him*. (Velleman 2009: 17; the underlined words reflect Velleman's emphasis)

In this passage, I have used bold type for those pronouns that refer to the aspect of the person constituted by the contingent behavioral dispositions that are the *objects* of the person's reflection. And I have used italics for those pronouns that refer to the person insofar as he is the *subject* of this reflection, motivated by the desire to understand his behavior. Velleman asks us to think of the relationship between these two aspects of the pre-agent as analogous to the relationship between an improvising actor and the character whose responses to the circumstances the actor is improvising. Because, he argues, **the character** and *the improviser* are one and the same person, **the character** is capable of accommodating *the improviser*; and for the same reason, **he** is motivated to accommodate *him*. Against this suggestion, I have been arguing that the division of labor between the two does not support the identity; and that, accordingly, it provides no basis for the claim that *the improviser's* aims are shared by **the character**.

Velleman is keen to stress that in our capacity as the ones who improvise the responses it would make sense for us to make if we really did have a certain character, *we* do not merely relate to this aspect of ourselves as the object of *our* reflection. *We* also try to see how things look from **this point of view**:

> *He* gathers materials for understanding possible courses of action by reacting to external circumstances in consciously valenced thoughts—thoughts that are consciously desirous or worried or fearful or joyous, as occasioned by the circumstances. Thoughts that are consciously desirous or fearful may illuminate intelligible lines of conduct better. (Velleman 2009: 22)

But since **the character** *he* is trying to understand does not attribute normative significance to anything, "reacting to external circumstances in [**the character's**] consciously valenced thoughts" does not involve seeing facts as having any normative significance. So, these reactions are necessarily external to the commitment that constitutes *his* point of view—namely, the commitment to responding to his circumstances in a way that makes sense. So, these reactions are not themselves *his* take on what makes sense. So, they are, for *him,* mere facts to be considered in determining how *he* has reason to react. To put the same point the other way around: a character and an improviser of this character's reactions are one and the same person, insofar, and only insofar, as they share reason between them. If they are two, then the best that *the improviser* can do is to act as if he were **the character**. In short, under the stipulated circumstances, it is not true that "*he* [is] fully merged with **his character**" (Velleman 2009: 16).

If action requires the cooperation of two parts of a pre-agent, one of which has no motive for cooperating, then action is not possible. Perhaps, however, we can abandon the model of intrapersonal cooperation without rejecting the basic elements of Velleman's account. (Many of Velleman's own comments suggest as much.) According to this suggestion, our desire for self-understanding generates intentions (self-fulfilling beliefs about what we will do) by generating new desires in response to our beliefs about what—independent of our desire for self-understanding—we are disposed to do. Given, for example, that I believe that I am a kind person who desires to help others in need, my desire for self-understanding will give rise to a desire to help this particular person who is struggling with her groceries; and so (assuming that I am aware of no stronger conflicting desires), I will come to believe that I will help this particular person who is struggling with her groceries; and so, given my desire to find my behavior intelligible, I will be moved to help her. (Alternatively, the desire for self-understanding will directly generate the belief that I will help this person, and this will then generate the desire to help her.)

On this account, action does not require a pre-agent's nonrational self to cooperate with her rational self. Accordingly, it does not require the pre-agent to overcome her alienation from the practical impulses she finds herself with. No such self-integration is necessary because, on this account, the aspect of the self from which the subject of self-reflection is alienated is relegated to the status of a very special feature of her circumstances: it is the feature of her circumstances that shares with her the power to directly affect the movements of her body, the feature that shares a body with her. On this account, integrating oneself with one's reason-independent practical impulses is not essential to one's identity as an agent. So, the self-alienation that persists is not an impediment to one's agency (provided, of course, that one's beliefs about one's reason-independent impulses are sufficiently on target to generate reason-dependent motivations that can exert a decisive influence on the movements of one's body). Nonetheless, the fact that, on this account, every agent is necessarily alienated from herself in this way seems to be a defect. (Is it really essential to rational agency that a rational being's understanding of why her action makes sense is completely independent of the way she represents this action to herself insofar as she is the subject of desires that are not grounded in her reason?)

More importantly, it seems to me that this account is plausible only insofar as it attributes to rational agents a substantive assumption about what is really worth doing, independent of their commitment to acting, or any other commitments they may have. Insofar as my desire to help the person struggling with her groceries is a response to my desire to understand myself, it is not simply one more disposition I find myself with when I reflect on the reason-independent constituents of my mind; it is not just one more mental state for me to take into account in deciding how it would be intelligible for me to behave. Nor is it simply the disposition to regard intelligibility as the key to what it is for a fact to be a reason for action. Nor is it simply the effect of such a disposition. As I have described the case, the desire to help the person struggling with her groceries is grounded in my belief that I desire to help people, together with my disposition to regard the intelligibility of my helping her as a *reason* to help her—a distinct, substantive consideration in favor of helping. As far as I can tell, unless my motivation can be construed in this way, my identity as a rational being cannot be unified with my identity as the subject of practical impulses in the way these two identities must be unified if the motivating force of my impulses is to be attributable to *me*—the one who is committed to acting for reasons. My behavior is my action when I am moved by my desire to help the person struggling with her groceries because and only because my desire to help is "under the guise of the intelligible," and because it thus reflects the very commitment that is constitutive of the point of view from which I reflect on the normative significance of my desires. To put the same point the other way around, in the case as I have described it, I am the agent of my behavior precisely because I—the one who demands a reason for doing what I do—regard the intelligibility of what I am doing "under the guise of the good." (Note that this point is analogous to a point I made in discussing Korsgaard, namely, insofar as a rational being does not stand in the sort of authority relation to herself that requires her to distance herself from her own normative verdicts, she is "governed" by these verdicts because she shares the commitment (to being governed by reason) that underlies them. Her "identification" with her reason just is her endorsement of the goals she sets herself insofar as she has the more basic goal of responding to reasons.)

The idea that the necessary conditions of action are the ultimate criteria of practical reasons is a very intriguing one. I have chosen to discuss two fascinating developments of this idea not only because they are fascinating (though this certainly played a part in my choice), but, more importantly, because they represent proposals regarding what rational agency could be if we assume both that (i) no rational agent qualifies as acting intentionally simply in virtue of engaging in instrumentally rational, goal-directed behavior, and that (ii) our reason is not a capacity for discovering mind-independent facts regarding which substantive goals we have reason to pursue. I do not know whether there are any alternative accounts of the source of normativity that are compatible with these conceptual commitments—any alternatives, that is, to (i) an account according to which practical reasons are grounded in laws whose practical authority is itself grounded in nothing but the fact that no one can coherently challenge this authority if she is committed to being the author of her behavior, or (ii) an account according to which practical reasons are grounded in the constitutive aim of theoretical reason, and the more specific aim of not being perplexed about why one is being moved to satisfy a given desire. I strongly suspect, however, that whatever other options there may be, they will have the same structural defects. They will presuppose that our identity as rational beings is distinct from our identity as beings with noninstrumental practical goals. And so they will lack the resources necessary to make sense of the possibility of rational agency.

Something has gone seriously wrong if agency itself ("constituting oneself as an agent") is conceived as the goal of every being with practical impulses—not, to be sure, a goal that such pre-agents conceive under the description "constituting myself as an agent," nor, indeed, any goal they have "in view" (Velleman 1996), but a goal nonetheless. There is an obvious—and trivial—sense in which we bring ourselves into existence as agents every time that we act. But, as far as I can tell, our actions are not events that occur because, and only because, we pursue the particular ends given us by our desires as a way of (or a means to) satisfying a condition without which we fail to act. To act is not to do things as a way of (or a means to) governing or understanding ourselves. Acting is something that just happens when we do things for reasons.

To be sure, actions are not just any old happenings. To act is to govern one's behavior in such a way that one knows what one is up to without observing oneself. This means, I have argued, that (1) there is more to the action of rational beings than the instrumentally rational pursuit of goals. As I have also tried to show, however, (2) no rational being can identify with her motivating desires if she is so alienated from them that she regards the conditions necessary for identification as a constraint on what she has reason to do. If I am right about this, then (3) an adequate account of agency will have to endorse an alternative account of practical reasoning—an account according to which our interest in responding to reasons is an interest in responding to considerations whose normative force depends on something more than our own commitments, be they our commitment to acting or our commitments to achieving certain substantive goals.

These three conclusions emerge from a train of thought that can be summarized as follows. (1) We need an account of rational agency that does justice to the possibility of inner conflict. This means that (2) an account of rational agency must do justice to the fact that if a rational being can question the desirability of satisfying her desires, then she does not qualify as an agent simply in virtue of engaging in instrumentally rational, goal-directed behavior; satisfying this condition does not suffice to ensure that the causal power of her motives can be attributed to *her*. From (1) it also follows, however, that (3) an account of rational agency must not assign to the "parts" of the soul essential to rational agency two roles so independent from each other that there is no way to explain how they could be two aspects of a single agent; if someone were alienated from herself in a way she could not overcome without acting, then she would lack the resources necessary for constituting herself as an agent. More particularly, (4) on any adequate account of rational agency, the nonrational part of the soul must be capable of appreciating the force of the claims/interests/concerns of the rational part, even though what is distinctive about it is precisely that it is not defined by these claims/interests/concerns. This means that (5) there must be something about the *content* of our nonrational desires/inclinations that makes them open to cooperate with, or obey, our normative verdicts.

That is, there must be something about *us* in our nonrational capacity that inclines us to cooperate with, or obey, the interests, or demands, that define us as "rational." It also means that (6) we cannot be so alienated from ourselves that the conditions necessary for unifying our normative verdicts and our nonrational dispositions—the conditions necessary for being an agent—determine what we have reason to do. In particular, (7) insofar as our point of view is not constituted by our normative verdicts, we must take ourselves to have reason to accommodate ourselves to these verdicts, and so we must take there to be reasons for action that are not grounded in the conditions necessary for agency.

REFERENCES

Anscombe, G.E. (1957). *Intention*, 2nd edn. (Ithaca, NY: Cornell University Press).

Bok, Hilary (1996). "Acting Without Choosing." *Noûs* 30 (2): 174–96.

Bratman, Michael (2002). "Hierarchy, Circularity, and Double Reduction." In Sarah Buss and Lee Overton (eds.), *The Contours of Agency* (Cambridge, MA: MIT Press), 65–85.

Brewer, Talbot (2009). *The Retrieval of Ethics.* (Oxford: Oxford University Press).

Buss, Sarah (1997). "Weakness of Will." *Pacific Philosophical Quarterly* 78: 13–44.

—— (2012). "Autonomous Action: Self-Determination in the Passive Mode." *Ethics* 122: 647–91.

—— "Norms of Rationality and the Superficial Unity of the Mind" (Unpublished manuscript).

Frankfurt, Harry. (1971). "Freedom of the Will and the Concept of a Person." *The Journal of Philosophy* LXVIII: 5–20. Reprinted in Frankfurt 1988, 11–25.

—— (1977). "Identification and Externality." In Amelie Oksenberg Rorty (ed.), *The Identities of Persons* (Berkeley, CA: University of California Press). Reprinted in Frankfurt 1988, 58–68.

—— (1988). *The Importance of What We Care About.* (New York: Cambridge University Press).

Hornsby, Jennifer (2004). "Agency and Alienation." In Mario De Caro and David Macarthur (eds.), *Naturalism in Question* (Cambridge, MA: Harvard University Press), 173–87.

Kant, Immanuel (1956). *The Groundwork of the Metaphysics of Morals.* Trans.. H. J. Paton. (London: Hutchinson & Co).

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

—— (2007). "Autonomy and the Second Person Within: A Commentary on Stephen Darwall's the Second-Person Standpoint." *Ethics* 118 (1): 8–23.

—— (2008). "The Normativity of Instrumental Reason." In *The Constitution of Agency.* (Oxford: Oxford University Press), 27–68.

—— (2009). *Self-Constitution: Agency, Identity, and Integrity.* (Oxford: Oxford University Press).

Moran, Richard (2001). *Authority and Estrangement.* (Princeton, NJ: Princeton University Press).

Pettit, Philip and Smith, Michael (1990). "Backgrounding Desire." *The Philosophical Review* XCIX: 565–92.

Railton, Peter (2011). "The Affective Dog and Its Rational Tail." Unpublished manuscript.

Rapoport, Judith (1991). *The Boy Who Couldn't Stop Washing.* (New York: Penguin Putnam Inc.).

Raz, Joseph (1985). "Authority and Justification." *Philosophy and Public Affairs* 14 (1): 3–29.

Schapiro, Tamar (2011). "Foregrounding Desire: A Defense of Kant's Incorporation Thesis." *Journal of Ethics* 15: 147–67.

Smith, Michael (2012). "Four Objections to the Standard Story of Action (And Four Replies)." *Philosophical Issues* 22, *Action Theory*: 387–401.

Velleman, David (1992). "What Happens When Someone Acts?" *Mind* 101 (403): 461–81.

—— (1996). "The Possibility of Practical Reason." *Ethics* 106 (4): 604–726.

—— (2005). "Precis of the Possibility of Practical Reason." *Philosophical Studies* 121 (3): 225–38.

—— (2009). *How We Get Along.* (New York: Cambridge University Press).

Wallace, R. Jay (2001). "Normativity, Commitment, and Instrumental Reason." *Philosophers' Imprint* 1 (3): 1–26.

Watson, Gary (1975). "Free Agency." *The Journal of Philosophy* 72: 205–20.

—— (2004). "Disordered Appetites: Addiction, Compulsion, and Dependence." Reprinted in Gary Watson, *Agency and Answerability* (Oxford: Oxford University Press): 59–87.

Williams, Bernard (1981). *Moral Luck*, 1–19. (Cambridge: Cambridge University Press).

2

The Fecundity of Planning Agency

Michael E. Bratman

1. THREE PRACTICAL CAPACITIES

As normal adult human agents we have a remarkable trio of capacities. First, we are capable of acting over time in ways that involve important forms of intentional cross-temporal organization and coordination. Think about growing vegetables in your garden. You need to prepare the soil, plant, water, cultivate, and harvest. These activities take place over time, and each of them is infused with the agent's understanding of and commitment to the larger temporally extended arc of the activity. This is a case of temporally extended intentional agency.[1]

Second, we are capable of acting together with others in ways that go significantly beyond standard forms of strategic interaction. Think about a piano quartet. It is not just that each is acting in a way that best promotes what she wants given her knowledge of what the other agents are doing, and given that the other agents' thoughts and actions have a parallel structure. Instead, each sees herself and her partners as acting together in ways that involve distinctive forms of commitment and responsiveness to the joint activity and so to each person's participation in that joint activity. This is a case of shared intentional activity.[2]

Third, we are capable of self-governance. On reflection, however, this can seem mysterious. The picture of a self stepping back from the psychic flow and putting a thumb on the scales is deeply problematic even though it does echo aspects of our experience of our agency. But in the absence of such a homunculus, what is it for the agent herself to be governing?

A theory of human agency should include an understanding of this trio of capacities: capacities for temporally extended, for shared, and for self-

[1] A related idea is in Ferrero 2009.

[2] Indeed, it is what I call a shared cooperative activity. See Bratman 1999a. But here I will work with the somewhat broader concept of a shared intentional activity.

governed intentional agency. And in each case we have three interrelated concerns. First, we want to know what concepts best support our theoretical understanding of these phenomena. Second, we want to understand how it is that we realize these capacities. Third, we want to understand the distinctive normative aspects of these capacities. Our concerns are, then, conceptual, metaphysical, and normative.

And my response to these concerns is to seek to understand these three human practical capacities by appeal to a common, core capacity for individual planning agency. This core, planning capacity itself suffices for the capacity for temporally extended intentional agency. Further, when appropriately supplemented, this core capacity suffices for capacities for shared intentional and self-governed agency; and these supplemental resources are broadly continuous with the resources already available within our theory of individual planning agency. In this way we provide a bridge from the conceptual, metaphysical and normative resources of our model of our individual planning agency to conceptual, metaphysical, and normative resources adequate for a model of each of the trio of practical capacities that we have highlighted. We thereby articulate structures of planning agency that suffice for central cases of temporally extended, shared, and self-governed intentional agency. The conjecture of the fecundity of planning agency is that this trio of human practical capacities, and perhaps others as well,[3] is in this way grounded in our capacity for planning agency.[4]

2. PRACTICAL THINKING AND TIME

Not all agents are planning agents. There can be goal-directed agents who do not structure their thought and action over time by way of planning. Dogs and very young human children are likely examples. Such agents engage in goal-directed activity in which their behavior tracks a goal—say, getting a drink of water. And this may involve complex plasticity of response—a thirsty dog may respond sensibly to your moving the bowl of water. But it seems unlikely that the dog, or the fifteen-month human, settles on plans for a temporally extended period, and thinks about and

[3] Allan Gibbard sees our capacity for planning agency as at the bottom of normative thinking (Gibbard 2003). And Scott Shapiro sees law as grounded in our planning capacities (Shapiro 2011). In these ways both Gibbard and Shapiro aim to extend the fecundity of planning agency.

[4] Bratman 2010. What is central is the claim to provide sufficient conditons for central cases of these forms of human agency. The conjecture of the fecundity of agency leaves open the possiblity of multiple such sufficient conditions.

guides his behavior over time in the light of those plans. Of course, whether the dog, or the young child, does or does not guide his thought and action in this plan-infused way is an empirical issue; perhaps psychological research on these matters will surprise us. But the central point here is the distinction between, on the one hand, temporally local goal-directed agency and, on the other hand, plan-guided diachronic agency.[5]

In planning we take seriously what to do later, as well as what to do now. But why worry about what to do later? Why not just worry about what to do now and cross your bridges when you come to them later? The commonsense answer is that we need to figure out what to do later in order to figure out what to do now and to coordinate our earlier with our later activities. Once I settle on going to New Orleans next month it is clear that I need soon to get an airline ticket. In contrast, if I just ask right now "should I right now get a ticket to New Orleans for next month?" I normally cannot sensibly answer without settling on whether to go there then. In the interest of cross-temporal coordination, I need many times to settle now what I will do later, and then plan accordingly about what to do now. In settling now to go to New Orleans next month I come to be in a plan state that is future-directed. And we understand what such a plan state is by explaining its role in the rational dynamics of planning agency.

This pressure in the direction of future-directed planning is especially acute for resource-limited agents like us.[6] We frequently need to settle now what to do later in part because we cannot reasonably trust that when the later time arrives we will be able, given the pressures of action, to sort out all the complexities efficiently and reach a sensible decision from scratch.

In developing this idea of the cross-temporal coordinating role of planning we need also to do justice to two further, commonsense ideas. The first is the idea of choice among alternative options each of which would be sensible—one aspect of, as we might say, the idea of "will."[7] We make such choices all the time, from different routes to a destination, to different careers, to the choice made by Sartre's young man between aiding his mother and fighting with the Free French.[8] And these choices, while underdetermined by the agent's view of her reasons, nevertheless settle relevant practical matters and play central, forward-looking roles in cross-

[5] For the idea of planning agency as a species of agency, see my use of Paul Grice's strategy of "creature construction." Grice 1974–5 in Bratman 2007 f.

[6] Simon 1983.

[7] A second aspect is the idea of willpower. See Holton 2009. And here too, as Holton argues, the planning theory can be of use.

[8] Sartre 1975, 354–6.

temporal coordination. The ability to make choices and settle such matters seems a basic feature of human agency, one that helps explain why we are not constantly suffering the plight of Buridan's ass. This is an ability to make a transition from a state of indecision between several alternatives, each of which seems sensible, to a state of being settled on one of those alternatives in particular. We shed light on this ability by saying more about the output state of being settled on one of those alternatives in particular. And we understand this output state in large part in terms of its role in planning agency: to be settled on an option is, or anyway is closely related to,[9] being in a plan state, where, again, we understand what a plan state is by explaining its role in the rational dynamics of planning agency. In this way we resist an overly simple time-slice conception of the will by embedding our account of the will within a larger model of diachronic planning agency.

There is, second, the very idea of intention. Here we need to be careful to distinguish adverbial forms—as in "he did it intentionally"—from the verb "to intend." The idea that is central here concerns the latter, and the corresponding phenomenon of intending; and we can allow for a more or less complex relation between X-ing intentionally and intending to X.[10] The central thought here is that to intend to do something is to be in a plan state, where—again—we understand what a plan state is by explaining its role in the rational dynamics of planning agency. Intending leads to action in ways that normally involve diachronic planning structures.

So a model of the rational dynamics, and the associated cross-temporal coordinating roles, of planning agency would also be a theory of the will and a theory of intending.[11] This promises a deeper understanding of the forward-looking nature of the will, and of the complex diachronic interconnections between our thought and our action. And the conjecture of the fecundity of planning agency is that such a planning model sheds light on the trio of agential capacities with which we began.

[9] This qualification is needed to allow for the complex relation between choice and intention that I discuss in Bratman 1987, ch. 10.

[10] This is where I have argued against the "Simple View" according to which it is always true that in intentionally A-ing one intends to A. (Bratman 1987, ch. 8). In that discussion I also considered the "single phenomenon view" according to which when one intentionally A's one intends something, though perhaps not, in particular, A. I said that something like this will be true in "the vast majority of cases," but I left it open that there might be marginal cases in which one acts intentionally and yet there is no intention—no plan-state—involved. (pp. 126–7.) What is important here is the conjecture, built into the planning theory, that cases of intentional action in which relevant plan states are completely absent, if such there be, are theoretically marginal for our understanding of our adult human agency. For a challenge to this conjecture See Velleman 2007.

[11] Holton 2009, 5, argues that "there is solid psychological evidence that we have intentions along something like these lines."

3. PLANNING AGENCY

Let me now sketch a model of our individual planning agency that I have discussed elsewhere.[12] We begin with plans of action. Such a plan specifies ways of acting now and into the future. It ties these ways of acting, together with what one believes about the world, into a more or less consistent and coherent web, though a web that will normally be both partial in important ways and to some extent conditional in structure. To plan to do something is to be appropriately committed—committed in a distinctive, practical way—to a plan of action that says to do that. Planning agents are systematically involved in such commitments to relevant plans, and these commitments shape their practical thought and action over time.

Intending to do something is, I have said, being committed in the relevant way to a plan that says to do that (perhaps conditionally). We try to explain this idea of being appropriately committed to such a plan of action by appeal to two general ideas. First: we try to characterize the normal roles—the normal ways of functioning—that are characteristic of such plan states; and second we try to articulate basic planning norms, the at-least-implicit acceptance of which by the agent is involved in those roles.

Begin with these planning norms. Intentions—that is, plan states—are subject to a characteristic quartet of norms. Intentions are to be internally consistent and consistent with the agent's beliefs. It needs to be possible to agglomerate one's various intentions together into an overall intention that is itself consistent and consistent with one's beliefs. Intentions in favor of ends engage demands of means-end coherence in the direction of filling in one's—normally, partial—plans appropriately, and as time goes by, with means and preliminary steps.[13] And structures of plan states, though potentially a target of reconsideration, are subject to a norm of stability over time: there is some sort of defeasible presumption in favor of one's prior plan states. In short, there are planning norms of consistency, agglomeration, means-end coherence, and stability. A planning agent will at least implicitly accept these norms and this will tend to guide that agent's thought and action in ways that support conformity to these norms.

To accept these norms is not simply to be disposed to conform to them. One is also set to see divergence from these norms as breakdowns, and so to

[12] See esp. Bratman 1987.

[13] Having noted the category of preliminary steps, I proceed to ignore it and focus on means.

respond to such cases with a kind of "Darn it!" reaction. One is thereby set to be guided in the direction of conformity with the norm.[14]

This norm-guidance is part of the explanation of characteristic roles that such plan states play. Given the pressure from an accepted planning norm to avoid means-end incoherence, prior, partial plan states pose problems about means; and the agent will be set to fill in her plans accordingly. This will normally induce end-means reasoning that is structured by prior, partial plans. Given the pressure from accepted planning norms in favor of agglomeration and consistency, plan states will tend to filter out from deliberation options intending which would introduce new plan inconsistencies. And given the pressure from an accepted planning norm of stability, a prior intention will tend to persist, other things equal. This persistence is sometimes a result of the snowball effect: once one begins acting on a prior plan one changes the world in ways that tend to support continuing with that plan.[15] And sometimes this is because the costs of reconsidering prior plans—and/or the risks that reconsideration poses of undermining important coordination previously forged—are too high in the absence of recognized reasons for change; and a prior plan will tend to persist unless it is reconsidered.[16] But our planning agency also involves, I think, a distinctive norm of stability that directly says to treat one's prior plans as a defeasible default.[17]

So we have a package of characteristic roles together with associated norms that are at least implicitly accepted. Such a planning psychology induces and supports complex forms of temporal organization of thought and action. And a plan state in this planning psychology can favor a specific option despite known underdetermination by prior reasons. Once Sartre's young man settles, say, on a plan to work with the Free French, demands of consistency and means-end coherence come to bear and support corresponding forms of temporally downstream functioning. This is true even though his decision in favor of the Free French was, by his own lights, underdetermined by his prior reasons. And we can make a similar point about less dramatic examples of career decisions, or even of different routes to an intended destination.

If all goes well, planning structures induce cross-temporal referential connections that are both forward and backward looking. My present

[14] For discussion of norm acceptance See Gibbard 1990. See also Railton 2006.

[15] See Bratman 1987, 82, where I note that I owe to John Etchemendy the label "snowball effect."

[16] Yet another potential explanation of this persistence is that we are frequently guided by habits of nondeliberative nonreconsideration, habits that are, in general, useful to us.

[17] Bratman 2010, 2012.

plan to paint the house this week at least implicitly refers to my later, then-present-directed intention to put on the final coat of paint; and that later intention at least implicitly refers back to my earlier intention. And the stability of my intention in favor of painting this week helps support a coordinated flow of activity over time. These cross-temporal referential interconnections and constancies help support an effective temporally extended structure of partial plans. These partial plans provide a background framework for further deliberation about means, deliberation shaped by accepted norms of coherence, consistency, and agglomeration.

This idea of cross-temporally stable and referentially interlocking attitudes is familiar from the Lockean tradition of understanding personal identity over time—or, at least, "what matters" in such identity[18]—by appeal to overlapping strands of continuities of attitude and referential connections across attitudes.[19] The standard functioning of plan states in our planning agency involves such broadly Lockean cross-temporal ties.[20]

Now, in speaking of accepted planning norms I have so far not raised the question whether these norms do indeed have normative force or significance. I have, up till now, only appealed to the explanatory role of their acceptance in a planning psychology.[21] But I do not think we should rest content simply with such a "positivist" theory of planning agency. There are two reasons for this. The first is that we ourselves are planning agents: our theory of planning agency is a theory of our agency. So the question whether these norms have normative force is one we face in such theorizing. After all, there are likely many patterns of thought that are pervasive in human agency as we know it, but which we would not, on reflection, endorse.[22] So we will want to know whether, and if so why, these planning norms do pass reflective muster. Second, it would be an important theoretical fact about planning agency if these structures of agency were indeed robust in the face of reflection on its basic principles. One important way to ascertain whether our planning agency has such robustness is directly to ask whether these planning norms really do have normative force.[23] And a positive answer would defuse a concern that the planning

[18] Parfit 1984, 217.

[19] See John Locke, *An Essay Concerning Human Understanding* Bk. 2 ch. 27.

[20] The idea is only that these planning ties are among the broadly Lockean ties, which also include, most prominently, memory. So one can survive a breakdown in these planning ties.

[21] For a closely related distinction See Schroeder 2003. And also See Broome 2007.

[22] See e.g. the essays in Kahneman et al. 1982. For a discussion of related matters, See Morton 2011.

[23] In saying this I remain neutral about the extent to which it is up to us whether to continue to be planning agents.

theory mis-describes our practical thinking insofar as it ascribes to us the acceptance of norms that are not defensible.[24]

A closely related issue is whether, and in what sense, these planning norms are norms of *rationality*. We can begin by observing that synchronic planning norms of consistency, agglomeration, and means-end coherence have obvious parallels with synchronic norms of consistency, agglomeration, and theoretical coherence of belief; and there is perhaps also a parallel between the diachronic planning norm of stability and a diachronic norm of belief conservation. I have argued elsewhere that the practical planning norms are not simply a matter of theoretical norms on belief.[25] Nevertheless, there remains an important parallel. Given their tight tie to belief, there is an initial case for seeing these norms of belief as norms of theoretical *rationality* for agents who have beliefs. Similarly, given their tight tie to intending/planning there is an initial case for seeing these planning norms as norms of practical *rationality* for planning agents; though, since not all agents are planning agents, they need not be rationality norms for all agents.[26]

More needs to be said; but we need first to return to the trio of practical capacities with which I began this essay.

4. TEMPORALLY EXTENDED AGENCY

The capacity for temporally extended intentional agency is the capacity to guide one's actions in light of one's grasp of their location in a larger temporally extended structure of activity to which one is practically committed. One plants the seeds as a part of a larger activity of growing beans in one's garden, a larger activity to which one is practically committed. This cross-temporal organization of this activity is mind-infused: this is not merely the cross-temporal biological organization of seeds developing into beans.

In what sense mind-infused? Here my proposal is that for us the normal explanation of this kind of temporally extended intentional activity involves the agent's plans and planning: one's plan is to grow the food by

[24] For a version of this concern See Kolodny 2009.

[25] See esp. Bratman 2009a.

[26] This idea of rationality norms for a specific kind of agent fits well within the strategy of Gricean creature construction that lies behind the planning theory. Kieran Setiya discusses a related idea, with skepticism, under the rubric of "pluralistic rationalism" in Setiya forthcoming. I am very much indebted to Setiya's discussion, though I reach different conclusions.

cultivating the garden; one recognizes that this needs to involve relevant means such as planting the seeds and watering; and so one includes those means within the plan to which one is practically committed. One's practical commitment to this plan involves one's being set intentionally to act in these specified ways in the course of following through with one's overall plan. In this way our temporally extended intentional activity is an exercise of our planning capacities.

Michael Thompson has criticized what he calls "the tendency of students of practical philosophy to view individual human actions as discrete or atomic or point-like or eye-blink-like units."[27] Now, I think that the planning theory in my 1987 book is not guilty of this purported tendency: a central concern of that book was, after all, to understand how planning rationally supports our temporally extended agency. In the present context the point is that the first dimension of the fecundity of planning agency—its support of our capacity for temporally extended intentional agency—continues to avoid an atomistic tendency that, as I agree with Thompson, we should avoid.[28]

5. SHARED AGENCY

Turn now to our capacity for shared intentional activity. The proposal here is that our capacity for shared intentional activity is grounded in our individual planning capacities in the sense that the proper functioning of those planning capacities, given relevant special contents of the plans, contexts of the plans, and interrelations among the participating planning agents, would constitute a basic case of shared intentional activity.

As anticipated, this does not mean that simply by having the capacity for planning agency one thereby has the capacity for shared intentionality. To get from planning capacities to the capacity for shared intentionality we need further resources in the form of distinctive contents, contexts, and interrelations. So there remains the possibility of planning agents who are not capable of shared intentionality. Indeed, it is some such gap between planning agency and shared intentionality that Michael Tomasello and his colleagues have highlighted as a potential key to a fundamental difference between humans and the great apes.[29] Nevertheless, the further resources—conceptual, metaphysical, and normative—involved in the step from

[27] Thompson 2008, 91.

[28] Though Thompson would not agree with my way of avoiding such atomism.

[29] See e.g. Tomasello 2009.

individual planning agency to shared intentional agency will be available within our theory of individual planning agency.[30] Or so I conjecture.

The key is the idea of shared intention. When we paint the house together, or dance together, or play a quartet together, or have a conversation together, or perform an experiment together, or—in Margaret Gilbert's example—walk together,[31] our activity is explained by our shared intention in favor of our so acting. We are walking together because that is what *we* intend to do—where talk of what we intend to do is talk of our shared intention so to act. Or at least this connection between shared activity and shared intention holds in the kinds of cases that are our central concern. In this respect shared intentional activity parallels individual intentional activity: in each case, the *explanatory* role of intention (individual or shared) is fundamental. And the idea is to articulate structures of interconnected individual planning agency whose rational functioning would constitute the rational functioning of shared intention.

If we are going to see certain cases of interconnected rational functioning of individual planning agents as constituting the rational functioning of shared intention, we need to say more about what such rational functioning of shared intention is. So let us ask the same questions about shared intention we asked about individual intention: why do we bother with shared intentions? What fundamental roles do they play in our lives, and what norms are associated with those roles?

My response to these queries is to highlight analogies with the coordinating and structuring roles of intentions and plans in the individual case, and with the associated norms of individual intention rationality. The characteristic roles of a shared intention to *X* will include interpersonal coordination of action and planning in pursuit of *X*, and the structuring of related bargaining and shared deliberation concerning how we are to *X*. And these roles will be associated with norms of social consistency and agglomeration, social coherence and social stability. Roughly, it should be possible to agglomerate relevant intentions into a larger social plan that is consistent, that in a timely way adequately specifies relevant means, and that is associated with appropriately stable social psychological structures. Failure to satisfy these social norms will normally undermine the distinctive social roles of shared intention.

We want then to specify a structure of interconnected plan states of individuals (and other related attitudes of those individuals) in appropriate

[30] Though see the qualification below about "out in the open."

[31] Gilbert 1990.

contexts that would, when functioning in the norm-guided ways highlighted by the planning theory of individuals, play these roles characteristic of shared intention in part by way of conformity with these associated norms of social rationality.[32] This would allow us to say that such individualistic planning structures constitute shared intention, or at least one important kind of shared intention.

We can call such a broadly individualistic structure of interconnected planning agents a *construction* of shared intention, and the view on order as a kind of constructivism about shared intention. The idea is not that the participants themselves construct the shared intention: it is we, the theorists, who do the constructing.[33] And the proposal is that some such construction will constitute shared intention, or at least a central kind of shared intention.

I have tried to develop such a construction elsewhere.[34] Here I proceed in broad strokes. What we need are a number of ideas that are within the domain of our theory of individual planning agency but when put together allow us to characterize a central form of shared intention and shared intentional activity. In particular, and focusing on our shared intention to paint the house, the initial ideas we need are:

a. Each of us intends that we paint the house.
b. Each of us intends that we paint the house by way of the intention of the other that we paint the house.[35] In this sense our intentions interlock.
c. Each of us intends that we paint the house by way of mutual responsiveness between us in sub-plan and in action. So, in particular, each intends that our sub-plans for our painting mesh with each other, in the sense of being co-possible.

[32] In putting the idea in this way I am signaling an asymmetry. In the individual case the basic explanation of conformity to the individualistic rationality norms involves the at-least-implicit acceptance of those very norms by the relevant individuals. In the shared case, in contrast, the initial explanation of conformity to the cited *social* norms appeals to the at-least-implicit acceptance of the norms of *individual* plan rationality by the relevant individuals. However, once these interconnected planning structures are on board it may well be that the individual participants also accept, and are guided by, the social norms themselves. I discuss these matters in Bratman forthcoming a.

[33] Though in special cases there will also be a sense in which the participants themselves construct the sharing. The theoretical construction I am envisaging would be a part of a larger strategy of Gricean creature construction.

[34] See e.g. the quartet of essays on this subject in Bratman 1999a, Bratman 2009c, and Bratman forthcoming a.

[35] I also think we need the idea that each intends that we paint by way of his own intention that we paint. But to keep things more manageable I put aside here this condition of reflexivity.

d. Each of us believes that our intentions in a. are interdependent in their persistence, and that given these intentions we will indeed paint the house.
e. There is in fact interdependence in persistence of the intentions in a.
f. These conditions are out in the open among us.

These conditions describe a structure of interconnected intentions of each, and associated beliefs. The constructivist proposal is that when this structure of attitudes functions in accordance with the norms of the planning theory of individual agency it thereby realizes the roles characteristic of shared intention in part by way of conformity with the associated social rationality norms. Further, when this structure in fact leads to our painting the house by way of the cited kinds of mutual responsiveness, there will be shared intentional activity. Such shared intentional activity is the activity of interconnected planning agents.

Without trying fully to defend this proposal here, let me make some brief comments. First, to avoid problems of circularity, the concept of our painting the house that appears in the contents of these intentions of each should (at least in basic cases) be understood as neutral with respect to shared intentionality. We might paint together in this neutral sense in a case in which we are each merely individually painting the same house while keeping an eye out not to bump into each other. (What will be crucial for shared intentionality is that this neutrally characterized pattern of our individually intentional activities is explained by the kinds of intentions cited in the construction.)

Second, these conditions appeal to the phenomenon of my intending *that we* act in a certain way: not all intending is intending *to*.[36] I think that once we understand intending within the framework of the planning theory this becomes a fairly natural idea,[37] so long as each participant also believes that her own intention appropriately settles whether they are going to act accordingly. And my conjecture is that each can sensibly have such a belief when each has the beliefs cited in d. about interdependence and efficacy.[38]

Third, the condition of intention-interlocking precludes cases in which each intends that the joint action proceed in a way that does not involve the intention of the other, as when a certain Mafioso, intending to throw the other into the trunk of the car, announces that "we are going to New Orleans together." In contrast with such a mafia case, the idea here is that

[36] Davis 1984. Though, as Randolph Clarke emphasized in conversation, a rational agent who intends *that* we *J* will normally also intend *to* act in certain ways.

[37] As Philip Cohen once suggested in conversation.

[38] For this last point See Bratman 1999b.

each intends that the joint activity proceed by way of the relevant intention of the other participant, and by way of associated mutual responsiveness between the intentions and actions of the participants. And this intended mutual responsiveness will need to involve mesh between the relevant sub-plans of each.

A fourth observation concerns the condition that all this is out in the open.[39] A central reason for this condition is that we want to model a kind of shared reasoning that is characteristic of shared intention, namely: shared reasoning about how to carry out our shared intention to paint the house together.

I leave it as an open question whether the relevant idea of "out in the open" goes somewhat beyond the conceptual resources of our theory of individual planning agency. Except for that possible qualification, however, I think that the conceptual, metaphysical, and normative resources at work in this proposed construction are within the domain of the resources of our theory of individual planning agents: the construction is conceptually, metaphysically, and normatively *conservative*.

The basic conjecture, then, is that these interdependent and interlocking intentions of each, taken together with the cited beliefs and in a public context, will, in responding to the rational pressures associated with individual planning agency, function in ways that constitute the social-norm-conforming social functioning of shared intention. So this structure of interdependent and interlocking intentions constitutes at least one central form of shared intention.

In partial support of this basic conjecture, note that we are supposing that each *intends* the joint painting by way of the other's intention, mutual responsiveness, and meshing sub-plans. It is not just that each intends his part and merely expects the other to play her part.[40] This means that the rational pressure on each to make her own plans coherent and consistent ensures rational pressure on each to support the success of the joint activity and the meshing role of the other in that activity. Given my intention that we paint by way of your intention, mutual responsiveness, and meshing sub-plans, I am under rational pressure to coordinate with you, to support your role—perhaps by way of helping actions—and to avoid ways of acting

[39] It is here that talk of common knowledge and/or mutual belief is potentially apt.

[40] This contrasts with a case in which each only intends his own activity while expecting the others to perform their part, and the social organization is out-sourced to some external managerial group. In such a case the participants will not each be committed to the joint activity in a way that is normally involved in shared deliberation about how to proceed. (Scott Shapiro discusses large-scale versions of this latter sort of case in Shapiro forthcoming.)

that are incompatible with all that. And given your analogous intention, you too are under analogous rational pressures. These rational pressures on each, given these distinctive contents and interrelations, induce pressures in the direction of social coherence and consistency, and associated coordination and effectiveness, pressures that are characteristic of shared intentionality.

This approach to shared intentionality highlights conceptual, metaphysical and normative *continuities* with individual planning agency. This contrasts with views that see the step to shared intentionality as necessarily involving a basic new conceptual and/or metaphysical and/or normative resource. For example, John Searle thinks that shared intentionality involves a special kind of "we-intention" in the heads of the individual participants. We-intentions, on Searle's view, are distinctive attitudes, not to be identified with ordinary intentions that concern our own activity.[41] Again, Margaret Gilbert thinks that shared intentionality involves a primitive interrelation of "joint commitment" between the participants, an interrelation that essentially involves distinctive nonmoral obligations.[42] Without pausing to examine the details of either of these views, we can say that if the plan-theoretic model I have proposed does succeed in providing sufficient conditions for shared intentionality,[43] then the burden of proof is on both Searle and Gilbert to explain why we also need these new primitives.[44]

6. SELF-GOVERNANCE

Turn now to our capacity for self-governance. Here my proposal is that a key to a nonhomuncular model of our self-governance is the idea that certain plan states both concern what matters in the sense of having weight in our deliberative thought,[45] and when functioning properly tie together our thought and action in relevant ways, both synchronically and diachronically. These plan-like commitments to weights speak for the agent, and when these plan states guide, the agent governs. This capacity for self-governance involves the capacity for guidance of thought and action

[41] Searle 1990.

[42] Gilbert 2009.

[43] What about the obligations that are commonly present in cases of shared intention and shared intentional action? On the plan-theoretic model these will normally be familiar moral obligations, such as an obligation to follow through with an assurance, or in accordance with reliance that is intentionally induced.

[44] As Michael Smith noted in conversation.

[45] I put to one side other possible ways in which certain considerations can matter.

by plan-like commitments to weights. Not all planning agents have this latter capacity. But its addition is a conservative extension of our model of planning agency, and is an important step towards sufficient conditions for self-governance.

I am led to this view by reflection on the idea that in self-governance the agent's relevant practical standpoint appropriately guides her thought and action. Such a practical standpoint serves as an anchor for deliberation, thereby supporting the idea that its guidance is a form of *governance*. And given the overall role of this standpoint in the agent's psychic economy, its appropriate guidance constitutes the *agent's* governance of his thought and action.[46]

Should we say that the agent's practical standpoint is constituted by her evaluative or normative judgments? Here I aim to be neutral as between different metaethical accounts of the nature of such judgments. I will, however, assume that such judgments are not merely personal preferences but are in some sense in the domain of the intersubjective.[47] Different metaethical theories will interpret this idea differently. Some will appeal to the idea of rational convergence; some will appeal to the idea of enjoining others to converge. But I aim to be neutral about these differences.

That said, I think that there are difficulties with the proposed appeal to evaluative or normative judgment. First, some such judgments may not be a part of one's relevant standpoint. As David Velleman observes, one can be alienated from one's evaluative or normative judgments.[48] Perhaps this is what we should say about Huck Finn's judgment that he ought to turn in Jim.[49] A related point is that one can sometimes be, as Allan Gibbard puts it, "in the grips" of a normative or evaluative judgment.[50] The subjects in Milgram's famous experiment may have believed that they should do what the person in the white coat told them to do; yet their wrenching conflict may indicate that this belief is not located securely in their standpoint. Yet further, one can judge that something is a good thing, but have no interest in it at all.[51] There are, after all, many good things and not that much time.

In these cases there is reason to exclude certain normative or evaluative judgments from the agent's practical standpoint. There are also cases in

[46] Bratman 2007b. This general approach—though not the details—is in the spirit of work of Harry Frankfurt. See Frankfurt 1988a and 1999a. (An important difference is my emphasis on the role of these commitments to weights in deliberation.) Concerning the agent's guidance, see Velleman 1992.

[47] Bratman 2007e, 151–4.

[48] Velleman 1992, 472.

[49] Bennett 1974, Arpaly and Schroeder 1999, and Driver, 2001, ch. 3.

[50] Gibbard, 1990, ch. 4. The example to follow comes from Gibbard's discussion.

[51] See Harman 2000b, 129.

which important elements in that standpoint resist identification with evaluative or normative judgment. Think about the kinds of love that have been highlighted by Harry Frankfurt.[52] What I love will normally play a central role in my practical standpoint. But in many cases it seems strained to see such loving as identical with a normative or evaluative judgment. Further, even when one seeks to respond to one's judgments about the right and the good, these judgments may underdetermine one's practical commitments. Sartre's young man may see considerations of loyalty to his mother and of loyalty to the Free French as noncomparable, and yet arrive at a commitment to give significantly more weight to the interests of his mother. Or he might be struck by broad disagreement in reflective views about the relative importance of these different forms of loyalty, and so be led, by way of a kind of humility of judgment, away from an intersubjectively accountable evaluative judgment in favor of one over the other. Yet he still might go on to a commitment to giving significantly more weight to one rather than the other. In each case such practical commitments go beyond evaluative judgment, but may still loom large in the agent's practical standpoint.[53]

These considerations suggest that a central element of a planning agent's practical standpoint will be plan-like commitments to weights in deliberation. Such commitments are missing in cases of alienated or noninterested value judgment. In contrast, love involves plan-like commitments to give significant weight to the interests of the beloved. (Which is not to say that it only involves that.) Commitments to certain personal ideals or projects involve plan-like commitments to give weight to relevant considerations. These commitments to weights will frequently be associated with relevant judgments about the right and the good. Nevertheless, these commitments are not in general identical with or ensured by such judgments, and they may settle matters in response to underdetermination by such judgments. The role of these commitments is to settle where one stands on questions about how to live one's own life; and they need not be intersubjectively accountable in the way in which evaluative judgment is.[54]

[52] Frankfurt 2006.

[53] Bratman 2007b, 233–8.

[54] Bratman 2007e, 153. Since these plan-like commitments to weights will normally have an important generality, I have called them self-governing *policies*; and I have also emphasized that they will include policies about the significance of one's own desires and/or what those desires are for. See Bratman 2007b and, 2007c. There are important similarities between my appeal to policies about weights and Alan Gibbard's appeal to plans "of 'treating R as weighing in favor of doing X'," in Gibbard 2003, 189. Gibbard's main concern is his metaethical proposal that judgments about normative reasons are expressions of such plans. My main concern is with the role of such plan states in self-governance.

What explains why such plan-like commitments speak for the agent[55] and are such that when they guide the agent governs? I have said that the agent's practical standpoint serves as an anchor for deliberation, and so guidance by this standpoint is a form of *governance*. But I have also alluded to other roles of that standpoint, roles that help support the idea that when it guides the *agent* governs. What roles are these?

Here my proposal is to embed the role of anchor in deliberation within the overall role of knitting together in distinctive ways the agent's practical thought and action, both synchronically and diachronically. What ways? Well, what we learn from broadly Lockean approaches to personal identity over time is that there is a close connection between such personal identity—or, anyway, "what matters" in such identity—and overlapping strands of continuities of attitudes at different times, and cross-reference among attitudes at different times. If certain attitudes that anchor deliberative thought had the role of knitting together thought and action in these broadly Lockean ways—in ways that essentially involve such diachronic continuities and cross-references—these attitudes would organize that agent's life over time and thereby help constitute the cross-temporal identity[56] of the agent whose life thereby has this organization. And my proposal is that it is in playing this role of anchoring deliberation in ways that constitute and support such broadly Lockean cross-temporal organization of one's thought and action that, in the absence of relevant conflicts,[57] an attitude speaks for that agent.[58]

And now the point is that plan-like commitments to weights do indeed realize these "design specifications": they anchor deliberation in ways that involve these broadly Lockean roles in the diachronic organization of the agent's practical life. This follows from the content of these commitments—namely, what is to have weight in deliberation—together with the general theory of the nature of plan states. Given their distinctive contents these plan-like commitments can provide premises about weights, premises that shape deliberation. And plan states organize our thought and action over time in Lockean ways. So there is a case for the idea that these plan-like commitments to weights, in the absence of relevant conflicts, do indeed speak for the agent. When plan-like commitments to weights function properly, they weave together the agent's practical thought and

[55] As I once put it, why do these commitments have "agential authority"? See Bratman 2007d.

[56] Or anyway, "what matters" in such identity.

[57] This appeal to the absence of relevant conflict draws on Harry Frankfurt's idea of satisfaction in Frankfurt 1999b; and See Bratman 2007c, esp. 34–5, 44.

[58] A related idea is in Yaffe 2000, ch. 3.

action in ways that, in the absence of relevant conflict, help constitute where the agent stands: their guidance is the agent's self-governance.

Granted, it might be that certain kinds of evaluative judgment also realize these design specifications.[59] But not all evaluative judgments do this; not only evaluative judgments do this; there are systematic ways in which a person's evaluative judgments will tend to underdetermine the practical commitments that do this; and these practical commitments are in important cases not intersubjectively accountable in the way in which evaluative judgment is. So it is reasonable to see plan-like commitments to weights as central to our nonhomuncular model of human self-governance.

7. WHY PLAN RATIONALITY MATTERS

Return now to the norms of individual plan rationality. Let's focus on the synchronic norms of consistency, agglomeration, and means-end coherence. And here reflection on the fecundity of planning agency gives us something to say about the force of these norms. After all, accepting and being guided by these norms is, on the theory, part and parcel of being a planning agent. Planning agency is fecund in the sense that it helps support, inter alia, important forms of temporally extended, shared, and self-governed agency. And, I take it that we sensibly value these forms of agency: they significantly enrich our lives. This means that we have good reason to value being a planning agent, and so to value our continued acceptance of these norms. Giving up these planning structures, if we could, would cascade through our lives and come at a high price.

This is a broadly pragmatic rationale for continuing to be a planning agent, and so for continuing to accept these synchronic norms—in a sense of "pragmatic" that includes both instrumental and noninstrumental relations between such planning agency and what we sensibly value. This pragmatic support for our continued acceptance of these planning norms does draw on a form of instrumental reasoning; but it need not draw specifically on the norm of means-end coherence of intentions that is among the norms it seeks to support. So there need be no circle.

There remains, however, a problem, one familiar from J. J. C. Smart's concerns about "rule worship."[60] Even if there are significant advantages that accrue to the acceptance of these general synchronic norms, it seems that there can still be particular occasions in which one does better—better as judged by the very values to which our pragmatic argument appeals—by

[59] Bratman 2007b, 249.

[60] Smart 1967.

violating one or more of these norms. Perhaps there are, on specific occasions, unusual benefits of inconsistency and/or incoherence. Given this possibility we cannot in general infer directly from the pragmatic support for the acceptance of these norms to normative support for conformity in the particular case.[61] So we are still without an explanation of the normative force of these norms in each particular case, an explanation to which a reflective planning agent can appeal to support her conformity to these norms in her particular case and to rebut the charge of unjustified consistency or coherence "worship."

One response to this is to see the idea that these norms in general have force in the particular case as a "myth."[62] In contrast, I think that our reflections on the relation between planning agency and self-governance point to a less skeptical view. Recall that a self-governing planning agent will have a plan-structured practical standpoint that speaks for him and whose relevant guidance is his self-governance. And note that for your plan-structured standpoint to speak for you about a particular practical matter it had better not be inconsistent or incoherent in relevant ways. After all, if you intend E, know that this requires that you intend means M, and yet do not intend M, you have no clear stance on E; there is, as Harry Frankfurt might say, no place where you stand with respect to E.[63] And similarly if you intend E but also intend F while knowing that E and F are not co-possible. In such cases you might manage to act intentionally and for a reason; but you have no clear standpoint concerning E whose guidance of your conduct can constitute your self-governance with respect to E. In this way the violation of these consistency and/or coherence norms precludes relevant self-governance for planning agents like us.

This is a point about the metaphysics of self-governance in the case of planning agents; it is not a point that depends on our having already established that conformity to these norms has normative force. So we can use this point about the metaphysics of self-governance to explain an aspect of this normative force. And we can do that if we suppose that a planning agent has a normative reason in favor of her self-governance in each particular context of intentional agency.[64]

We begin with the point, just noted, about the role of plan consistency and coherence in the metaphysics of self-governing planning agents. We then appeal to a principle that if there is normative reason for S to achieve

[61] So we cannot take for granted what Michael Thompson calls a "general-to-particular transfer principle." (2008), 171.

[62] Raz,2005, Kolodny, 2008.

[63] Frankfurt,1988b, 166.

[64] Though we may need to qualify this supposition for contexts of trivial decision.

E, and if S has the capacity to achieve E, and if X is constitutively necessary for E, then there is normative reason in favor of X.[65] We then infer, given our assumption of a reason for self-governance, that if you are a planning agent who has the capacity for relevant self-governance then you have a reason of self-governance to conform to the planning norms of consistency and coherence in this particular case.[66]

In partial summary: The fecundity of planning agency leads to a pragmatic rationale for continuing to be a planning agent, one who accepts the cited norms. Given that we are—as we have pragmatic reason to be—planning agents, our self-governance involves, as a necessary constitutive element, relevant plan consistency and coherence. So, on the assumption that we have reason to govern our lives in the particular case, and that this self-governance is in our power, we have a reason of self-governance to conform to these synchronic norms in the particular case, a reason over and above the specific reasons we have for the specific options in play. In this sense these capacities of planning agency are *self-reinforcing*: if we do have these planning capacities (capacities supported by a pragmatic rationale), and if we do have the capacity for relevant self-governance, then we thereby have a distinctive reason of self-governance to conform in the particular case to the synchronic norms whose acceptance is an element in these very planning capacities.[67]

In this way a reason for self-governance, taken together with the metaphysics of self-governing planning agency, supports a tendency toward equilibrium between one's pragmatically grounded acceptance of the cited planning norms and one's reasons for conformity to those norms in the particular case. And my tentative conjecture is that this combination of pragmatic support, self-reinforcement, and tendency toward normative equilibrium lends further support to the idea that these norms are, indeed, norms of practical rationality for planning agents. In these multiple ways, the fecundity of our planning agency helps us understand the complex normative significance of norms the acceptance of which is at the heart of that planning agency.[68]

[65] This is not to say that *ought* works this way. Nevertheless, there are hard issues that I cannot discuss here about this proposed principle.

[66] Bratman 2009b. Compare David Copp's thought that "*rationality* is in the service of self-*government*" (Copp 2007, 351), and Kenneth Stalzer's thought that means-end incoherence is a failure of "self-fidelity." (Stalzer 2004, ch. 5).

[67] I discuss a related idea of self-reinforcement, in the context of a discussion of views of David Gauthier, in Bratman forthcoming b.

[68] Thanks to the participants at the 2011 New Orleans Workshop on Agency and Responsibility. Special thanks to David Shoemaker both for his deft organization of the workshop and for his helpful comments on an earlier draft.

REFERENCES

Arpaly, Nomy and Schroeder, Timothy (1999). "Praise, Blame and the Whole Self," *Philosophical Studies* 93, 161–88.

Bennett, Jonathan (1974). "The Conscience of Huckleberry Finn." *Philosophy* 49, 123–34.

Bratman, Michael E. (1987). *Intention, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press; reissued 1999 by CSLI Publications).

—— (1999a). *Faces of Intention* (New York: Cambridge University Press).

—— (1999b). "I Intend that We *J*." In Bratman, 1999a, 142–61.

—— (2007a). *Structures of Agency: Essays* (New York: Oxford University Press).

—— (2007b). "Three Theories of Self-Governance." In Bratman, 2007a,. 222–53.

—— (2007c). "Reflection, Planning, and Temporally Extended Agency." In Bratman, 2007a, 21–46.

—— (2007d). "Two Problems About Human Agency." In Bratman, 2007a,. 89–105.

—— (2007e). "A Desire of One's Own." In Bratman, 2007a, 137–61.

—— (2007f). "Valuing and the Will." In Bratman, 2007a, 47–67.

—— (2009a). "Intention, Belief, Practical, Theoretical." In Robertson (ed.), 2009, 29–61.

—— (2009b). "Intention, Practical Rationality, and Self-Governance." *Ethics* 119, 411–43.

—— (2009c). "Modest Sociality and the Distinctiveness of Intention." *Philosophical Studies* 144, 149–65.

—— (2010). "Agency, Time, and Sociality," *Proceedings and Addresses of the American Philosophical Association* 84, 7–26.

—— (2012). "Time, Rationality, and Self-Governance," *Philosophical Issues* (Action Theory) 22.

—— (Forthcoming a). *Shared Agency* (New York: Oxford University Press).

—— (Forthcoming b). "The Interplay of Intention and Reason," *Ethics.*

Broome, John (2007). "Is Rationality Normative?" *Disputatio* 2, 161–78.

Cohen, Philip R., Morgan, Jerry, and Pollack, Martha E. (eds.) (1990). *Intentions in Communication* (Cambridge, MA: MIT Press).

Copp, David (2007). *Morality in a Natural World* (New York: Cambridge University Press).

Davis, Wayne (1984). "A Causal Theory of Intending." *American Philosophical Quarterly* 21, 43–54.

Driver, Julia (2001). *Uneasy Virtue* (New York: Cambridge University Press).

Ferrero, Luca (2009). "What Good is a Diachronic Will?" *Philosophical Studies* 144, 403–30.

Foot, Philippa (ed.) (1967). *Theories of Ethics* (New York: Oxford University Press).

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

—— (1988b). "Identification and Wholeheartedness." In Frankfurt, 1988a, 159–76.

—— (1999a). *Necessity, Volition, and Love* (New York: Cambridge University Press).

—— (1999b). "The Faintest Passion." In Frankfurt, 1999a, 95–107.

—— (2006). *Taking Ourselves Seriously and Getting It Right* (Stanford, CA: Stanford University Press).

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

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

Gilbert, Margaret (1990). "Walking Together: A Paradigmatic Social Phenomenon." *Midwest Studies in Philosophy* 15, 101–14.

—— (2009). "Shared Intention and Personal Intentions." *Philosophical Studies* 144, 167–87.

Grice, Paul (1974–5). "Method in Philosophical Psychology (From the Banal to the Bizarre)." *Proceedings and Addresses of the American Philosophical Association* 48, 23–53.

Harman, Gilbert (2000a). *Explaining Value and Other Essays in Moral Philosophy* (New York: Oxford University Press).

—— (2000b). "Desired Desires." In Harman, 2000a, 117–36.

Holton, Richard (2009). *Willing, Wanting, Waiting* (New York: Oxford University Press).

Kahneman, Daniel, Slovic, Paul, and Tversky, Amos (eds.) (1982). *Judgments Under Uncertainty: Heuristics and Biases* (Cambridge: Cambridge University Press).

Kaufmann, Walter (ed.) (1975). *Existentialism: from Dostoevsky to Sartre* (New York: Meridian/Penguin).

Kolodny, Niko (2008). "The Myth of Practical Consistency." *European Journal of Philosophy* 16, 366–402.

—— (2009). "Reply to Bridges." *Mind* 118, 369–76.

Leist, Anton (ed.) (2007). *Action in Context* (Berlin: de Gruyter/Mouton).

Morton, Jennifer (2011). "Towards an Ecological Theory of the Norms of Practical Deliberation." *European Journal of Philosophy* 19, 561–84.

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

Railton, Peter (2006). "Normative Guidance." In Shafer-Landau, 2006, 3–33.

Raz, Joseph (2005). "The Myth of Instrumental Rationality," *Journal of Ethics and Social Philosophy* 1, 1–28.

Robertson, Simon (ed.) (2009). *Spheres of Reason: New Essays on the Philosophy of Normativity* (New York: Oxford University Press).

Sartre, Jean Paul (1975). "Existentialism is a Humanism." In Kaufmann (ed.) (1975), 345–69).

Schroeder, Timothy (2003). "Donald Davidson's Theory of Mind is Non-Normative." *Philosophers' Imprint* 3, 1–14.

Searle, John R. (1990). "Collective Intentions and Actions." In Cohen, Morgan, and Pollack (eds.) (1990), 401–15.

Setiya, Kieran (Forthcoming). "Intentions, Plans, and Ethical Rationalism." In Vargas and Yaffe (eds.) (forthcoming).

Shafer-Landau, Russ (ed.) (2006). *Oxford Studies in Meta-Ethics* 1.

Shapiro, Scott (2011). *Legality* (Cambridge, MA: Harvard University Press).

—— (Forthcoming). "Massively Shared Agency." In Vargas and Yaffe (eds.) (forthcoming).

Simon, Herbert (1983). *Reason in Human Affairs* (Stanford, CA: Stanford University Press).

Smart, J. J. C. (1967). "Extreme and Restricted Utilitarianism." In Foot (ed.) (1967), 171–83).

Stalzer, Kenneth (2004). *On the Normativity of the Instrumental Principle* (Ph.D. Thesis, Stanford University).

Thompson, Michael (2008). *Life and Action* (Cambridge, MA: Harvard University Press).

Tomasello, Michael (2009). *Why We Cooperate* (Cambridge, MA: MIT Press).

Vargas, Manuel and Yaffe, Gideon (eds.) (Forthcoming). *Rational and Social Agency: Essays on the Philosophy of Michael Bratman* (New York: Oxford University Press).

Velleman, J. David (1992). "What Happens When Someone Acts?" *Mind* 101, 461–81.

—— (2007). "What Good is a Will?" In Leist (ed.) (2007), 193–215.

Yaffe, Gideon (2000). *Liberty Worth the Name: Locke on Free Agency* (Princeton, NJ: Princeton University Press).

3

Can I Only Intend My Own Actions?

Intentions and the Own Action Condition

Luca Ferrero

1. INTRODUCTION

1.1

The possible objects of one's desires and wishes seem to be virtually unlimited: one might desire unattainable states of affairs including, perhaps, even known logical impossibilities. By contrast, the proper objects of one's intentions appear to be much more limited. As Baier (1976: 214) writes: "My intentions must not only be 'made' by me, when I make up my mind, they must be directed upon, they must concern, my own future actions." It seems that the agent can only intend *to do* something *herself*: intentions appear to be necessarily *de actu suo* (Wilson 1989).

This restriction on the proper object of intentions is reflected in the grammar of attributions of intention: an agent is usually said "to intend *to do* such-and-such." The syntactical complement of attributions of intention is usually an infinitival verb phrase whose subject is implicitly understood to be the very agent of that intention. Although it is not ungrammatical to say that an agent *S* intends *that p* be the case, it seems that the propositional complement "that *p*" should be interpreted as making implicit reference to the agent's own action. That is, "*S* intends that *p*" is elliptical for "*S* intends *to do* what it takes for her to bring about that *p*."

This restriction on the admissible objects of intentions—which sometimes goes under the name of the "own action condition"—suggests that there is a special relationship between an agent, her intentions, and her own actions.[1] The own action condition (OAC, hereafter) might strike the reader as trivial, in the sense of both obviously true and uninteresting. Nonetheless, this thesis has not gone utterly uncontested.[2] In addition, many philosophers claim that the status of OAC bears on important issues in the philosophy of agency. For instance, according to both Thompson (2008) and Setiya (2011) a proper characterization of the relation between intentions and actions depends on the fate of OAC. Boyle and Lavin (2010) defend OAC as central to their rejection of causal theories of action and their defense of the "guise of the good" thesis. Finally, OAC matters for the characterization of shared intentions in joint agency (See Bratman 1997 and Stoutland 2002).

Hence, in spite of its apparent triviality, OAC is worth a closer look. In fact, in this paper I will argue against it: the genuine object of intentions is neither necessarily nor primarily restricted to the agent's own actions. Although the agent's own agency plays a necessary role in carrying out her intentions, this role is not reflected in the restriction of the intention's objects to her own actions. The object of an intention is not necessarily cast in the infinitival/agential form—"I intend *to do* so-and-so." The more inclusive propositional clause—"I intend *that* such-and-such"—is actually the proper characterization of the logical form of intentions. Or so I will argue.[3]

1.2

I will begin with some preliminary considerations on the notion of the object of an attitude in general (Section 2). I will then discuss how to characterize the content of a simpler kind of practical attitude, which I call "aiming." I will argue that OAC does not hold of aimings (Section 3). Then I will move to intentions and show that the features that make them different from aimings are not sufficient to support the application of OAC to intentions (Section 4). I will then consider whether one's own actions might still be taken to be the standard although not necessary object of

[1] OAC is defended by Baier (1970: 650), Meiland (1970), Castañeda (1972: 140), Castañeda (1975: 25, 169–75), Baier (1976: 214), Searle (1983: 100, 105), Gustafson (1986: 104, 121, 204), Wilson (1989), Perloff (1991: 403), Velleman (1997), Stoutland (2002), Thompson (2008: 120–3, 127–8, 130–1), Moran and Stone (2009: 143, 147).

[2] See Bratman (1997), Tuomela (2005), Setiya (2011).

[3] In this paper, I use "propositional" as an umbrella term that covers various possible characterizations of content—whether in terms of propositions, sentences, states of affairs, and the like. For present purposes, what matters is only the contrast between the characterization in broadly conceived propositional terms and the one in the agential/infinitival form.

intention. I will argue that although one's own actions might play a prominent role they do not do so as distinct pieces of conduct and separate targets of one's intentions (Section 5). Finally, I will consider the different degrees to which one's own agency might be involved in intentions. I will claim that one's own agency is necessarily involved but in a way that is formal and generic, and as such lends no support to OAC (Section 6). Space constraints prevent me from discussing how my conclusions bear on the various philosophical disputes that invoke the fate of OAC but for the discussion of the distinction between intending and acting. I will argue that my rejection of OAC puts pressure on the tenability of a stark distinction between intending and acting. This is a somewhat surprising development of the standard dialectic, since this substantive claim is usually supported on the basis of the acceptance of OAC (5.8).

2. THE OBJECT OF AN ATTITUDE

2.1

Talk of the "object" or "content" of an intention is not as straightforward as it might appear. Grammatically speaking, the object is the syntactic complement of expressions of intention, but what we are interested in here is rather the object as the complement of the *logical* form of intentions.

Before investigating this logical form, however, we need to make some preliminary considerations about the notion of the content of attitudes in general and various notions of the conditions of satisfaction or success of an attitude.

Let's begin with the simpler case of belief. Take the belief that *p*. The content of this belief—the proposition *p*—*individuates* the attitude. It differentiates this particular belief from other beliefs, such as the belief that *q*. (This individuating content is also something that can be "shared" with attitudes of a different kind, for instance, the desire that *p*, the assumption that *p*, etc.) The individuating content of a belief also seems to indicate its *conditions of success*: the belief that *p* succeeds as a belief when *p* obtains.

At first, it seems that the same holds true of practical attitudes. Consider the desire that *p*. The proposition *p* individuates it by differentiating from other particular desires. In addition, a desire that *p* is satisfied—succeeds as a desire—only when *p* obtains. Likewise for intentions. The object of the intention (whether characterized in infinitival or propositional terms) seems to both individuate a particular intention and indicate its conditions of success.

These considerations, however, are too hasty since they do not pay attention to the different senses in which an attitude can be successful.

First, an attitude is *constitutively successful* when it meets the standards that govern the acquisition, retention, and abandonment of attitudes of its kind. For instance, beliefs are regulated by the standards of veridicality: the belief that *p* is constitutively successful when *p* is true. This explains why the obtaining of the conditions of individuation of a belief coincides with the obtaining of its conditions of "constitutive" success.

This coincidence, however, might not hold of other attitudes: conditions of constitutive success need not be the same as the attitude's individuating content. For instance, assumptions are not regulated by veridicality: the assumption that *p* is individuated by *p* but, unlike the belief that *p*, it might be constitutively successful even if *p* is false.

Something similar holds of desires. The constitutive standards of desires are those of "desirability": one is supposed to acquire, retain, and abandon a desire in light of the "desirability" of its content—whatever that turns out to be according to the substantive accounts of the nature of desires. The desire that *p* can be constitutively successful even if *p* does not obtain and never will. The obtaining of *p*, however, matters for a different kind of success, the desire's *fulfillment* or *satisfaction*.

2.2

The satisfaction of a desire is a form of success distinctive of *practical* attitudes, but it is not to be confused with another kind of practical success, the "achievement" as the success of *executive* attitudes. An executive attitude, such as an intention, is an attitude that *moves* the agent toward a goal. When the agent has an executive attitude aimed at the state of affairs *g*, the mere obtaining of *g* is not sufficient for the attitude's executive success. The obtaining of *g* does not count as an achievement if *g* comes about independently of the attitude or via its deviant operation (although the agent's non-executive desire that *g*, which might accompany the executive attitude toward *g*, can still be successful—in the sense of "satisfied"—no matter how *g* comes about).

Thus, for executive attitudes like intentions, there are three kinds of conditions: (i) of individuation (formulated in terms of the goal *g*), (ii) of constitutive success (to be met in order for the agent to be correct in having that attitude), (iii) of executive success (the non-deviant obtaining of *g by way of* the proper operation of the executive attitude). The distinction is not simply between these notions but also between the substantive conditions: the goal *g* might individuate the attitude, but its obtaining is neither necessary for constitutive success nor sufficient for the executive one.

The difference between these conditions matters for the attitude's logical form. Talk of its "object" should be cast in terms of individuating content. The conditions of individuation are those that matter for the discussion of restrictions on the proper objects of an attitude. Considerations about constitutive and executive success pertain instead to discussions about *kinds* of attitudes. They matter for the understanding of the distinctive operation of attitude-types—their distinctive *mode* or *force*—rather than the content of their tokens.

It can be quite tempting to conflate the different conditions, especially when talking about the objects of attitudes. Particularly dangerous is the nowadays common philosophical talk of the attitudes' aims. This expression when properly used indicates what regulates attitudes as a matter of their constitutive conditions (for instance, to say that beliefs aim at truth is to single out veridicality as their constitutive standard). This "constitutive aim," however, should not be confused with the "individuating aim" of particular *executive* attitudes, that is, with the particular goal *g* that the agent has when she is set on a particular pursuit. Only executive attitudes have individuating aims, since having aims in this sense is what makes them executive. (Additionally, executive attitudes are also under conditions of executive success and, as a result, there might be a sense in which they "aim" at achievement but the target of this aiming is not to be confused with the substantive goal that provides their individuating content—see 5.7.)

3. AIMING

3.1

Before discussing whether OAC holds of intentions, let's consider whether it is true of a simpler kind of executive attitude, which I call "aiming." Aiming is the distinctive attitude of the *basic* form of representational agency. When an agent aims at a state of affairs *g*, she has *g* as her goal, that is, she is set on making *g* true in light of her representation of it.

The pursuit of *g* by aiming at it consists of the combined exercise of several *executive powers.*

1. Power of *self-motion.* The subject *as a whole* is the source of its conduct (the conduct is not the mere product of external forces or uncoordinated operations of the subject's subsystems).[4]

[4] See Frankfurt (1978), Burge (2009).

2. Power to *represent* the goal and orient one's conduct in view of it.[5]
3. Power to *respond to interferences* by (i) adjusting conduct in the face of perturbations and (ii) persisting in the pursuit in the face of some setbacks.
4. Practical *ingenuity and opportunism*: the ability to take advantage of favorable conditions, including: (a) refraining from interfering with advantageous courses of events, (b) reliance on other agents, (c) exploration of novel ways to progress toward the goal.
5. *Two-way powers.* The exercise of executive powers might not be merely *reactive*: the subject might not simply react automatically and in fixed ways. It might respond to circumstances in light of its perception of or belief about the fit between its response and the circumstances, rather than being simply triggered by them (See Kenny 1992: 70).

Some of these powers might be exercised in isolation from each other and, as such, underpin even simpler forms of agency. In addition, agents might have these powers in domain-specific forms and exert them with different degrees (if any) of conceptual sophistication and reflection. Nonetheless, I contend that some combination of these powers, even if in the absence of conceptual sophistication and reflection, is constitutive of basic goal-directed agency, of the agency exhibited in "aiming."

3.2

A particular aiming is individuated in terms of its goal *g*, the state of affairs toward which the aiming is *executively* directed. When *g* obtains nondeviantly *by way of* the combined operation of executive powers constitutive of aiming, the aiming at *g* is executively successful—it achieves *g*.

The conditions of executive success are not part of the individuating content, of the goal *g*: when one aims at *g*, one is orienting the executive capacities toward making the states of affairs *g* obtain. One is not orienting them toward "the bringing about of *g* in a nondeviant manner and by way of this very aiming." The nondeviant bringing about of *g* is not the individuating goal of the attitude. Rather it spells out what aiming consists in as an attitude *kind*. Hence, it needs to be spelled out only when *articulating* the characteristic force of aiming as a kind, rather than the distinction between token aimings.

[5] The orientation might be minimal: it requires neither a plan toward *g*, nor the representation of *g* as a goal. At bottom, it amounts to the capacity of taking instrumental steps and registering one's progress toward *g* (at least, registering when *g* obtains so as to stop its pursuit).

These considerations suggest that the following is the canonical logical form of aiming:

(A) S aims at *g*

where *g* stands for the state of affairs whose obtaining constitutes the aim's achievement—*g* is the goal of this executive attitude. The "object" of the aiming is specified in (broadly conceived) propositional terms, not in agential/infinitival ones.

3.3

Limiting the content of the simpler form of goal-directed agency to one's own actions is overly restrictive. Although it is possible to aim just at one's own actions, this restriction does not seem supported by reflection on the operation of goal-directed agency. Executive capacities can be oriented toward goals quite remote from the agent's own actions, that is, from the agent's exercises of these capacities. As long as the cognitive and conceptual abilities allow for it, the agent's goals can extend as far as the remote *consequences* of her actions, including the results of her omissions, the actions of other agents, and the effects of other agents' actions. That a state of affairs *g* might obtain either by omission or the mediation of others' actions does not make it illegitimate as a goal.

A state of affairs qualifies as a goal as long as it can provide suitable orientation for the exercise of executive capacities. There might be limitations on *how* remote the object might be to qualify as a goal rather than the object of a wish. But this is a concern with the *upper* boundary of aiming's possible objects, which puts no pressure on restricting acceptable goals to the lower boundary of the agent's own actions.

These considerations rule out OAC for aimings. If, as I maintain, talk of the object of an attitude should be interpreted in terms of individuating content, and an aiming's individuating content is its goal, then the object of aimings cannot be restricted to the agent's own actions.

One might try to reject this conclusion by interpreting talk of the aiming's "object" in terms of what is especially relevant in the attitude's psychological operation rather than in its individuating content. For instance, one might argue that the main focus of the exercise of one's executive capacities in the aiming at *g* is not the (explicit) representation of *g* but of the instrumental actions to be taken toward it. And one might continue with the claim that the object of an attitude is to be equated with the content of this focus. But this suggestion is problematic. It is not sufficiently systematic and comprehensive. First, the focus of the psychological operation of aimings might vary widely; second, oftentimes the remote goal rather than the means seems to be at the center of one's attention (this is especially so, I surmise, for the simpler executive attitudes like aimings).[6] In addition, even if the goal is not the focus of orientation, it is hard to deny that it plays some role in the actual psychological operation of aimings.

3.4

To reject OAC for aimings is not to deny that they display important *de se* features: these features can be found both (i) in the ownership and exercise of executive capacities, and (ii) in the path toward the achievement of *g*.

First, the immediate exercise of the executive powers that underpins the agent's aiming cannot but be the agent's own. For it is *constitutive* of the subject's identity as an agent that she is the locus of this immediate exercise. This is an instance of a more general phenomenon: the fundamental *de se* involvement that characterizes the existence and identity of subjects as separate loci of individual psychologies. A subject is none other than the locus of the *unmediated* exercise of psychological powers, including the executive ones (See Burge 2000).

Second, the path toward achievement originates in the agent. Execution originates *in* the agent and proceeds *from* her: from her specific location in space and time and from her point of view on her surroundings. In addition, execution is subjected to the limitations imposed by the agent's present executive powers and circumstances. In this sense, execution is always ego-centered and perspectival.

These features do not entail that the object of one's aiming should be restricted to the exercise of one's executive capacities. Execution is necessarily *ab se*, but its object/target need not be *de se*.

The psychological work is, ultimately, always of one's own. For the subject of this working is nothing other than the locus of the unmediated psychological operations. This is not to say that one's executive attitudes are either exclusively or primarily oriented toward these operations or their immediate effects. The distal and outward-looking orientation of aiming is rather a manifestation of the proper operation of goal-directed agency.

The *ab se* character of execution does not bear directly on the object of executive attitudes, but it matters for determining the conditions of both executive and constitutive success. First, executive success can only be secured via the initial and unmediated exercise of one's executive capacities.

[6] The focus of the psychological operation on instrumental steps might affect the content of intentions as more complex executive attitudes; see plans-as-recipes in 4.5.

It must originate in the agent. Second, constitutive success is related to what counts as an acceptable goal, which is partly a matter of what is attainable by the agent *ab se* from her present circumstances.

The *ab se* character of execution is undeniable, but it does not support OAC. For it is the *de se* aspect of agency in its source rather than in its target. Although any achievement has to go through the agent's contribution via the exercise of her executive powers, her targets can be at much remove from that origin.

A state of affairs might not count as a genuine goal if the connection between its eventual occurrence and the exercise of executive capacities is excessively remote—too thin and fragile. But a concern about excessive distance gives no reason to push the objects of executive attitudes all the way back to the immediate exercises of executive powers.[7]

The relation between the exercise of executive powers and the obtaining of the goal *g* can be quite indirect. For *g* might just be a remote and indirect effect (including via the intermediation of other agents) of processes initiated by these exercises. The ability to aim at such distal states of affairs is one of the remarkable products of practical ingenuity and opportunism, when coupled with the agent's predictive ability.

Opportunistic capacities and the two-way volitional powers also make it possible to pursue goals by relying on already favorable circumstances. Agents often make progress toward their goal simply by refraining from "antagonistic" interventions in the natural course of events. These nonantagonistic omissions can contribute to genuine achievement as much as acts of commission. In the limiting case, an agent might achieve *g* simply by monitoring the unfolding of a favorable course of events. She is to monitor it with an eye toward making possible corrective interventions, which she is able and ready to perform but might not to be required.[8] This nonantagonistic attainment is still a *bona fide* achievement, i.e. an executive success.

3.5

To sum up: the discussion of aiming shows that there are fundamental *de se* elements in the location and working of executive attitudes. But these elements do not necessarily make the content of these attitudes to be exclusively or primarily about one's own actions: the admissible goals of aimings are

[7] My claims concern the logical form not the metaphysics of executive attitudes. It might well be that possessing an executive attitude amounts to nothing more than a disposition to or the actual exercise of certain immediate executive capacities.

[8] Compare the case of the driver going downhill in Frankfurt (1988).

not limited to the immediate exercises of one's executive powers or their proximal outcomes. OAC is not true of aimings.

4. INTENTION

4.1

Let's return to intentions. Intentions are executive attitudes. They are considerably more complex than aimings, but does this difference affect their logical form and support OAC?

Intentions differ from aimings in four ways:

1. Intentions are under distinctive rational pressures for stability and agglomeration, whereas aimings are not rationally criticizable when unstable or unagglomerable (See Bratman 1987).
2. Intentions extend the temporal reach of agency, by allowing the pursuit of very distal goals and engagement in temporally extended and "unified" activities (See Ferrero 2009b).
3. Full-fledged intentional agency usually goes together with more sophisticated conceptual capacities. It is the agency characteristic of agents who are (a) able to articulate their goals and their plans to reach them, (b) in the business of both offering and asking for folk-psychological explanations, and (c) capable of at least some reflection about their agency and psychology.
4. Intentions are *planning* attitudes (Bratman 1987): (a) they frame further deliberation and help coordinate conduct over time (in part by fixing expectations about future conduct); (b) they often come with plans as (partial and hierarchical) recipes that list some of the steps to be taken toward one's goal.

None of these distinctive features of intentions support OAC. This is easy to show for the first two features. First, the rational pressures on the intention do not concern the form of their individuating content. Likewise for the temporal reach of agency. It does not affect the content's form, although it allows for the pursuit of more complex and distal substantive goals.

4.2

Perhaps, one might devise a way to support OAC starting from the *de se* character of the alleged self-referential nature of intentions, which relates to the third feature of intentions, their alleged conceptual sophistication.

Some philosophers maintain that the content of an intention includes the role that the intention plays in securing its executive success.[9] If so, its logical form is something like:

> *(I-sr) S* intends that: *this very intention* nondeviantly
> results in the obtaining of *g*.

The content of (I-sr) has a necessary although indirect *de se* element. For the intention included in the content is none other than the agent's own intention (that is, "*S* intends that: this very intention *of one's own* nondeviantly results in the obtaining of *g*"). But this *de se* element does not support OAC. For it does not bear on the characterization of what the intention is supposed to result in. One needs a separate argument to prove that *g* is to be restricted to the agent's own actions. The self-referentiality of the intention only pertains to the origin of executive success. It spells out the *ab se* character of achievement (see 3.4).

In any event, I find the self-referentiality of intentions problematic for two reasons. First, it is based on a confusion between the individuating content of a particular intention (which is not self-referential) with the reflective articulation of the conditions of executive success of intentions as a kind of attitude.[10] Second, it is psychologically unrealistic insofar as it demands that only agents with the capacity for entertaining self-referential content can have genuine intentions.[11]

The last concern ultimately undercuts any strategy of attempting to derive OAC from the conceptual sophistication of subjects capable of intentional agency. Although intentions might be the characteristic executive attitudes of reflective and conceptually sophisticated agents, it is

[9] See Searle (1983: 85), Harman (1986: 86), Harman (1993: 141).

[10] This confusion underlies the "content satisfaction view" used by Searle (1983) to support self-referentiality. According to him, the content of an attitude should be equated with its conditions of satisfaction. But as shown above (2.2), the idea of satisfaction/success is ambiguous. The content satisfaction view seems to work for some attitudes, such as beliefs and (nonexecutive) desires—the content of beliefs happens to correspond to the conditions of their constitutive success; the content of desires happens to correspond to the conditions of their success as satisfaction. But this equivalence is accidental; there is no principled reason to think that it holds of attitudes in general, let alone of executive attitudes given that the conditions of constitutive and executive success of executive attitudes are much more complex than their individuating content. (My point here is stronger than the one made by Mele (1987: 316–17), who only claims that the content satisfaction view does not hold of intentions but might be fine for beliefs and desires.)

[11] See the criticisms of Mele (1992: 204–6), Kapitan (1995: 154, fn. 8), Roth (2000), and Harman's (1993: 145) acknowledgment of the problem. In addition, Kapitan (1995: 163) correctly remarks that the efficacy of intending does not seem to depend on the representation in its content of the conditions of its success.

problematic to limit their possession to these agents. Conceptual and reflective sophistication might help with shaping one's agency by helping in devising more complex goals and improving one's rate of success thanks to one's understanding of how the intention contributes to the attainment of one's goals. But explicit understanding of executive success does not change the logical form of intentions. In particular, it does not induce the agent to take her own achievement of *g* as her goal: having the goal *g* already equates with being set on *achieving g*, on obtaining it in the mode of executive success (see 5.7).[12]

4.3

Consider now the last distinctive feature: the planning character of intentions. Their role in framing and coordinating deliberation and conduct does not concern their content. But the plan-as-recipe component does. An agent might not simply intend to pursue a goal *g* but to pursue it *by way of* a plan-as-recipe *r*; that is, she might pursue *g* guided by a specification of some of the steps that she expects to take toward it. The intention with a plan-as-recipe plays a somewhat different role in organizing the agent's deliberation and conduct than an intention directed at the same goal without a plan or via a different one. To this extent, the recipe is part of the individuating content of the intention.

However, this qualification of the content does not support OAC. First, the change does not restrict the goal, which still needs not to be formulated in terms of one's own actions. Second, a recipe need not be formulated in terms of the agent's own actions. Sometimes it only specifies intermediate nonagential goals on the way to *g*.

4.4

Putting these considerations together, we might conclude that none of the ways in which intentions differ from aimings make a difference to the

[12] A conceptually sophisticated agent might have *second-order* intentions about the efficacy of her first-order intentions. For instance, she might become aware of failures of her executive capacities and address them by intending to improve her rate of executive success. By engaging in the self-policing of her executive capacities, she acquires *new* goals that make reference to the role of her intentions in securing executive success. But she does not thereby change the content of her first-order intentions. In addition, the second-order intentions about the success of one's first-order intention do not have as a goal the securing of their own conditions of executive success (if not in the indirect sense in which they themselves might benefit from the self-policing of executive capacities that the second-order intentions might put in place).

structure of their content. Intentions are just a more psychologically and normatively sophisticated form of aimings, but they retain the basic logical form of the individuating content of aimings. Therefore, by analogy with *(A)* in 3.2, I maintain that the canonical form of intentions is propositional rather than agential:

(I) S intends that *g*

where *g* stands for a state of affairs in the role of the goal of intending as an executive attitude.

Are there restrictions on the acceptable goals of intentions? The restrictions parallel those imposed on the content of aimings. At most, there might be an *upper* limit: the goal must be in principle attainable or, more restrictively, its attainment should not be *too* remote from the exercise of the agent's executive powers (and the remoteness, like for aimings, allows for executive success even when the agent relies on nonantagonistic waiting, monitoring, and the intermediation of other agents, see 3.4). But this limit does not impose the lowest possible restriction to the agent's own actions.

Likewise for *de se* features. The executive powers whose exercise underpins intending are necessarily *de se*. The subjects of intentions are none other than the loci of possession and exercise of these executive powers. But this constitutive relation between agents and their executive powers only makes the path to the intention's success *ab se* rather than the intention's object *de actu suo*.

4.5

The only modification to the logical form warranted by the distinctive features of intentions is the introduction of a recipe component. Some intentions have the following form:

(I+r) S intends that: (by way of *r*) *g*

where *r* stands for a plan-as-recipe, the specification of some of the instrumental steps that the agent expects to take in the pursuit of *g*.[13]

The recipe's instrumental character is indicated by the parenthesis. Unless *r* is entirely composed of necessary means to *g*, it is possible genuinely to *achieve g* even if the agent does not follow the recipe (not to be confused with the accidental or deviant obtaining of *g*, which does not count as an achievement). In this sense, the agent might be executively

[13] For a similar suggestion about the role of plan components in the content of intentions, see Mele (1987: 326).

successful in carrying out the more generic intention *(I)* directed at the same goal but without the specification of a plan of action. (For recipes that are partly constitutive of the goal and incorporated into it, see 5.4.)

4.6

To sum up, in this section, I argued that there are no differences in logical form and the *de se* involvement between simpler aimings and intentions. Aimings and intentions share these features in virtue of their common executive character. Their only differences concern (a) the rational norms that govern them and the specific ways in which the executive powers are called upon to meet these norms; (b) the possible presence of the recipe component in intentions. None of these differences, however, support OAC for intentions.

5. ACTION AS THE OBJECT OF INTENTION

5.1

Rejecting OAC does not imply that one's own actions might not be the object of one's intentions. One might actually argue that they are the paradigmatic objects at least of *future-directed* intentions, which ordinarily seem directed at one's own future actions. A future-directed intention to ϕ at f seems to have as its genuine object the action of ϕ-ing, which is to be initiated at the future time f and whose inception marks the transition from the mere intending to ϕ to the actual ϕ-ing.

This reading might seem obvious at first, but it needs to be handled carefully since it might induce a misleading picture of the relation between intending and acting, a picture that exaggerates their differences by taking the intention to be directed at the action as a truly *distinct* piece of conduct. This distinction is perfectly in order when an agent intends that *another* agent does something. For one intends that the other agent acquires a separate intention and carries it out successfully as a matter of a separate course of action. But as I am going to argue, a similar distinction is problematic for ordinary intentions directed at one's own actions.

In this section, I will discuss how best to understand the relation between an intention to ϕ and one's ϕ-ing. This discussion will reinforce my case against the OAC but also offer a diagnosis of its intuitive appeal. In addition, it will allow a more fine-grained characterization of the *de se* elements of intention.

5.2

As soon as one acquires the intention that *g*, one puts oneself under various rational pressures, including the continuous demand to secure the possibility of eventual success and, when appropriate, to make actual progress toward it. What is specifically required to meet these demands depends on one's particular circumstances (including one's present and expected skills, opportunities, and information). At times, one has to take specific steps, including making particular bodily movements and using specific tools; at other times, one might simply take advantage of favorable conditions and let the natural course of events unfold unperturbed. At times, one might have to engage in particular deliberations about implementation and coordination; at other times, one might automatically implement prior plans and policies, or let habits determine one's conduct. Discharging the rational pressures of intention is a matter of the agent's continuous "intelligent guidance" toward *g*, which requires a mix of antagonistic interventions, nonantagonistic monitoring, and the management of attention and deliberation. This mix of bodily and mental events is what I will call a *Course of Active Intelligent Guidance* (CrAIG, henceforth) directed at *g*. Responding to the rational pressures of the intention that *g* is thus a matter of engaging in the appropriate CrAIG.

Although a CrAIG is produced by exercising one's executive powers, not all portions of the CrAIG need to be "actions" in the sense of *antagonistic* bodily interventions. In the limiting case, an agent might secure that *g* simply by monitoring the favorable unfolding of a natural course of events that eventuates in *g* without requiring any antagonistic intervention. In this case, there is no action in the narrow sense of some antagonistic intervention, but the achievement is still the agent's *doing*.

5.3

Imagine that I intend that I *be* at a party tonight but I don't care about *how* I get there (this includes not caring about getting there by antagonistic interventions of mine; it would be fine, say, if someone were to take me there while I am asleep). However, given my circumstances, it is reasonable to expect that some antagonistic intervention of mine is required. Yet, I do not need to commit to any specific implementation. My intention only commits me to exercising my executive powers toward *g*, i.e. to engage in a CrAIG directed at *g*—whatever shape this CrAIG might take in response to the specific demands imposed by my present and future circumstances until the goal is either achieved or abandoned.

Imagine now that my goal is still simply to *be* at the party, but I also plan on driving there as the most reasonable or likely means to my goal. Not only does my intention include a recipe (my driving) but it might also be expressed in its terms as "I intend to drive to the party," even if the driving is only instrumental to my goal (that is, "I intend that: (by way of my driving) I be at the party"). The driving-as-recipe provides the default focus for the organization of my executive capacities. As I embark in the CrAIG directed at *g*, I exert these capacities toward *g* in large part by being oriented toward my driving. As a result, I am going to engage in an episode of "driving" as a course of action with a characteristic mix of antagonistic interventions (including the use of various tools and the performance of distinctive bodily movements), nonantagonistic monitoring, and correlated management of deliberation and attention.

A plan-as-recipe offers a default yet revisable orientation for especially salient albeit instrumental stages of an otherwise *continuous* CrAIG directed at *g*. The CrAIG does not necessarily consist only of my driving, nor does it necessarily start or end with my driving. My actual driving should not be equated with the executive success of the intention. But my starting to drive marks the point where the progress toward *g* begins to unfold as "intended" in the sense of "according to the plan-as-recipe;" the point where the course of active intelligent guidance directed at my being at the party begins to take its expected shape.

5.4

There are cases where the recipe becomes part of the goal. For instance, I intend to be at the party *only* by my driving (maybe, I want to impress the partygoers by showing off my new car and I do not care to be at the party otherwise). In these cases, the intention takes this form:

(I+gr) I intend that: *g*-by-way-of-*r*

Although *g* might be a goal by itself, here it is only a portion of the goal, which now also includes some of the means to *g*—the recipe *r*. When so, the inception of *r* marks a more momentous transition in the CrAIG: when *r* begins, the *achievement* of the intention begins as well. From that moment on, the intention is "in achievement," so to say. But this transition does not mark the acquisition of a new intention or the inception of a novel CrAIG. It is not a transition between the intention and a genuinely separate action. Rather, the transition is only the beginning of the *internal* culmination of the original CrAIG, the inception of its *finale*.

Oftentimes, the recipes that are part of the goal are what might be called "performances": specific and characteristic combinations of mostly antagonistic interventions in the form of certain bodily movements and the distinctive uses of specific tools—for instance, playing a piano sonata or dancing the tango. The beginning of the performance marks the inception of the achievement but it is still a transition *internal* to the CrAIG that began when the intention was first acquired (which might occur well in advance of the performance's inception).

Whenever the execution of a recipe begins (whether it is instrumental to or constitutive of the goal), there is no break in the continuity of the agent's active guidance toward the goal. The execution of any of the actions included in the recipe does not mark the termination of the intending and the inception of a distinct piece of conduct, the acting. It is only an internal transformation of the CrAIG's shape, as required by the dynamics of intentional progress toward *g*.

To claim that there is a single course of active intelligent guidance that begins with the acquisition of the intention and continues until the intention is given up, voided, or carried out is not to deny that this course of guidance has distinct stages. These stages might be quite different from each other given that the specific demands on the agent's executive capacities might vary widely as she progresses toward *g*. For instance, if she uses her planning abilities well, earlier stages of a CrAIG tend to demand few, if any, antagonistic interventions. At the outset, many well-planned projects require minimal, if any, interference with the natural course of events. The agent might only need to monitor for the persistence of currently favorable circumstances. At the earlier stages, the specific effects of the intention tend to be limited to the framing and organization of further practical deliberation rather than antagonistic interventions in the outer world. Conversely, later stages tend to keep the agent busier with antagonistic interventions since more interference is usually needed at the later stages to secure that the course of events eventuates in the intended state of affairs. But this is only a simplified illustration of a typical but not necessary progression of the demands imposed by an intention. The distribution of these demands can differ widely (for instance, the agent might be kept busier earlier rather than later, or cycle through the different stages).

The transitions between the stages can vary from the smooth and subtle to the abrupt and sudden, depending on the changes in the demands imposed by the circumstances. In any event, these transitions do not mark a hard and fast metaphysical boundary between two utterly *distinct* kinds of practical engagement with the world—intending vs. acting. When describing the unfolding of an intentional pursuit, the distinction between merely intending and actually acting is between two stages of an underlying unitary process of active intelligent guidance, stages whose boundaries need not be hard and fast.[14]

5.5

The temptation to think of the acting stage as a distinct item to which the intention is directed might arise from the possibility of engaging in the same kind of conduct independently of the future-directed intention. It is often possible to imagine circumstances where one might initiate the action of ϕ-ing without any prior intention directed at it. But from this it does not follow that in acquiring a prospective intention to ϕ at f one aims at initiating a distinct piece of conduct in the future. The prospective intention and its correlated CrAIG are not successfully carried out by "passing the baton" to another intention and its correlated CrAIG at f. What occurs at f is rather a transition to a different stage within the same intention and CrAIG. It is a matter of *internal* transformation dictated by the dynamic unfolding of the pursuit of the goal g to which the ϕ-ing is directed.

When one acquires the prospective intention to ϕ at f one is—so to say—"stretching over time" the pursuit of g by bringing its inception forward in time. One is not preparing in advance for a *distinct* future undertaking. Rather, one is advancing the time when one first acquires the goal g and, thereby, puts oneself under the rational pressures distinctive of that particular intentional pursuit. In turn, if one is rational, this is also the time when one begins discharging these pressures by engaging in a CrAIG directed at g.

5.6

We are now in a position to better appreciate what it means for an intention to be directed at an action as a genuinely distinct item, as it happens when one intends the action of *another* agent. The latter intention is directed at the initiation of a *separate* course of active guidance. The target of my intention that *you* ϕ is the success of your distinct CrAIG directed at your goal. Your ϕ-ing is not a stage of my CrAIG. When an action is a genuinely distinct goal, the inception of the action marks a break in the continuity of intentional guidance rather than a transformation induced by its internal dynamics.

[14] One might think of the two stages of intending and acting as akin to phase-sortals, and to their transitions as akin to metamorphosis (See McDowell (2010)).

In the first-person case, the continuity of active guidance is what normally precludes taking one's own actions as genuinely distinct items targeted by one's intentions. For one's ϕ-ing to be a genuinely distinct object of one's intention, either (i) one intends only to *prepare* for ϕ-ing without being already committed to its actual pursuit, or (ii) one is alienated from one's future self and treats one's future conduct third personally, i.e. as if it were of another agent. Unless one is in either of these two scenarios, it is paradoxical to intend to initiate an action of one's own but not to be successful at carrying it out. This is not problematic, instead, for the genuinely distinct action of another agent (or of an alienated future self).[15] This is another way to show the lack of separation between having an intention and the normal first-personal engagement in the correlated CrAIG.

Under normal circumstances, when one's own actions are presented, both in thought and speech, as the objects of one's intentions, they are not the distinct targets of one's intentional agency. Rather, they are just the descriptions of the more specific shapes—that is, of the characteristic patterns of bodily movement, tool-use, monitoring, attention management, and appreciation of the situation—that the CrAIGs are supposed to take at some crucial stages in their unfolding (whether as a matter of instrumental recipes or of goal-constituting performances).

5.7

Let's consider one last attempt at defending OAC. One might concede all the points I made but claim that there is still a sense in which the object of intention is necessarily one's own action. The argument would revolve around the necessary *ab se* character of executive success. As argued above (3.4), executive success can be secured only via the agent's exercise of her own executive capacities: achievement is necessarily *ab executione sua*, so to say. Hence, any achievement amounts to the culmination of an actual *doing* on the agent's part, that is, the culmination of a CrAIG directed at *g* (i.e. of a sequence of exercises of the agent's own executive capacities that eventuates in the nondeviant obtaining of *g*). Why can't we claim that the object of the agent's intending is necessarily this CrAIG, which is by its very

[15] There are some special cases in which one might genuinely intend to pursue a goal *g* and yet hope, without any paradox, that one will never be in a position to actually succeed at carrying it out. For instance, I am fully committed to go the hospital if I get a life-threatening injury but also hope that I will never have to carry out this precautionary conditional intention. If I can have some control on the antecedent, I might even intend to avoid that it ever comes true (that I ever get a life-threatening injury), see Ferrero (2009a).

nature always of her own? In other words, according to this suggestion, the object of intentions is necessarily the agent's own doing, in the broad sense of "doing" that encompasses the various modes of intelligent guidance.

The problem with this suggestion is that it indicates as the object a *formal* and *generic* target. The course of action that leads to the achievement is not something to which one can aim in the substantive and specific way in which one aims at a particular goal *g*. As the necessary object of the intention, the CrAIG in question is nothing other than the very operation of an executively successful intention; it is the same as the successfully completed process that constitutes the "perfection" of one's intending that *g*. A formal target cannot be the individuating goal of an intention, since there is nothing specific to it. The executive capacities are oriented toward substantive targets, not toward executive success as such. In intending that *g* one necessarily "aims" at the formal target as well, at the achievement as the successful culmination of one's own doing. But in this purely formal sense of aiming, it is uncontroversial but also uninformative to claim that in intending one aims necessarily at the doing of one's own that would amount to the executive success of that intention.

5.8

Time to take stock. In the first part of the paper, I discussed whether OAC could be supported on the basis of considerations on the individuating content of intentions and their conditions of success. I started by considering whether OAC might hold of the simpler kind of executive intentions—aiming. I argued that it doesn't. Intentions are a special kind of aiming. Their distinctive features, however, do not make a difference to the structure of their individuating content. Hence, I argued that OAC does not hold of intentions either.

I then moved to a distinct but related question. Could action at least be the paradigmatic target of future-directed intentions? I argued that this is not so, if the action is understood to be the *separate* target of the intending. The actions of other agents are genuinely distinct pieces of conduct that can be made into the proper object of one's intentions. But one's own actions do not normally play this role. In the first person mode of full temporal identification, there is a deeper continuity between intending and acting. This continuity is deeper than it might appear at first, especially if one were to read the logical structure of intentions out of the grammatical structure of ordinary expressions of intentions with their infinitival complement, since the latter seems to suggest that one is targeting one's actions as distinct pieces of conduct.

My conclusion supports a picture of diachronic agency that takes as fundamental the unity of what I call courses of active intentional guidance (CrAIG), which correspond to sequences of exercises of the agent's planning executive capacities directed toward a particular goal. According to this picture, intending is not directed at acting. Rather, intending is a matter of engaging in courses of active intentional guidance, some *stages* of which we describe as actions or activities. These stages provide a focal point for the organization of the courses of active guidance (whether as central recipes or as goal-constituting performances). Further development of this picture must be left to another occasion, but let me notice that this conclusion is a somewhat surprising twist in the debate about OAC. The picture that I have just sketched has some affinity with the one championed by some on the basis of the very views that I have rejected: the infinitival reading of the object of intentions and the defense of OAC (See Thompson 2008, Boyle and Lavin 2010, and Moran and Stone 2009). Hence, more work needs to be done to explore the implications of my rejection of OAC on those additional issues where the fate of OAC is supposed to bear, such as the status of the causal theory of action, of the "guise of the good" theory, and the nature of shared intentions (see 1.1).

6. DEGREES OF *DE SE* INVOLVEMENT

6.1

Although I have argued that the content of intentions is neither necessarily nor paradigmatically cast in terms of one's own actions, I do not deny that intending necessarily involves one's own agency. This necessary involvement is a matter of the metaphysics of executive attitudes, which are necessarily of their own agent, both in ownership and exercise. Any executive attitude is necessarily (but also trivially) *de executione sua*: possession of an executive attitude is a matter of the agent's exercise of her relevant executive powers. This *de se* involvement of agency is fundamental but also unspecific. This is just the *ab se* dimension of execution (see 3.4), a dimension that does not impose any formal restriction on the objects of executive attitudes.

This fundamental but generic degree of *de se* involvement of agency is not reflected in the individuating content. It is rather implicit in the reference to the subject of the executive attitude: in a first-person attribution of intention, we might say that it is implicit in the "I" of the "I intend."

There are two other degrees of involvement of one's agency in intending. They concern the role that the agent *might* play in the individuating content of her intentions. The agent might plan on playing an instrumental role in the pursuit of *g*. That is, she plans on relying on her taking certain specific steps (say, her ϕ-ing) toward *g* but she is open to the possibility that someone else might take her place in promoting *g*. Alternatively, an agent might take specific exercises of her own agency to be constitutive of her goal—putting herself under a rational pressure to prevent anyone else from taking those steps instead.

Here is how the three degrees of *de se* involvement of agency appear (in boldface) in the logical form of intentions as formulated in the first person:

(1st) **I** intend that: *g*
(2nd) I intend that: (by way of **my** ϕ-ing) *g*
(3rd) I intend that: *g*-by-way-of-**my**-ϕ-ing

It is only in the first degree that one's own agency is necessarily involved, although in the form of the *generic* exercise of one's executive capacities. By contrast, one's specific agential involvement in the recipes and goals of one's intentions is not required. These goals or recipes might actually involve other agents, both instrumentally and constitutively (for instance, I might intend that: (by way of *your* ϕ-ing) *g*; or I might intend that: *g*-by-way-of-*your*-ϕ-ing).

6.2

One might be involved to the second and third degree in a temporally "alienated" form. Let's imagine that, out of a concern for my own health, I intend to work out tomorrow. My intention is that my body undergoes the strenuous exercise. However, the satisfaction of my desire to be healthy does not require that I work out *directly out of* my prior intention rather than just *as a result* of it. Given that I expect that tomorrow I will be so lazy that I might irrationally abandon my intention, I can still carry out my intention by setting up a pre-commitment device that manipulates me into working out tomorrow in spite of my reluctance to do so at that time.

In this case, my intention involves me to the third degree (the goal is my performance rather than someone else's) even if tomorrow I need not endorse the considerations that supported my original intention. I might act in response to that intention in the same way as *another* agent might be cajoled by it. If so, my lack of full identification with myself in the past makes my working out a *distinct* action, which is the genuine object of my prior intention. This action is the culmination of a separate CrAIG guided by a new intention, the intention to work out that I acquire tomorrow as a result of the manipulating effect of my prior intention.

Standard intentions are not of this alienated kind. We ordinarily take ourselves to continue to identify over time, to continue to embrace and sustain the same intention throughout its unfolding and, thereby, to engage in a single continuous CrAIG without self-directed goading, cajoling, or manipulating. Hence the strongest degree of *de se* involvement is that of *full temporal identification* rather than that of mere (and potentially alienated) temporal identity.[16]

6.3

Although the second and third degrees of *de se* involvement are not necessary, it is very common for ordinary intentions to take these forms.[17] Hence, many of our recipes and goals are formulated in terms of specific manifestations of our own agency (usually in the nonalienated form). I surmise that this explains why ordinary expressions of intention usually take the infinitival form. For two reasons:

First, verbal expressions help characterize the specific ways in which the agent's executive capacities are involved in pursuing her goals. Nongeneric verbs of action describe distinct and characteristic patterns of bodily movement, tool-use, monitoring, attention management, and appreciation of the situation. Ordinary expressions usually make explicit the communicatively most salient elements of the intention, in terms of (some aspects) of its recipe and/or goal. But we should not expect these expressions necessarily and fully to articulate the content of the attitude.

Second, usually the intention's most salient elements concern the agent's exercise of her executive capacities in the mode of full temporal identification. Hence it is often pleonastic to make explicit the subject of the infinitival

[16] To mark this stronger form of first-personal transtemporal relation, one might introduce an *augmented* quasi-indicator, S^{**}. Hence, "S intends that: S^* works out tomorrow," leaves it open that one might manipulate one's future self, whereas "S intends that: S^{**} works out tomorrow" doesn't.

[17] The existence of the various degrees of *de se* involvement raises one important question. Even if many of our ordinary projects involve by default the agent to the third degree, should this involvement matter to us? Are there *intrinsically personal* projects—projects that cannot be pursued but *de se* to the strongest degree? And if there are, should we care about them? (See Perry 1976, Whiting 1986.) Reflection on the nature of the fundamental first-degree of *de se* involvement does not seem to help with these questions. The fundamental form of *de se* involvement makes it metaphysically impossible for the *source* of agency to be but particular agents involved in the *immediate* exercise of their own executive capacities. What does it take for this involvement to extend over time? That is, to extend over its immediate and momentary exercises? Is some temporal extension metaphysically required by the nature of agency itself? (See Burge (2004).) And if not, what does it take to secure this extension? And is it worth it? These questions cannot be addressed in this paper, but the discussion of the object of intentions and their relations to the first person makes them particularly vivid. (For some initial considerations about what makes extended intentional agency valuable to us, see Ferrero 2009b.)

complement. In saying that "I intend to ϕ," it is implicit that I am talking about my own ϕ-ing. But this does not imply that the agent is *necessarily* involved in the recipes and/or goals of her intentions. The only necessary *de se* involvement is that of the first degree, which is not reflected in the content of the attitude, but in its subject. The standard grammatical form of the expressions of intention in the infinitival form lends some initial support to the "own action condition," but—as I hope to have shown—these expressions, although perfectly in order in their everyday use, are a misleading guide to the logical form of intentions.[18]

REFERENCES

Baier, Annette (1970). "Act and Intent." *Journal of Philosophy* 67: 648–58.

—— (1976). "Mixing Memory and Desire." *American Philosophical Quarterly* 13: 213–20.

Boyle, Matt and Lavin, Doug (2010). "Goodness and Desire." In Sergio Tenenbaum (ed.), *Desire, Good, and Practical Reason.* (Oxford: Oxford University Press), 171–91.

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

—— (1997). "I intend that We J." In R. Tuomela and G. Holmstrom-Hintikka (eds.), *Contemporary Action Theory*, Volume 2 (Dordrecht: Kluwer), 49–63.

Burge, Tyler (2000). "Reason and the First Person," In C. Wright, B. Smith, and C. Macdonald (eds.), *Knowing Our Own Minds.* (Oxford: Oxford University Press).

—— (2004). "Memory and Persons." *Philosophical Review* 112: 289–337.

—— (2009). "Primitive Agency and Natural Norms." *Philosophy and Phenomenological Research* 79: 251–78.

Castañeda, Hector-Neri (1972). "Intentions and Intending." *American Philosophical Quarterly* 9: 139–49.

—— (1975). *Thinking and Doing.* (Dordrecht: Reidel).

Ferrero, Luca. (2009a). "Conditional Intentions." *Noûs* 43: 700–41.

—— (2009b). "What Good is a Diachronic Will?" *Philosophical Studies* 144: 403–30.

Frankfurt, Harry (1978). "The Problem of Action," In Harry Frankfurt, *The Importance of What We Care About* (Cambridge: Cambridge University Press), 69–79.

Gustafson, D. F. (1986). *Intention and Agency.* (Dordrecht: Reidel).

Harman, Gilbert (1986). *Change in View: Principles of Reasoning.* (Cambridge, MA: MIT Press).

[18] Thanks to Michael Bratman, Rahul Kumar, Matthias Haase, David Shoemaker, the audience at the NOWAR meeting in 2011, and two anonymous referees.

—— (1993). "Desired Desires." In R. G. Frey (ed.), *Value, Welfare, and Morality.* (Cambridge: Cambridge University Press).

Kapitan, Thomas (1995). "Intentions and Self-Referential Content." *Philosophical Papers* 24: 151–66.

Kenny, Anthony (1992). *The Metaphysics of Mind.* (Oxford: Oxford University Press).

McDowell, John (2010). "What is the Content of an Intention in Action?" *Ratio* 23: 415–32.

Meiland, Jack W. (1970). *The Nature of Intention.* (London: Methuen).

Mele, Alfred R. (1987). "Are Intentions Self-referential?" *Philosophical Studies* 52: 309–29.

—— (1992). *The Springs of Action.* (Oxford: Oxford University Press).

Moran, Richard and Martin Stone (2009). "Anscombe on Expression of Intention." in C. Sandis (ed.), *New Essays on the Explanation of Action* (New York: Palgrave Macmillan).

Perloff, Michael (1991). "STIT and the Language of Agency." *Synthese* 86: 379–408.

Perry, John (1976). "The Importance of Being Identical." In A. Rorty (ed.), *The Identity of Persons* (Berkeley, CA: University of California Press),. 67–90.

Roth, Abraham (2000). "The Self-Referentiality of Intentions." *Philosophical Studies* 97: 11–51.

Searle, John R. (1983). *Intentionality.* (Cambridge: Cambridge University Press).

Setiya, Kieran (2011). "Intention." In Edward N. Zalta (ed.), *The Stanford Encyclopedia of Philosophy*, Spring 2011. <http://plato.stanford.edu/archives/spr2011/entries/intention/>.

Stoutland, Frederick (2002). "Critical notice of Bratman's Faces of Intention." *Philosophy and Phenomenological Research* 65: 224–38.

Thompson, Michael (2008). *Life and Action.* (Cambridge, MA: Harvard University Press).

Tuomela, Raimo (2005). "We-intentions Revisited." *Philosophical Studies* 125: 327–69.

Velleman, J. David (1997). "How to Share an Intention." *Philosophy and Phenomenological Research* 57: 29–50.

Whiting, Jennifer (1986). "Friends and Future Selves." *The Philosophical Review* 95: 547–80.

Wilson, George M. (1989). *The Intentionality of Human Action.* (Stanford, CA: Stanford University Press).

4

Regret, Agency, and Error*

Daniel Jacobson

"How much better if it had been otherwise." Bernard Williams characterizes this proposition as the *constitutive thought* of regret in general. Although he never explicates the notion, presumably this is a thought that must somehow be credited to an agent, though perhaps not consciously entertained, in order for her to be in that emotional state. According to Williams (1976: 27): "In this general sense of regret, what are regretted are states of affairs, and they can be regretted, in principle, by anyone who knows of them." Thus one can regret unfortunate states of affairs to which one bears no special relationship. Although Williams only briefly characterizes generic regret, before narrowing his focus to those cases directed at one's own action—which he calls *agent-regret*—it's easy to think that he has gotten off to a false start by describing even the more general state too broadly.[1]

Since the thought supposedly constitutive of regret is available to anyone, it follows that, in Williams's view, everyone has equal standing to regret any unfortunate event. He recognizes no difference between agent and spectator when it comes to what they can (aptly) regret. Yet one might

* This paper issues from an ongoing collaborative project undertaken with Justin D'Arms. Many of its arguments were developed in discussion with him, though responsibility for the paper's faults falls to me alone. I am grateful to Bruce Brower, Sarah Buss, Rae Langton, Michael McKenna, Peter Railton, Tim Scanlon, David Shoemaker, two anonymous referees, and audiences at the New Orleans Workshop on Agency and Responsibility, Syracuse University, and the University of Michigan for helpful comments on previous drafts. My greatest debt is to Bernard Williams and Michael Stocker whose work not only inspired this paper but sparked much of my interest in moral philosophy. This paper was supported by a grant from the John Templeton Foundation.

[1] It should be noted that Williams' primary concern is with the narrower notion of agent-regret. The problems I will raise concern agent-regret too, but this initial complaint reflects mostly a semantic (and hence superficial) difference with Williams: we are working with different conceptions of regret. But it should become clear that we are not merely fighting over words.

think for instance that Neville Chamberlain, the architect of the 1938 appeasement of Hitler in Munich, who returned to England triumphantly declaring "peace with honour," was peculiarly well situated to regret that policy in light of the German blitzkrieg of Poland in 1939. By contrast, Winston Churchill famously announced, on Chamberlain's return from Munich: "You were given the choice between war and dishonour . . . you chose dishonour and you will have war." Obviously Churchill thought it would have been much better had Britain confronted Hitler earlier and under more favorable conditions. As Williams understands regret, Churchill and Chamberlain both have reason to regret the Munich agreement—but this seems odd. Indeed, Williams can hardly claim that Chamberlain has *more* reason to regret than does Churchill, and hence that his feelings should be more intense, since even in retrospect it was surely Churchill who believed more strongly that another course of action would have been preferable. Yet agent-regret might seem to solve this problem on the cheap, since, by definition, it is an emotion that Chamberlain but not Churchill could feel about the Munich appeasement.

However, I will argue that the deep worries in this neighborhood cannot be met in this way, by designing new emotions to meet a philosophical job description. In particular, I reject the method Williams employs in characterizing generic regret and then constructing agent-regret out of it. This method of type-identifying emotions is characteristic of the *cognitivist theory of the emotions*, which identifies a constitutive thought, belief (or perhaps some weaker propositional attitude) in which is a necessary condition for having any given emotion.[2] Williams does just this by explicitly identifying a constitutive thought of regret and implicitly suggesting another for agent-regret. Perhaps having this thought isn't sufficient for being in a state of regret, since it might need to be accompanied by the right sort of feeling. Nevertheless, Williams (1976: 27) expressly claims that "in principle" anyone who can make the judgment can have the emotion and—since the judgment he associates with regret expressly pertains to states of affairs—"anyone who knows of them" can make the relevant judgment. While I will not argue directly against cognitivism here, I do want to use this discussion to illustrate two serious problems with the

[2] Such cognitivist theories come in several varieties, most notably neo-stoicism (Solomon 1988, Nussbaum 2001), judgmentalism (Taylor 1985), and quasi-judgmentalism (Greenspan 1988, Roberts 1988). These theories share the defining propositions methodology of type-identifying the emotions, though they differ on whether the thought must be accepted or merely entertained, and whether it is wholly or only partly constitutive of the emotion.

theory, and to suggest an alternative conception of how emotions are to be type-identified and when they are rational.[3]

The first problem is evident in Williams's characterization of generic regret by way of its constitutive thought, without regard to whether the emotional states that can be expressed with this thought resemble each other in their affective and motivational aspects. As a result, his notion of generic regret turns out to be a genus rather than a species of emotion. It includes not just agent-regret and what Williams refers to as the regret of the spectator, but also such disparate sentiments as guilt, shame, sorrow, and anger—all of which can (sometimes) be manifested in the thought, "How much better if it had been otherwise." Philosophers can type-identify emotions however they like, of course, but not all such constructions will be fruitful and some prove misleading. The defining-proposition methodology of cognitivism threatens to obscure matters by conflating some importantly different emotions, as well as other practical attitudes that lack the distinctive affective and motivational features of emotion. So it is with agent-regret, which inherits this problem with generic regret and has another significant problem of its own.

This second problem reflects a more general difficulty facing the cognitivist theory of emotion: it threatens to create chimerical emotions whenever a sentiment's underlying motivational component—its *action tendency* (Frijda 1986)—conflicts with its posited constitutive thought. This problem emerges dramatically for agent-regret, because the rationality conditions implicitly given by its constitutive thought contradict Williams's claims about his two primary examples. In light of what Williams says about agent-regret, its constitutive thought must be: "How much better, in some respect, if I had done otherwise." Notice that this thought concerns the *outcome* of one's agency rather than any *decision* about acting; and that it does not imply any conviction that things would be better otherwise, all things considered. I will suggest that the chimerical emotion here is regret accompanied by the thought: "While I regret doing what I did, I endorse doing it again in similar circumstances." This would make good sense if directed at a bad but unlucky outcome of one's action, but it would be confused if directed at one's decision. Hence part of my dispute with Williams concerns whether regret is fundamentally about bad outcomes or bad decisions, loss or error. A closely related disagreement

[3] The issue of what is meant by calling an emotion rational will be taken up presently. In order to follow the literature under discussion, I will use the term "rational" to refer to what D'Arms and I elsewhere called its fittingness, in order to disambiguate this judgment from other kinds of assessment of the appropriateness of emotions.

with both Williams and Michael Stocker concerns whether one can rationally regret taking (what one considers) the best option available.

Nevertheless, their discussion of regret and its cognates exhibits an important virtue, which I greatly respect and wish to maintain: its psychological realism. Since I find this one of the characteristic virtues in the work of both Williams and Stocker, my two principal foils here, I should begin by noting my admiration for the psychological insight that informs their moral philosophy. Two facets of this realism are particularly significant for present purposes, what I'll call the *agency principle* and the *residue principle*, both of which concern the relationship between acting and feeling. These two principles are illustrated respectively by Williams's two paradigms of agent-regret: the lorry driver who blamelessly kills a child, and Agamemnon who chooses to sacrifice his daughter in order to save his becalmed fleet.[4]

Consider first the agency principle: the claim that *we are connected, through our feelings, to even the involuntary aspects of what we do*—as opposed to what simply happens around us. Williams has in mind things for which we are merely causally responsible, but in virtue of what we intentionally did. He illustrates this insight with his famous lorry driver case. The driver who blamelessly kills a child can surely be expected to feel something different than does a mere spectator to the accident, though neither is culpable for the tragedy. This core insight is one of his primary motivations for differentiating the "regret of the agent" from that of a mere spectator. Whereas Williams insists that the lorry driver feels rational agent-regret, I will argue that he can be expected to feel irrational but praiseworthy guilt. This is not to deny that the driver might have another bad feeling, directed at the state of affairs or even his role in bringing it about; but this emotion should not be considered a form of regret either. Although I thus reject both Williams's claim that the driver feels regret and that this emotion is rational, I think he grasps an important kernel of truth. The agency principle, understood as an empirical rather than a normative claim, is a psychological insight with profound philosophical implications. But the further claim that such predictable feelings are rational needs to be considered separately, as does the question of exactly what emotion we should expect from the driver in this scenario.

[4] Before considering these scenarios, however, it's worth noting that they are chosen for their philosophical interest. This strategy proves double-edged. In trying to gloss an emotion—that is, to understand its distinctive concern—I contend that one should look first to commonplace rather than novel cases. Though philosophers tend to focus on philosophically interesting examples, the more mundane cases are also more perspicuous, and unusual cases can be misleading—often for the very reason that they're interesting.

Second, both Williams and Stocker observe that sometimes when we face conflicting values or obligations, which cannot be jointly satisfied, some residue of bad feeling (or emotional "remainder") can be expected *even when we have done the best we could.*[5] This residue principle contains a crucial psychological insight as well, though I will describe it differently and draw different conclusions from it than do Williams and Stocker. Williams (1965: 181) adduces as an example Agamemnon, who "ought to discharge his responsibilities as a commander, further the expedition, and so forth . . . [but] ought not to kill his daughter." We are of course to imagine that he can satisfy either obligation but not both. Williams insists further that both incompatible obligations bind Agamemnon, and that therefore it is rational for him to regret whatever he does, despite thinking he has most reason to do it. In fact, Williams and Stocker go so far as to claim that anything Agamemnon may be able to do in this tragic situation would be wrong.

While I do not accept this conclusion, the cases on which I will focus concern conflicts of value rather than obligation, so as to avoid issues of guilt and wrongness. The problematic claim here, developed most perspicuously by Stocker, is that in some such conflicts of value it will be rational to regret choosing what one knows is the better option: the greater good or the lesser evil. Let us call this *residual regret.* As with the agency principle, one can accept the residue principle as an empirical insight without granting the rationality of residual regret. I will argue, analogously, that some of these cases don't involve regret but another emotion, which may or may not be rational; and others are cases of irrational regret that is nevertheless, in a different sense, an appropriate emotional response. This normative corollary to the residue principle, concerning the putative rationality of residual regret, plays a crucial role in a recent debate over conflicts of value, to be discussed later. But this argument obscures the insight at the heart of the residue principle: the phenomenon of *loss.* I grant that something significant can be lost in making the better choice, and that this loss can indeed justify residual bad feeling in various respects.

A similar problem plagues discussion of both the agency and residue principles. The problem lies not in the principles themselves but with the

[5] Sometimes this remainder is supposed to be a new obligation arising out of the old (forsaken) one. But that kind of remainder poses no problem to those who think that the lesser of two conflicting (prima facie) obligations does not bind, and hence that violating it would not be wrong. It is easy to grant that when I miss our lunch date for some obviously good reason, I incur an obligation to explain myself to you, to reschedule at your convenience, or even to pick up the tab. But these new obligations can be seen as issuing from the inconvenience I've caused you, however justifiably, rather than from any wrongdoing on my part.

conception of regret (or agent-regret) these philosophers adopt, and the normative conclusions this conception leads them to draw about its rationality. I contend that agent-regret is a chimerical emotion, an amalgam of two fundamentally different sentiments with disparate conditions of rationality: guilt and regret. But if we adopt a more narrowly focused conception of regret and pay closer attention to the different respects in which emotions can be evaluated, that will allow us to accommodate these authors' psychological insights while avoiding certain philosophical excesses. Specifically, the agency principle does not imply that it is rational to regret tragedies for which one was merely causally responsible; and the residue principle does not imply that it is rational to regret making the correct choice. Both these conclusions should be rejected.

1. REGRET AS A SENTIMENT

As a purely semantic point, it must be granted that the word "regret" can be used extremely broadly. However strange it sounds for a present-day historian to say that she regrets the Munich agreement, she could make use of what Williams elsewhere calls the thoroughly adverbial sense of the word to say: "The Allies regrettably chose appeasement at Munich."[6] Even if any unfortunate situation can be called regrettable, though, should we really conclude that everyone has equal standing to regret it? It will be more perspicuous to focus on when it is apt to ascribe regret to an agent, rather than on what could be called regrettable but might just as well be called unfortunate. Indeed, if "regrettable" and "unfortunate" are not synonymous, when used in this broadest sense, then it is because regret involves *someone's* agency or choice, whereas misfortune need not. This linguistic observation suggests that Williams's notion of generic regret, which is directed at states of affairs rather than actions, is a misnomer. The fact that the storm did so much damage is more felicitously termed unfortunate than regrettable. Although in my view all regret is agential, one can certainly judge *another* person's action regrettable, meaning thereby that the agent should (rationally) regret it.

Perhaps the best lesson to draw from these reflections is that ordinary language is especially misleading with respect to emotion terms. Even if the adverbial usage counts as a literal ascription of regret, certain familiar

[6] Williams (1973b: 112, fn. 1) writes of the "thoroughly adverbial" account of pleasure, on which anything that can be done pleasurably counts as a form of pleasure.

"expressions of regret" could not qualify by any plausible standard.[7] Many of these broad uses of emotion terms should not be understood as ascriptions of any actual emotion, understood (as is standard) as an occurrent, affect-laden, object-directed state. Understanding regret as an emotion helps explain the oddness of placing Chamberlain, Churchill, and the imagined historian in the same position with respect to regretting the Munich agreement. I will use these three characters to explicate the distinction between *evaluative judgments*, *emotions*, and *sentiments*.

The historian could write a speculative alternative history, detailing how much better it would have been had Hitler not been appeased but confronted in 1938, while making this conjecture dispassionately.[8] Though she thereby has the thought supposedly constitutive of regret, she is unlikely to be seriously bothered by it. If any unwanted feeling keeps her up at night, it is more likely anxiety about how her work will be received than obsessive thoughts about what might have been, if not for Munich. Indeed it would be neurotic for her to be greatly pained by disastrous decisions in the distant past to which she bears no special relationship. This is a fundamentally *aretaic* assessment of when emotional responses are admirable or not. So the historian makes an evaluative judgment without having the emotional response characteristically (but not inevitably) associated with it. Neither the intimacy nor the disparity between evaluation and emotion should be surprising. When one has the emotion, one typically makes the characteristic judgment; and when one makes the evaluative judgment, one sometimes has the emotion. The latter connection is weaker because various contextual features such as one's place in the scenario, one's temperament, and even temporal factors play a crucial role.

By contrast with the historian, Churchill was *pained* by the Munich agreement precisely because he thought appeasement so much worse than the alternative. Undoubtedly Churchill had an emotional response to the event, in virtue of believing the supposedly constitutive thought. But was it regret? There is another emotion term ready to hand, which also involves being moved—both in the sense of having some feeling and being motivated—by the thought of how much better things might have been. We

[7] My favorite example comes from a postcard used by a famous scholar to reply to requests which begins, "Edmund Wilson regrets that it is impossible for him to: Read manuscripts . . . ," and then goes on to list some twenty other scholarly tasks that he was, regrettably, unable to perform. We should all be so fortunate as to have such regrets.

[8] But cf. Deigh (2000) who suggests that this assertion must issue from a Humean theory of mind on which only passions and not judgments can motivate. On the contrary, this is a commonsense observation about specific cases. Indeed, I claim to make the relevant judgment about Munich myself, without affect. It seems ad hoc to deny that I make this judgment sincerely.

might instead say that Churchill felt *dismay* over the Munich agreement (in addition to whatever else he may have felt, such as anger at Chamberlain's self-righteousness, contempt for his naivety, fear for England's future, and so forth). Whereas the historian's judgment that things would have been better otherwise is simply an evaluation, dismay and regret are emotions because they are ways of feeling. The question is whether they are the same emotion. Is the regret that Chamberlain can be supposed to have felt, in retrospect, about *his* decision to appease Hitler at Munich simply a "particularly important species" of the same emotion Churchill felt about what happened despite his objections, as Williams implies? More abstractly, is this question a matter of psychological fact, as it purports to be, or simply linguistic stipulation about "regret" and "dismay"?

At this point I need to broach two additional claims, one general and the other more specific, for which I cannot yet argue. These are substantial claims, open to falsification, and ultimately in need of elaboration and defense. The first concerns a distinction between emotions and sentiments. Although the disparate class of states called emotions have little in common (compare approval and surprise, for example), my focus here will be on a core group I'll term *sentiments*, each of which constitutes a natural psychological kind of state whose nature can be discovered empirically rather than stipulated semantically.[9] The second claim specifically concerns regret: that it is a sentiment concerned with the agent's own *errors*. If this gloss is correct, then the reason it would be odd to attribute regret over Munich to Churchill is simply that he didn't make the mistake of trying to appease Hitler—Chamberlain did. But since both Williams and Stocker's central claims about regret are inconsistent with my gloss, it would beg the question against them to assume it. For now I simply aspire to make these two claims plausible: that regret is a sentiment, and that it is best glossed as being about one's own error.

What am I claiming by calling regret a sentiment? Like other emotions, sentiments are syndromes of thought, feeling, and motivation; but unlike some members of the disparate class of emotional states, the sentiments play a distinctive role in the human mental economy.[10] Sentiments are discrete sources of motivation, which can be in tension with our overall

[9] Philosophers of the emotions as disparate as Amélie Rorty (1978: 104) and Paul Griffiths (2004) doubt that emotion counts as a natural kind. Although I agree, I contend both that the sentiments as a group, and each individual sentiment such as regret, forms a natural psychological kind. In part, this claim amounts to adopting a broad construal of the emotions and a narrow, more technical notion of the sentiments. Cf. Frank (1988), Frijda (1986).

[10] See D'Arms and Jacobson (2003, 2006). Note the unfortunate change of terminology: what we call natural emotions in (2003) become the sentiments in (2006).

beliefs and desires. For this reason, two related features distinguish the sentiments from mere emotions: they are prone to *stable recalcitrance*, and they issue in *acting without thinking* in a sense to be explicated. First, whereas emotions that are not sentiments evaporate when one disbelieves their associated judgment, the sentiments can be recalcitrant, in that an agent can be in the grip of a sentiment contrary to his better judgment.[11] Those who are afraid of flying despite knowing that their flight is safer than the drive to the airport, which does not frighten them, are prone to stably recalcitrant fear. By contrast, the emotion of resentment, understood as a form of anger that involves a critical moral judgment, cannot coexist with the belief that no wrong was done; what can persist recalcitrantly is just anger.

The phenomenon of recalcitrance reflects the epistemic aspect of the sentiments' partial encapsulation from our beliefs and desires; whereas acting without thinking reflects the motivational aspect. For instance, the sentiment of anger is notoriously insensitive to whether the behavior it motivates coheres with our overall ends. This gives anger and the other sentiments an essential role in psychological explanation of behavior that would otherwise be mysterious. These explanations are obvious, not especially insightful. They might be called shallow (as opposed to depth) psychological explanations, so long as this isn't taken pejoratively. Obvious explanations need not be false.

Consider for example the behavior of the French footballer Zidane, perhaps the greatest player of his era, in his final match: the 2006 World Cup Final. This was not only the most important match of Zidane's career, it was his last, and the hopes of the French side depended almost entirely on him. Yet when an Italian player taunted him, Zidane responded with a flagrant and illegal head-butt, certain to get him thrown out of the match, thus ending his career in selfishness and disgrace. He thereby ruined any chance of attaining what had to be his greatest desire: leading France to a World Cup victory in the final match of his spectacular career. Instead he got the immediate gratification of revenge at the cost of defeat and ignominy. Why would anyone behave so irrationally—that is, so contrary to his overall beliefs and values? The answer should be obvious: he was in a rage and acted on the characteristic motivation of anger, namely retaliation. Sentiments thus explain the phenomenon of acting without thinking in this

[11] See D'Arms and Jacobson (2003) for discussion of recalcitrance and why there isn't recalcitrant "fear of flying," understood as an emotion, only recalcitrant fear. The crucial claim here is that whenever an emotion figures in such an explanation, some underlying sentiment does the real explanatory work.

specific sense: the agent does not consider how his action coheres with his aims.[12]

In short, the sentiments are a core class of emotions that can be found across cultures and times, albeit with some variation in their eliciting conditions, which have some characteristic motivational tendency.[13] We can thus say, with Aristotle, that anger is about undeserved slights—even though what counted as a slight for Aristotle differs from what a modern youth would consider a diss (i.e. show of disrespect). Whereas any combination of thought, feeling, and motivation counts as an emotion, the sentiments figure essentially in psychological explanations and predictions of behavior. Only some of the varied states that get called emotions will count as sentiments, while others cut across different sentiments. For example, we have stipulated that one can be dismayed in various ways, all of which can be expressed by a thought such as, "How much better if it had been otherwise"; but sometimes such dismay will involve sorrow, anger, or shame rather than regret. Although we can say that dismay is an emotion that includes various sentiments expressed by that constitutive thought, this way of speaking can be treacherous. Each of these sentiments has distinctive content and issues in different characteristic behavior. Hence what explains why Jack lashed out in his dismay, while Jill tried to make amends and Jan withdrew completely, is that Jack's dismay was a form of anger, Jill's of guilt, and Jan's of shame. What I'm calling dismay resembles the emotional manifestations of Williams's notion of generic regret, but I contend that regret is best understood as a sentiment.

Whatever Chamberlain actually felt after the blitzkrieg shattered the illusion of "peace in our time," there is a familiar psychological syndrome that he might have suffered but Churchill could not. Only Chamberlain could have had a bout of the sentiment I will hereafter simply call regret: the syndrome of painful feelings of self-reproach focused on his blunder and its disastrous consequences, accompanied by the wish to undo the error and the intention to act differently next time. Part of the regret syndrome thus involves a motivation toward what might loosely be called *policy change*, although in this case there could be no question of Chamberlain getting another chance. (Indeed, the motivation should be expected even when it is futile and manifests itself only in symbolic gestures and fantasy;

[12] This is not always to condemn such action, even when it conflicts with considered judgment, much less to bemoan the fact that we humans have such discrete, fast-and-frugal motivational systems in addition to slower and more deliberative ways of thinking (Gigerenzer and Todd 1999).

[13] See Ekman (2003) on the universality of facial expression of the sentiments (or basic emotions).

this partly explains the social importance of such gestures and the psychological importance of such fantasy.) This characterization is loose because in some cases the regretted act does not accord with any previous policy: either it was done without thinking; or not as an expression of a general policy; or even, in cases of weakness of will, contrary to one's policy. In all these cases, though, regret has a motivational component concerning one's intentions for future action. The lessons learned from regret involve dwelling painfully, perhaps obsessively, on some (putative) mistake and the decision leading up to it.[14] In sum, our historian makes a dispassionate evaluative judgment, and Churchill's painful dismay is an emotion, but only Chamberlain's hypothesized regret counts as a sentiment. Moreover, agent-regret will prove to be chimerical because it manifests itself in disparate sentiments, which can be in tension both with Williams's characterization of the state and with his paradigm cases: the lorry driver and Agamemnon.

2. WHAT IS AGENT-REGRET?

I have suggested that regret is a familiar, painful sentiment focused on error and motivating the intention to act differently. This gloss is admittedly contentious, since Williams and Stocker expressly assert that to regret an act, even rationally, need not involve repudiating it as mistaken. As Williams (1976: 31) writes:

> Regret necessarily involves a wish that things had been otherwise, for instance that one had not had to act as one did. But it does not necessarily involve the wish, all things taken together, that one had acted otherwise. An example of this . . . is offered by the cases of conflict between two courses of action each of which is morally required, where either course of action, even if it is judged to be for the best, leaves regrets—which are, in present terms, agent-regrets about something voluntarily done. We should not assimilate agent-regret and the wish, all things taken together, to have acted otherwise.

My proposal for how to understand regret closely resembles the conception (of agent-regret) Williams rejects, but it is important to note that I am *not* assimilating regret and the wish to have acted otherwise. One can find oneself regretting something without wishing, all things considered, that

[14] The word "putative" does important work here, since there is no guarantee that the action genuinely regretted was really mistaken. Indeed, when things go badly, one can expect that the reasons against doing what was done will become especially salient, and the reasons in favor will appear correspondingly diminished.

one acted otherwise. When this is really regret about the choice rather than dismay about the outcome, however, I hold that it is *recalcitrant* regret, which is irrational by the agent's own lights. Hence the criterion for the rationality of regret is whether or not one erred. That is one of my central disagreements with Williams and Stocker.

Recall that Williams coins agent-regret out of his capacious conception of regret in general, with the additional restriction that it must focus on the agent's own actions and their consequences, and the insistence that this need not be an all-things-considered wish. He thus implies that agent-regret has a constitutive thought along the lines, "How much better, in *some* respect, if I had done otherwise." Moreover, he (1976: 27–8) writes:

> The sentiment of agent-regret is by no means restricted to *voluntary* agency. It can extend far beyond what one intentionally did to almost anything for which one was causally responsible in virtue of something one intentionally did. Yet even at deeply accidental or non-voluntary levels of agency, sentiments of agent-regret are different from regret in general, such as might be felt by a spectator, and are acknowledged in our practice as being different. The lorry driver who, through no fault of his, runs over a child, will feel differently from any spectator, even a spectator next to him in the cab, except perhaps to the extent that the spectator takes on the thought that he himself might have prevented it, an agent's thought.

Here Williams seems to be making a psychological claim about what we can expect an ordinary, decent person in the unfortunate driver's position to feel: something different from what a mere spectator would feel. He also claims, plausibly, that we would expect the driver's bad feelings to move him to make some gesture of reparation to the victim. As Williams (1976: 28) notes, "The lorry-driver may act in some way which he hopes will constitute or at least symbolize some kind of recompense or restitution, and this will be an expression of his agent-regret." While a spectator would likely be moved to console the victim's parents, it would be strange for her to offer restitution for the accident.

Moreover, although Williams grants that people are prone to "irrational and self-punitive excess" in this area, he insists on the propriety of some such expression from the driver. Despite the driver's (stipulated) blamelessness, he claims that such a response would not only be normal but proper, indeed rational. "[I]t would be a kind of insanity never to experience sentiments of this kind toward anyone," Williams (1976: 29) writes, "and it would be an insane conception of rationality which insisted that a rational person never would." Thus the leading thought of his discussion of the lorry driver case—with which he introduces the notion of agent-regret—is what I've called the agency principle. As he (1993: 92) notes elsewhere, it is an "utterly familiar fact that what happened to others through our own agency can have its own authority over our feelings, even though we brought it about involuntarily." This seems especially likely when the outcome is negative.

Williams moves quickly from the predictability of such a response to the insanity of being entirely immune to it, and then to the rejection of any conception of rationality on which a rational person would be so immune. Surely "insanity" is hyperbole, and what he really means is something more like inhumanity. Any decent person will display such tendencies. Thus Williams commits himself to the rationality of the lorry driver's regret (which I will deny), on the grounds that it would be unacceptable to adopt standards of rationality for the emotions that fail to cohere with human psychology (which I accept). One issue between us concerns the specific respect in which one should endorse this regret. This will be the topic of Section 3. Another question, to which I will now turn, is whether we should accommodate the agency principle by positing a novel emotion, namely agent-regret.

Williams' notion of agent-regret seems to differ from guilt, which is typically taken to be rationally restricted to the voluntary—or, at any rate, the blameworthy.[15] As Williams (1993: 89) himself observes, "What arouses guilt in an agent is an act or omission of a sort that typically elicits from other people anger, resentment, or indignation." This is just what distinguishes the blameless driver from the negligent or malicious one: no one has any justified complaint against him. Although Williams (1976: 27) considers agent-regret to be a species of the more capacious bad feeling he calls regret in general, he also claims that it "is not distinguished from regret in general solely or simply in virtue of its subject-matter," namely one's own action, but is also characterized by a different form of expression in motivation: the desire to make reparations. But note that this is *exactly the motivation characteristic of guilt.*[16] In the throes of guilt, one feels like one owes a debt, and that discomforting feeling issues in the desire to make reparations. Thus Williams's notion of agent-regret seems precariously placed between (generic) regret and guilt. What then is the point of giving the emotion we can expect from the lorry driver a distinct name? And why call it agent-regret?

[15] It is worth noting that, years later, Williams (1993: 93) tosses off the remark that the lorry driver's agent-regret is "psychologically and structurally a manifestation of guilt." On one hand, this suggests that my treatment of the case does not differ as much from Williams as first appears; on the other, this seems an admission that even his restricted notion of agent-regret conflates cases of regret with those of guilt.

[16] See, e.g. Gibbard (1990), Tangney and Dearing (2002), Baumeister et al. (1994, 1995).

This question becomes even more puzzling when we consider what some other philosophers have written about the lorry driver case. Gabrielle Taylor describes the driver's emotion simply as guilt and concludes that guilt concerns causal rather than moral responsibility. "Causal responsibility is the type that is sufficient for guilt, and that much is also necessary," Taylor (1985: 91) writes; hence, guilt "cannot be vicarious, and feelings of guilt similarly cannot arise from the deeds or omissions of others," not even one's children or compatriots. But almost no one else agrees with her, and for good reason: counterexamples seem rampant.[17] Taylor's gloss of guilt forces her to hold that vicarious guilt is not merely irrational but impossible, which would make survivor guilt and liberal guilt into counterexamples to her theory—if they are indeed forms of guilt, not just states casually called by that name.[18] Of course, she could always posit another emotion to accommodate these phenomena rather than denying their possibility outright. But to posit something like "liberal compunction" to cover cases of vicarious guilt is an ad hoc maneuver designed to salvage the theory by stipulation. Worse, she would no longer be talking about guilt, the natural kind of psychological state, but a subset of its instances chosen precisely to fit her claims. This renders tautological a claim—that guilt cannot be vicarious—which seems to be put forward as an empirical observation. It would be better to advance a claim that is neither empirical nor semantic but normative: that causal responsibility is necessary for guilt to be rational. This would make survivor guilt possible, even explicable, but not rational; and that combination of claims seems quite plausible. Moreover, Taylor's implicit suggestion that we should expect the lorry driver to feel guilt rather than some variety of regret is exactly right.

Though we can evaluate claims about the sentiments, those about constructed emotions are much more difficult to assess, because they seem precariously placed between empirical hypothesis and semantic stipulation. Consider Marcia Baron's (1988) discussion of agent-regret and the lorry driver, in which she claims to detect a difference between regret and agent-regret that Williams fails to note. Baron (1988: 262) writes: "Agent-regret has an ethical dimension which plain regret lacks. Agent-regret is felt toward the sorts of things which, if done deliberately, would properly occasion guilt." Baron clearly means to be talking about the same state as

[17] See Greenspan (1992) for counterargument. When Taylor is being more careful she allows that the requirement is, rather, that an agent must believe himself causally responsible for something in order to feel guilty over it. Even so, many counterexamples arise.

[18] The fact that these states motivate apology and attempts at reparation, even where there is no causal responsibility, is a reason to think that they are indeed forms of guilt.

Williams, not just expropriating an emotion term he coined. Whereas for him agent-regret is something like *regret about what I did*, she thinks it more like *guilt about the involuntary*. What Baron gets right is that we should expect ordinary decent people to feel guilt over terrible results of their actions even when they are involuntary. This is what I've called Williams's agency principle, but with the connection to guilt made explicit. Even so, not all of Williams's paradigms of agent-regret look like guilt over the involuntary—the Agamemnon case does not.

The ethical dimension Baron claims to find in agent-regret also involves a threshold of significance, below which she thinks it impossible to have the emotion. In contrast with what she calls plain regret, Baron (1988: 262; emphasis added) holds that "agent-regret simply toward my failure to buy myself a pair of shoes that I liked *does not seem possible*. What I failed to do was too trivial and too remote from ethical concerns." Yet surely a sufficiently shallow person could regret forgoing the shoes, and it is unclear how Baron can deny that this is agent-regret except by definition. Although she can restrict her use of "agent-regret" stipulatively to cases that are of ethical import and reach a threshold level of significance, it is hard to see what this gains. Not only would it trivialize her (seemingly empirical) impossibility claim, she would then simply be talking about a different emotion than is Williams, while taking herself to be correcting him about its nature. Again the normative hypothesis seems more tenable. Why not say instead that only a little regret is rational over minor errors?[19]

Moreover, Williams does not merely "fail to note" that agent-regret must have an ethical dimension, as Baron contends; he denies it outright. Her claim contradicts his initial characterization of the emotion as a refinement of regret in general. If one can (rationally) feel agent-regret toward anything about which one can truthfully think, "How much better if I had done otherwise," then surely this can be thought and felt about missing out on a nice pair of shoes at a great price. Indeed, Williams expressly acknowledges that one can feel agent-regret over mistakes that harmed only oneself. Since he thinks that the lorry driver's emotion will be expressed in the desire to make restitution, he has to adduce a different sort of expression for these cases. "Granted that there is no issue of compensation to others in the purely egoistic case," Williams (1976: 32–3) writes, agent-regret can only be expressed in "the agent's resolutions for his future deliberations." This,

[19] This point illustrates another infelicity of ordinary language: some words do have connotations of great significance, such as "remorse" and "misery," both of which suggest great intensity. But surely this is just a semantic point and, what is more, it does not hold about "regret."

of course, is the motivation characteristic not of guilt but regret—that is, regret understood in my sense, as a sentiment that focuses on error.

However, Williams *does* fail to notice that this expression too will be out of place in certain cases of putative agent-regret: those where the agent does *not* repudiate her action, either because she chose the least-bad alternative in a bad choice situation, or because she made the right decision but things happened to turn out badly nonetheless. In those cases there is no opportunity for either reparation or repudiation, so agent-regret cannot seem to find any appropriate expression in action. And yet the thought supposedly constitutive of agent-regret is very much in order. It would be much better, at least in some respect, if I had done (or had been able to do) otherwise. These cases strongly suggest that the most striking psychological phenomenon in the neighborhood is not one that can attach to any way things might have been better. Rather, it is Stocker's notion of *loss*, which we will consider in more detail in Section 4.[20]

The problem for Williams is that his posited notion of agent-regret has no coherent expression in behavior. Worse, his paradigms have two quite different expressions, which cut across the distinction between egoistic and altruistic cases, and which happen to be precisely the motivations characteristic of regret and guilt. This strongly suggests that Williams has conflated (at least) two distinct sentiments under his broad class of generic regret, characterized in terms of a constitutive thought. The problem infects agent-regret as well, since that subspecies is sharpened out of the more capacious notion by the stipulation that it concerns one's own action. Since both guilt and regret *typically* focus on one's own action, agent-regret too will cut across the distinction between the two sentiments. This explains the disparate action tendencies found in the different cases Williams groups together as agent-regret. Such chimerical emotions, which are not distinct sentiments but constructed groupings, cannot sustain empirical generalizations. Although philosophers can type-identify emotions however they like—and hence there is no point in denying that agent-regret counts as an emotion—only some emotions turn out to be sentiments which explain human behavior that would otherwise be mysterious, such as Zidane's enraged head-butt.

Although there are no definitive arguments in this domain, I can offer some reasons to recognize regret as a sentiment that plays an important and

[20] When one makes the right decision but it turns out badly, one loses out on something genuinely valuable, with only the cold comfort of having chosen correctly. And when one makes a hard choice, where there are different reasons on both sides of the question—even if the reasons for one choice clearly outweigh the other—then we often do painfully feel the loss of the goods rationally forgone.

familiar role in the explanation of human behavior. The function of regret is to focus us on our ends and to help us avoid temptation and distraction. The regretful agent recalls his action painfully and without bidding; his regret motivates him to undo the action if possible, and to change policy for the future. This gloss of regret locates its concern in putative mistakes. If this syndrome of thought, feeling, and motivation seems familiar and useful in shallow psychological explanation, then that would give us reason to posit a sentiment in the neighborhood of the pre-theoretic notion of regret. Further evidence that regret is a sentiment should be found in the presence of stably recalcitrant regret, and in regretful agents being prone to acting without thinking. First, to be in a state of recalcitrant regret would be to be motivated to undo some previous action despite believing it to have been the best available option. Do people excoriate themselves over things they've done—in particular, things that have turned out badly—despite knowing that there wasn't anything mistaken about their decision-making? Surely they do. Second, for regret to issue in acting without thinking would be for it to move people, in the throes of regret, to punish themselves counterproductively over their putative mistake, even to the extent of harming themselves more than did the regretted error. This too seems quite familiar.

Notice that because regret has obvious similarities with another sentiment, guilt, the two distinct states are often confused. An agent in the throes of guilt also focuses on his action painfully and without bidding, but guilt primarily functions to placate the anger of others. It has a social function of reconciliation that regret does not. Although guilt is typically glossed as being about wrongness, I actually think this too simple; but for present purposes I will adopt the simpler gloss in order to avoid unnecessary complications.[21] Recalcitrant guilt would arise when someone feels guilty despite thinking he has not acted wrongly. The lorry driver case is just such an example. Guilt also motivates acting without thinking, the other hallmark of a sentiment, when people are moved to make reparations contrary to their overall set of beliefs and desires. Are guilt-ridden people prone to apologize, for instance, even when so doing actually hurts the wronged person even more? Again this seems all too familiar.[22]

[21] I think guilt should instead be given a dual aspect interpretation, as about either impersonal wrongness or personal betrayal. This dual-aspect view can explain the rationality of guilt in situations like Agamemnon's, where the demands of duty conflict intractably with our deepest personal commitments and relationships. See D'Arms and Jacobson (1994), Baumeister et al. (1994, 1995), Tangney and Dearing (2002).

[22] While some of these cases call for depth psychological explanation, such as when the confession is best seen as a further act of hostility rather than atonement, surely this is not always the case.

One reason why guilt and regret are so often confused—or conflated by way of an artificially constructed emotion like agent-regret—is that *typically* when one feels guilty, one will also feel regret (though not vice versa). That is to say, painful thoughts of having done wrong, which motivate the desire to make reparations and thereby placate the anger of others, tend to be accompanied by painful thoughts that one has made a mistake, which issue in repudiation of the action and resolve to act differently in the future. Nevertheless, guilt need not be accompanied by regret, since it seems psychologically possible to think that one acted immorally without repudiating what one did.

Because agent-regret conflates two distinct sentiments, it is a misleading emotion category that obscures more than it illuminates. However, its construction is not a mere mistake; it is motivated by the insights of the agency and residue principles. Yet Williams's attribution of agent-regret to the lorry driver carries some tacit normative implications which, when made fully explicit, point in a different direction than he suggests. He aims to find an emotion that it would be rational for the lorry driver to feel, but which *cannot* rationally be felt by a mere spectator to the event. I've already granted the plausibility of Williams's agency principle, which holds that the driver can be expected to have bad feelings disproportionate to those of a mere spectator. If we assume, further, that the difference between agent and spectator is not merely a matter of degree but of what kind of emotion they can be expected to feel—as Williams clearly implies by contrasting agent-regret with the regret of the spectator—then his motivation for coining agent-regret becomes clear.

Calling the lorry driver's emotion agent-*regret* suggests that he need not be blameworthy in order to feel it rationally. According to Williams, one can feel rational regret over actions with bad outcomes for which one is merely causally responsible, in virtue of what one intentionally did—such as driving the lorry blamelessly. And calling it *agent*-regret suggests that a mere spectator cannot feel it rationally, since he could feel that only if he thought himself to be a participant in the action (which really would be insane). Agent-regret thus seems to fit the job description perfectly. Yet I will argue in the following section that this is a specious victory. Nothing important is gained by attributing agent-regret rather than guilt or dismay to the lorry driver, and Williams's (hedged) commitment to defending the rationality of the lorry driver's emotion is understandable and plausible but ultimately proves misguided.

3. THE RATIONALITY OF REGRET

Williams does not argue for the rationality of the lorry driver's regret so much as peremptorily reject any conception of rationality that denies it. Since there are several different respects in which philosophers have claimed emotions to be appropriate—as rational, admirable, fitting, and so forth—it isn't immediately obvious what Williams thereby rejects. We might differentiate between these forms of criticism by saying that an emotion is *unreasonable* whenever there is most reason not to feel it, that it is *not admirable* when it conflicts with what a virtuous person would feel under the circumstances, and that it is *unfitting* when it misrepresents its object. It is possible that an emotion could be fitting, in this sense, despite neither being what a virtuous person would feel nor what one has most reason to feel. Because cognitivist theories of the emotions posit a constitutive thought to each, they have an obvious affinity for accounts of fittingness in terms of the truth of that proposition. Although I here follow Williams and Stocker in speaking of the *rationality* of emotions such as regret, I take them to be talking about fittingness. And I understand them to hold that an emotion is irrational when its (supposedly) constitutive thought is false.[23]

Williams suggests that there would be something wrong with the lorry driver if he felt merely what a spectator could feel over the tragedy for which he was causally responsible. While acknowledging the likelihood of excessive self-reproach in such situations, he insists that a driver *should* be prone to an initial bout of bad feeling that expresses itself in the impulse to make (even symbolic) reparations. But he also expresses some proper caution about this conclusion. Williams (1976: 28) writes:

> Doubtless, and rightly, people will try, in comforting him, to move the driver from this state of feeling . . . to something more like the place of a spectator, but it is important that this is seen as something that should need to be done, and indeed some doubt would be felt about a driver who too blandly or readily moved to that position.

I can accommodate Williams's leading thought about this case, which corresponds to the modest, empirical version of the agency principle alongside an ethical norm endorsing those feelings. We can expect even the blameless driver to feel worse than a spectator, and to want to make some kind of restitution; moreover, *in some sense* he should have such

[23] See D'Arms and Jacobson (2000). An additional reason not to use "rational" in this sense arises from cases of misleading evidence, where the epistemic flavor of that term is inapt. But we will not be considering such cases here.

feelings. Yet Williams wants to defend the driver's emotion as rational regret, and this motivates him to coin the notion of agent-regret so as to construct an emotion whose constitutive thought is true of the driver but not the spectator. The blameless driver can think "How much better if I had done otherwise" without imagining that he is at fault, but the spectator cannot think this—supposing that she could not have averted the accident. But if the driver simply feels regret or guilt (or most likely both), then his sentiment is irrational: in fact he made no mistake and did nothing wrong. Nevertheless, I can accept that it should be endorsed as admirable.

The first thing to note about Williams's discussion of the case is that he displays some ambivalence about what the driver should feel. While he suggests that we would feel "some doubt" about a driver who was not inclined to some participatory feeling, he also thinks that we should try to talk him out of such feelings and into something more like what a spectator would feel. Someone might hold this view for merely strategic reasons—such as the consideration that there is nothing to be gained by blaming himself—but I see no reason to attribute to Williams the mistake of conflating the reasonable and the rational emotion. Instead, I suggest that his core normative conviction is aretaic. One would have doubts about the *character* of a driver who too easily adopted the attitude of a spectator, when he was the causal agent of some disaster. People tend to beat a hasty retreat from responsibility for bad outcomes, often under the cover of the passive voice. In light of this dubious tendency, we should acknowledge that there are grounds for self-scrutiny whenever one is the causal agent of a bad outcome.[24] An admirable person in a real-life situation would be prone to search for grounds of self-reproach: he cannot help himself to the stipulation that he is blameless.

In fact, Williams utilizes something very much like this thought when he considers what someone should feel about cases of conflicting obligations. There he (1965: 173) insists that the standards for "an admirable moral agent cannot be all that remote from that of a decent human being." He goes on to claim, plausibly, that decent human beings will be disposed to feel badly about doing what is best in some such situations, for instance when that entails breaking a serious promise. I suggest an analogous conclusion about the lorry driver case: a decent person who is causally responsible for some disaster will be prone to guilt, at least initially, though he should be movable to something more like the position of a spectator. (And he will be so movable, if he is not subject to "self-punitive excess.")

[24] I take this to be the leading thought of Rosebury (1995) on moral luck.

Moreover, if this is what decent people feel, then any notion of what an admirable agent will feel cannot deviate too far, on pain of becoming overly idealized and unrealistic. As Williams (1965: 175) holds about cases of conflicts of obligation,

> A tendency to feel regrets . . . at having broken a promise even in the course of acting for the best might well be considered a reassuring sign that an agent took his promises seriously. At this point, the objector might say that he still thinks the regrets irrational, but that he does not intend 'irrational' pejoratively: we must rather admit that an admirable moral agent is one who on occasion is irrational. This, of course, is a new position: it may well be correct.

But this is precisely my view of the lorry driver scenario. It is admirable for the driver to feel guilt about the accident that is, strictly speaking, irrational. The driver's guilt attributes some grounds for blame to himself that are not present, by stipulation—a stipulation to which the driver cannot help himself. The driver's initial tendency should be considered, to paraphrase Williams, as a reassuring sign that he takes his agency seriously.[25]

This solution will remain elusive, however, so long as the different forms of normative assessment are conflated with the term "rational" (or any other word). Only by differentiating between the rational and the admirable can we explicate what Williams calls the non-pejorative sense of irrationality. Moreover, once this distinction is made, it no longer seems necessary to coin an emotion that the driver can rationally feel; hence there is no need to coin agent-regret to do that job. And this, in turn, would allow Williams to acknowledge that it is irrational guilt that we can expect the driver to feel, not rational agent-regret. It also solves the problem noted earlier, that his catch-all notion of agent-regret makes it mysterious why that emotion has such disparate forms of expression in action: it sometimes issues in the motive to make reparations, sometimes to change policy, and other times to do neither. That can be easily explained once we notice that the emotion is sometimes guilt and sometimes regret, and that it is not always rational even when it is predictable, and indeed admirable, to feel.

Even if guilt can be rationally felt only over actions for which we are responsible, it can be reliably predicted in other circumstances as well. The most admirable sort of person—within the constraints imposed by a realistic moral psychology—will initially be prone to unjustified guilt when he is an innocent causal agent of disaster; but he will also be "moveable"

[25] I want to thank Michael McKenna for pressing the point that the agency principle does not operate solely under conditions of uncertainty. In my view, uncertainty is sufficient for justifying some such feelings, but I do not hold that it is necessary. Much will depend upon the context, and people can disagree about what virtue demands.

from that guilt in the face of sufficient evidence of his blamelessness. Though I claim this guilt to be irrational, I have not adopted an insane conception of rationality. It would be closer to the truth to say that I have explicated the notion that Williams refers to, approvingly but opaquely, as irrationality in the non-pejorative sense.

Suppose one accepts the distinction between sentiments and (mere) emotions, and my account of regret and guilt as sentiments. One might still wonder if there is an emotion that would be rational for the lorry driver to feel but which would not be rational from a mere spectator. After all, there are actions that are appropriate for the driver but not the spectator, so why not feelings as well?[26] Note first that, in my view, there is a sentiment (guilt) that would be in some sense appropriate (that is, admirable) for the driver to feel, at least initially—though I agree with Williams that he should not remain in that state. Our thoughts about the appropriateness of actions and especially feelings need to be focused; the question is whether any such emotion is rational in the sense we have fixed upon. My answer is yes and no. There is no such sentiment, I claim—not regret, guilt, shame, or sadness (the most obvious candidates). Each of these is either too narrow to be rational for the faultless lorry driver or too broad not to be rational for the spectator. But there is an emotion that can fit the bill—or at any rate one can be constructed. It will be useful to consider this point further.

I have used the term "dismay" stipulatively to refer to an emotion generically directed at bad outcomes, which can naturally be expressed with the thought, "How much better if it had been otherwise." By the agency principle, we are more inclined toward dismay when we have some salient causal role in the outcome, even if we made no error and thus are not at fault. Consider a case where there can be no question of culpability, unlike the lorry driver, because the outcome is so distant from one's agency. Suppose you have parked your car on the street outside your house, and it is thrown by a tornado into your neighbor's house, killing their child. You will be dismayed by this event, as no doubt will other uninvolved neighbors. The thought of how much better it would have been otherwise is available to all of you. But the agential thought, "How much better if I had done otherwise"—say, parked on the other side of the street—is available only to you. Given the permissive stance I have taken toward the category of emotions, we can use the constitutive thought methodology to coin an emotion that only you can feel: call it agent-dismay. Hence there is an

[26] I am indebted to Tim Scanlon for this way of putting the worry, which I find particularly challenging. It has been extremely helpful to me to think through my response to this challenge.

emotion that you can feel rationally but the mere spectator cannot. Yet this constructed emotion explicates nothing. We might just as well have said that although both you and the spectator can be dismayed at this terrible event, only you can (rationally) be dismayed while having the thought that it would have been better had you parked elsewhere.

Contrast dismay with regret. Dismay has no coherent motivational component; there is no characteristic action-tendency common to its disparate instances. One can be dismayed about all sorts of misfortune, and one will act and express such dismay very differently depending on the context. These actions and expressions will be explicable in terms of the underlying sentiment producing them: if I feel guilty, I will act on my dismay in one way; if I am ashamed, in another; and if I am angry, in an altogether different manner. Moreover, when dismay is recalcitrant, it will be recalcitrant as a sentiment: as regret, guilt, shame, or anger. And when it issues in acting without thinking, the action motivated will again be the characteristic motivation of some particular sentiment. Of course what I have been calling agent-dismay is precisely Williams's notion of agent-regret. But the dispute over what to call this emotion is not mere semantics, because calling it regret obscures what is distinctive about regret as a sentiment: that it focuses on our errors and motivates policy change. This discussion has also aimed to illuminate why, although it is cheap to construct new emotions by their constitutive thoughts, that approach does not advance our understanding.

4. CONFLICTS OF VALUE AND RESIDUAL REGRET

Thus far I have argued for a particular conception of regret as a sentiment that concerns error, and I have focused on the agency principle and Williams's first paradigm of agent-regret, the lorry driver. Now I propose to turn to the residue principle and a variation on Williams's second paradigm, where I will draw an analogous conclusion. The second case Williams uses to characterize agent-regret is Agamemnon's tragic choice between his daughter and his fleet, love and duty, personal and impersonal obligation. My argument for diagnosing the conflicted agent's residual bad feeling as guilt is strongest when there are incompatible obligations in play, as here, but I grant that Agamemnon might feel both guilt and regret simultaneously. In order to avoid this complication, I will focus on conflicts of value rather than of obligation, where guilt is not at issue. Suppose that someone who made a hard choice regrets his decision despite thinking it the best of his options. If my gloss of regret is correct then this emotion, like the lorry driver's, must be irrational though it might be appropriate in some other sense. (We would not want a parent to be too ready to sacrifice his child even for the best of reasons.) Such conflict of value cases present the strongest objection to my gloss of regret in terms of error.

The objection is that my conception of regret cannot make sense of the intuition, insisted upon by both Williams and Stocker (1990: 191), that in such cases it is "rational to regret what clearly is to be done." Recall that the empirical version of the residue principle predicts that we will often feel residual regret upon making hard choices between (somehow) competing values. Situations where one has strong reasons for making each of two incompatible choices can be taken to support an alternative conception of regret, suggested by Stocker, on which it is an emotion that fundamentally concerns not error but *loss*. When one makes the right decision but it turns out badly, one loses out on something genuinely valuable, with the cold comfort of having chosen correctly. And when one makes a hard choice, where there are strong conflicting reasons, then we often painfully feel the loss of those forgone goods. This much seems impossible to deny. But are they cases of regret, and is it rational?

Consider the ramifications of any conception of regret broad enough to vindicate such residual regret as rational. Stocker clearly finds the normative corollary of the residue principle intuitively obvious. "It is beyond argument," he (1990: 125) writes, "that the world as we know it gives us grounds for regret and conflict even if we do what is to be done." In order to grasp his view of regret, one must appreciate his concern with the psychological state of being *conflicted*: emotionally ambivalent. This state of psychological conflict is not justified whenever there are reasons on both sides of a choice, as is typical even of easy choices. Rather, for Stocker (1990: 87) the distinction between "rational conflict" and "conflict over unimportant matters" marks the crucial difference between choices that should be emotionally conflicting and those where one should simply forgo some good without regrets. He considers this the rationality criterion for residual regret. Hence Stocker's conception of regret is fundamentally agential, but it focuses not on error but significant loss of value.

Were regret understood as being about the kind of loss that often renders us conflicted even about doing what clearly should be done, then obviously it would sometimes be rational no matter what we do. Specifically, in certain circumstances where you have chosen the greater good or the lesser evil, it is nevertheless rational to feel regret. But in order for regret to be rational over doing what clearly is to be done, Stocker (1990: 267) writes, "[t]he agent must see something good in the lesser option not seen in the greater, which gives an evaluative reason to do the lesser rather than the greater."[27] This suggests a gloss of regret in terms of the thought, "How much better, in some respect, if I had done otherwise"—which is exactly the constitutive thought of agent-regret according to Williams.

We should not accept an account of regret that marginalizes the central scenario in order to capture the peripheral, even if the latter are more philosophically interesting. The gloss of regret in terms of this thought threatens to do just that. Since the thought is true both when the agent errs and when he makes the right choice in a hard case, the account assimilates error and loss. This threatens the idea that regret serves as a sentiment whose function involves learning from one's mistakes, because cases of loss without error should *not* motivate any change of policy. However we use the term regret, we must not lose sight of this sentiment or fail to notice its importance. Even in difficult choice situations, where we will be emotionally conflicted no matter what we do, our aim is surely to make the right decision. In deliberation we seek to avoid error and the sentiment concerned with sanctioning errors, even if a painful emotion concerned with loss is unavoidable.

Like Williams's agency principle and the lorry driver, Stocker's argument moves very quickly and with great rhetorical insistence from the predictability of an emotion to its rationality. Indeed, we ordinarily grant a presumption of warrant to our actual feelings and, especially, to what we are stably prone to feel. Hence the burden lies with those arguing against the rationality of feelings to which almost everyone would be prone. Yet recall the lesson of the lorry driver example. Even though we can predict that ordinary, decent people will feel guilty over being the causal agent of disaster despite their lack of culpability, there is an explanation ready-to-hand that undermines the rationality of such predictable guilt. This explanation invokes the deeply unrealistic idealization of such examples. Someone actually in the driver's unfortunate situation cannot be certain of his lack of culpability, let alone stipulate it. The trouble with the lorry driver case is that it is *designed* to be unusual and, in this crucial respect, deeply unrealistic.

A similar problem infects the dilemma cases. In ordinary life, we make decisions under conditions of uncertainty. Even if the reasons I'm aware of

[27] This is the core intuition behind Stocker's (1996) argument from rational regret for value pluralism. However, I do not think that the distinction between choices that justify emotional conflict and conflict over unimportant matters tracks the distinction between kinds of value. So although I accept value pluralism, I do not find this argument for it compelling, because the sort of loss that can rationally be emotionally conflicted can involve one value instantiated in qualitatively different ways or distributed to different people. In just this respect, I agree with Hurka (1996a, 1996b), but see fn. 29.

favor buying the country house rather than renting in the city, one never knows—the well could be about to run dry. The general problem with much of this literature about conflicts and dilemmas is that it focuses on the most extreme sort of cases and ignores the commonplace uncertainty under which we make decisions. I suggest we look to more mundane scenarios where we can expect to feel regret no matter what we do, such as buying a house. Psychological realism here dictates that we can expect to vacillate between buyer's remorse and renter's regret (as it were) without any new evidence. This vacillation should not be mysterious, as it has an explicable source. Our anxiety over such large but hardly heroic or tragic decisions makes the problems with whatever is currently the likely outcome more salient, as well as the attractions of the alternative likely forgone. When the likelihoods change so does the salience, even when nothing else about the choice alters.

This rather obvious point goes underappreciated in this literature. There are many circumstances in which you can reliably expect to feel regret no matter what you do, not because of some deep fact about value or morality, but because of familiar human tendencies such as those present when deliberating over buying a house. No important lesson for value theory can be drawn from the all-too-human phenomenon of shifting salience of goods and the anxiety produced by prospective regret.[28] The fact that you will feel regret no matter what you do must serve as a defeater of the presumption of warrant one normally accords to one's emotions and dispositions to feel. If I will regret whatever I do, then it would be a mistake to accord that regret any evidential weight with respect to my decision. How can I know if it is ordinary regret over making the wrong choice or residual regret over the right choice? Moreover, we can reliably predict regret over the better choice when, for reasons that could not be foreseen, the choice turns out badly. The recognizable human tendency to commit a "bad outcome, therefore bad decision" fallacy also serves as a defeater in many mundane cases concerning regret. These are strong reasons not to accept the normative corollary vindicating the rationality of residual regrets, even where we accept the residue principle predicting them.

Nevertheless, a profound insight underlies the examples Stocker and Williams consider, which might be taken to motivate a broader conception of regret than my error-focused gloss. Their insight is that there is an ethically important difference between true sacrifices and mere forgone goods. Consider this example drawn from Stocker. Suppose a man must

[28] To be clear, some very interesting results in decision theory and experimental economics might issue from such phenomena. But these are psychological not axiological conclusions.

give up his favored occupation, to which he has committed himself—perhaps it is philosophy—in order to support his family by some more mundane but lucrative enterprise. It seems he may well look back on his life and think he made the right choice, all things considered, yet feel a deep sense of loss at the personal good sacrificed for the sake of his family. By contrast, someone unable to satisfy her desire for a convertible because her family needs a minivan, although she gives up something greatly desired, would be shallow to dwell on the foregone good. This is not a point about materialism, and it can be framed in other terms. But I'm prepared to grant that the convertible could be a long-standing desire, a potential source of considerable pleasure, and not in principle an indefensible indulgence. Even so, to bemoan this hardly tragic sacrifice appears less than admirable. There seems an important distinction between significant loss and mere doing without, even when the choice was for the best in both cases. But I contend that this point marks a fundamentally aretaic distinction. There are sacrifices an admirable person can regret honorably, and those he must accept gracefully as yet another way in which the world does not bend to our wishes. If I am right then no coherent conception of regret can track Stocker's fundamental distinction, because the difference between conflicting and insignificant loss is not a matter of value but virtue.

Does my argument imply that no painful emotion over making the best choice in a situation of value conflict can ever be rational? As with the lorry driver, the answer is yes and no. No such sentiment is rational. The claim that regret is rational in such cases assimilates making a better choice that is not preferable in *every* respect with making a worse choice that has *something* going for it. In almost every choice situation there will be some reason to choose the lesser good, so this view leads inevitably to a proliferation of rational regrets.[29] But what we care about in deliberation is making the better choice: avoiding error. Even in cases where we can expect to be emotionally conflicted no matter what we choose, this is the crucial concern of the sentiment that deserves the name regret.

Stocker's key insight—that we do feel conflicted when forced to make *some* such choices and sacrifices, and it seems that in some sense we should

[29] The counterargument given by Hurka (1996a, 1996b) looks like a reductio ad absurdum of this claim. In his view, whenever you choose the best vacation out of ten good choices, for example, you should rationally regret not choosing the other nine (albeit not "to excess"). But any actual bad feeling about the unlimited number of lesser but still good alternatives would be unseemly. Hurka's (2001) view makes regret not an emotion but an evaluative judgment, which is just what he does with love as well, following Brentano (1969).

feel this way—does not apply to many cases where we sacrifice one kind of value for another, simply because the good forgone is insufficiently important. Yet it does hold about some cases where the same kind of value is at stake in both options, when the choices are qualitatively different in important respects.[30] To be seriously dismayed over genuine losses of value is compatible with being an admirable person facing an uncooperative world that demands sacrifice; whereas to be dismayed over not getting everything you want, though a recognizably human tendency, falls into the sphere of vice. As Stocker has aptly put the point: it makes you a whiner. While we could stipulate that dismay is rational only in cases of *significant* loss, this will be true by definition; moreover, the rationality conditions of this residual emotion will be parasitic on an aretaic judgment about the agent.

Williams's lorry driver scenario and Stocker's argument about conflicts of value share certain virtues and vices. Keen psychological insight underlies their core claims, particularly about human emotional propensities. These insights are sometimes obscured, though, by an insufficiently discriminating theory of emotion. The failure to differentiate adequately between regret, guilt, and dismay prevents them from drawing the right philosophical lesson from the psychological data. It is true that we can expect the decent lorry driver to feel something different from a mere spectator to the tragedy. He will feel guilt, even though his guilt is (by a stipulation unavailable to him) unfitting. It is irrational guilt that an admirable person can nevertheless be expected to feel, at least until he becomes convinced of his blamelessness. Analogously, Stocker's cases of conflicts of value reflect the insight that bad feelings are often driven by uncertainty and anxiety over our decisions. Sometimes they reflect a good forgone that deserves tribute from our feelings, but is not best understood as regret. Here we can say that bad feeling over making the best choice is either rational dismay about some good forgone or irrational regret over a correct decision—or perhaps some vacillation between the two.

[30] We would need a coherent account of the distinction between value monism and pluralism to make this argument rigorous, and I'm not sure that one can be found. See Chang (2001) for doubts along these lines. It suffices for present purposes that different varieties of sensual pleasure—which must be granted to be instances of the same type of value if anything is to count as monism—can be sufficiently qualitatively different as to justify some dismay over forgone choices. This much seems clear to me, though I am reluctant to go into detail about cases.

REFERENCES

Baron, Marcia (1988). "Remorse and Agent-Regret." In *Midwest Studies in Philosophy* vol. XIII: "Ethical Theory: Character and Virtue," eds. French, Uehling, and Wettstein. (Notre Dame, IN: University of Notre Dame Press).

Baumeister, R. F., A. M. Stillwell, and T. F. Heatherton (1994). "Guilt: An Interpersonal Approach." *Psychological Bulletin* 115: 243–67.

—— —— —— (1995). "Interpersonal Aspects of Guilt: Evidence from Narrative Studies." In *Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*, eds. June Tangney and Kurt Fischer. (New York: Guilford Press).

Bittner, Rüdiger (1992). "Is it Reasonable to Regret Things One Did?" *The Journal of Philosophy* 89: 262–73.

Brentano, Franz (1969) [1889]. *The Origin of our Knowledge of Right and Wrong*, ed. Roderick Chisholm. (London: Routledge & Kegan Paul).

Chang, Ruth (2001). "Value Pluralism." In *International Encyclopedia of the Social and Behavioral Sciences*," eds. Neil Smelser and Paul Baltes. (New York: Elsevier Press).

D'Arms, Justin and Daniel Jacobson (1994). "Expressivism, Morality, and the Emotions." *Ethics* 104: 739–63.

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

—— —— (2003). "The Significance of Recalcitrant Emotions (or, Anti-Quasijudgmentalism)." In *Philosophy and the Emotions*, ed. Anthony Hatzimoysis. (Cambridge: Cambridge University Press).

—— —— (2006). "Anthropocentric Constraints on Human Value." *Oxford Studies in Metaethics* 1: 99–126.

—— —— (2009). "Regret and Irrational Action." In *Reasons for Action*, eds. David Sobel and Steven Wall. (Cambridge: Cambridge University Press).

—— —— (2010). "Demystifying Sensibilities: Sentimental Values and the Instability of Affect." In *The Oxford Handbook of Philosophy of Emotion*, ed. Peter Goldie. (Oxford: Oxford University Press).

—— —— "Guilt and Wrongness Reconsidered." (Unpublished manuscript).

Deigh, John (1994). "Cognitivism in the Theory of Emotions." *Ethics* 104: 824–54.

—— (2000). "Symposium: The Works of Martha C. Nussbaum: Nussbaum's Defense of the Stoic Theory of Emotions." *Quinnipiac Law Review* 19: 293–307.

de Sousa, Ronald (1974). "The Good and the True." *Mind* 83: 534–51.

—— (1987). *The Rationality of Emotion*. (Cambridge, MA: MIT Press).

Ekman, Paul (2003). *Emotions Revealed.* (New York: Henry Hold and Co).

Frank, Robert (1988). *Passions Within Reason: The Strategic Role of the Emotions.* (New York: W. W. Norton and Co).

Frijda, Nico (1986). *The Emotions.* (Cambridge: Cambridge University Press).

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

Gigerenzer, Gerd and Peter Todd (1999). "Fast and Frugal Heuristics: The Adaptive Toolbox." In *Simple Heuristics That Make Us Smart*, eds. Gigerenzer, Todd, et al. (New York: Oxford University Press).

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

—— (1992). "Subjective Guilt and Responsibility." *Mind* 101: 287–303.

Griffiths, Paul (2004). "Is Emotion a Natural Kind?" In *Emotion, Evolution and Rationality*, eds. P. Cruse.and D. Evans. (Oxford: Oxford University Press).

Hume, David (1978) [1739]. *A Treatise of Human Nature*, 2nd edn., ed. L. A. Selby-Bigge. (Oxford: Clarendon).

Hurka, Thomas (1996a). "Monism, Pluralism, and Rational Regret." *Ethics* 106: 555–75.

—— (1996b). "Reply to Critics." *BEARS Symposium*, eds. Jamie Dreier and David Estlund. <http://www.brown.edu/Departments/Philosophy/bears/homepage.html>.

—— (2001). *Virtue, Vice, and Value.* (New York: Oxford University Press).

Jacobson, Daniel (1997). "In Praise of Immoral Art." *Philosophical Topics* 25: 155–99.

—— (2005). "Seeing by Feeling: Virtues, Skills, and Moral Perception." *Ethical Theory and Moral Practice* 8: 387–409.

Knobe, Joshua (2006). "The Concept of Intentional Action: A Case Study in the Uses of Folk Psychology." *Philosophical Studies* 130: 203–31.

Nussbaum, Martha (2001). *Upheavals of Thought: The Intelligence of Emotion.* (Cambridge: Cambridge University Press).

Roberts, Robert (1988). "What an Emotion Is: A Sketch." *The Philosophical Review* 97: 183–209.

Rorty, Amélie (1978). "Explaining Emotions." In *Explaining Emotions*, ed. Amélie Rorty. (Berkeley, CA: University of California Press, 1980).

Rosebury (1995). "Moral Responsibility and Moral Luck." *The Philosophical Review* 104: 499–524.

—— (1980). "Agent Regret." In *Explaining Emotions*, ed. Amélie Rorty. (Berkeley, CA: University of California Press).

Solomon, Robert (1998). "On Emotions as Judgments," reprinted in *Not Passion's Slave: Emotions and Choice.* (Oxford: Oxford University Press, 2003).

Stocker, Michael (1971). "'Ought' and 'Can'." *Australasian Journal of Philosophy* 49: 303–16.

—— (1990). *Plural and Conflicting Values.* (Oxford: Clarendon Press).

—— (1996). "Symposium: Monism, Pluralism, and Rational Regret." *BEARS Symposium*, eds. Jamie Dreier and David Estlund. <http://www.brown.edu/Departments/Philosophy/bears/homepage.html>.

Tangney, June (1995). "Shame and Guilt in Interpersonal Relationships." In *Self-Conscious Emotions: The Psychology of Shame, Guilt, Embarrassment, and Pride*, eds. June Tangney and Kurt Fischer. (New York: Guilford Press).

—— and Ronda Dearing (2002). *Shame and Guilt.* (New York: Guilford Press).

Taylor, Gabrielle (1985). *Pride, Shame, and Guilt: Emotions of Self-assessment.* (New York: Oxford University Press).

Williams, Bernard (1965). "Ethical Consistency," reprinted in Williams (1973a).

—— (1973a). *Problems of the Self.* (Cambridge: Cambridge University Press).

—— (1973b). "Against Utilitarianism." In J. J. C. Smart and Bernard Williams, *Utilitarianism: For and Against.* (Cambridge: Cambridge University Press).

—— (1976). "Moral Luck," reprinted in Williams (1981).

—— (1979). "Conflicts of Values," reprinted in Williams (1981).

—— (1981). *Moral Luck.* (Cambridge: Cambridge University Press).

—— (1993). *Shame and Necessity.* (Berkeley, CA: University of California Press).

5

Phenomenal Abilities: Incompatibilism and the Experience of Agency*

Oisín Deery, Matt Bedke, and *Shaun Nichols*

1. BACKGROUND

1.1 Introduction

Agents act. They buy detergent at the store, they go to work, they celebrate holidays, they cheat on their taxes. Sometimes we hold agents morally responsible for what they do, or what they fail to do, meting out credit and blame as the occasion merits. In typical cases, when agents act they are thought to have an *ability to do otherwise*. This is a point on which most parties to the free-will debates agree. When it comes to characterizing the ability to do otherwise and asking whether this ability is compatible with determinism, however, there is no consensus.

In the ensuing debates, the *experience* as of having an ability to do otherwise occupies a central role.[1] Many libertarians, for instance, maintain that the ability experienced is incompatible with determinism (Campbell 1951; O'Connor 1995). Of course, some compatibilists have challenged this idea (Mill 1865; Grünbaum 1952; Nahmias et al. 2004). Despite the centrality of the *phenomenology of agency* in all this, there has been strikingly little work on its characteristics. Of particular significance, there is almost no empirical work on whether the experience of agency

* We thank Eddy Nahmias, Dana Nelkin, Paulo Sousa, David Shoemaker, Uriah Kriegel, and Bertram Malle for comments on earlier drafts.We are also grateful to participants at the New Orleans Workshop on Agency and Responsibility (November 3–5, 2011) and two anonymous reviewers at OUP for suggestions and comments.

[1] For some of the literature on the ability to do otherwise, see Moore (1912); Berofsky (2002); J. Campbell (2005); Perry (2004); Vihvelin (2004); Smith (2004); Fara (2008); John M. Fischer (2008); Randolph Clarke (2009). The claim that moral responsibility requires alternative possibilities has been disputed since Frankfurt (1969). However, it is still widely contended that free will requires being able to do otherwise.

involves a phenomenology of being able to choose among alternative possibilities or whether people take their agentive experiences to have incompatibilist elements.[2]

This paper reports a series of experiments that investigates the phenomenology of agency. To anticipate, we found remarkably consistent results across three sets of studies: participants regarded their experience of the ability to do otherwise as incompatible with determinism. Now that we've spoiled any suspense, we will locate the issue in the broader literature.

1.2. The Experience of the Ability to Do Otherwise

Let us characterize determinism as follows: a statement of the facts of the world at an instant, together with a statement of the laws of nature, entail all truths about the world, including those about future human actions.[3] Granting that we often feel that we have an ability to act other than we do, for present purposes incompatibilists think that the experience as of an ability to do otherwise is incompatible with determinism, while compatibilists think the opposite.

There are a number of influential "introspectors" on both sides of this issue. John Searle is a representative incompatibilist:

> [R]eflect very carefully on the character of the experiences you have as you engage in normal, everyday human actions. You will sense the possibility of alternative courses of action built into these experiences . . . that we could be doing something else right here and now, that is, all other conditions remaining the same. This, I submit, is the source of our own unshakeable conviction of our own free will. (1984: 95)

Similarly, Keith Lehrer has claimed that the incompatibilist "accurately describes what I find by introspecting, and I cannot believe that others do not find the same" (1960: 150). Even such a paradigmatic compatibilist as David Hume (1960/1739) agrees with this sentiment when he writes, "There is a false . . . experience . . . of the liberty of indifference" (Bk. II, Part III, §II).

The appeal to an incompatibilist phenomenology plays a particularly important role in libertarianism. Many libertarians maintain both that we experience our agency as incompatible with determinism, and that this experience provides reason to think that our agency defies determinism. C. A. Campbell writes:

[2] One exception in the recent literature is a paper by Nahmias and colleagues (2004), which we discuss in Section 1.3. Although we challenge their experimental results, we are indebted to them for pioneering the investigation. We also draw on their scholarship in setting out some of the historical statements below. See also Monroe and Malle (2010).

[3] In the experiments below, we will present this idea in terms of causation to make it more intuitive and accessible.

> Everyone must make the introspective experiment for himself: but I may perhaps venture to report... that I cannot help believing that it lies with me here and now, quite absolutely, which of two genuinely open possibilities I adopt. (1951: 463)

Campbell goes on to argue that, unless we have good reason to doubt the impression that "it lies with me" which of two possibilities I adopt, we should accept the impression to reflect the truth. Timothy O'Connor makes this move as well. First, O'Connor describes the character of the experience of decision-making:

> [T]he agency theory is appealing because it captures the way we experience our own activity. It does not seem to me (at least ordinarily) that I am caused to act by the reasons which favor doing so; it seems to be the case, rather, that *I* produce my decision *in view of* those reasons, and could have, in an unconditional sense, decided differently. (1995: 196)

Next, O'Connor says that we should take these experiences to reflect something important about the *nature* of decision-making:

> Such experiences could, of course, be wholly illusory, but do we not properly assume, in the absence of strong countervailing reasons, that things are pretty much the way they appear to us?... Skepticism about the veridicality of such experiences has numerous isomorphs that, if accepted, appear to lead to a greatly diminished assessment of our knowledge of the world, an assessment that most philosophers would resist. (1995: 196–7)

A number of compatibilists have challenged the basic phenomenological claim. These compatibilists deny that we experience our agency as incompatible with determinism. John Stuart Mill, for instance, writes,

> Take any alternative: say to murder or not to murder.... If I elect to abstain: in what sense am I conscious that I could have elected to commit the crime? Only if I had desired to commit it with a desire stronger than my horror of murder; not with one less strong. When we think of ourselves hypothetically as having acted otherwise than we did, we always suppose a difference in the antecedents: we picture ourselves as having known something that we did not know, or not known something that we did know; which is a difference in the external motives; or as having desired something, or disliked something, more or less than we did; which is a difference in the internal motives. (1865: 285)

On Mill's view, the feeling of the ability to do otherwise is always contingent on our supposing that the situation prior to the decision was somehow different. Adolf Grünbaum repudiates any incompatibilist element with equal vigor:

> Let us carefully examine the content of the feeling that on a certain occasion we could have acted other than the way we did.... Does the feeling we have inform us

> that we could have acted otherwise *under exactly the same external and internal motivational conditions*? No, . . . this feeling simply discloses that we were able to act in accord with our strongest desire at that time, and that we could indeed have acted otherwise if a different motive had prevailed at the time. (1952: 672)

Grünbaum's last sentence here gestures at the payoff of denying the phenomenological claim of incompatibilist agency: if Mill and Grünbaum are right, then the feeling of being able to do otherwise is consistent with determinism, and this would undercut a crucial motivation for libertarianism.

This situation might seem to be a dialectical stalemate (cf. Fischer 1994: 84). However, these philosophers are making general claims about the nature of our experience of agency. These are empirical claims, and they can be illuminated by taking up empirical methods.

1.3. Previous Work on the Phenomenology of Free Will

We are not the first to recommend a more systematic investigation that is partly empirical. Nahmias and colleagues (2004) suggest that we find out how people actually tend to describe their agentive experience (what they call the phenomenology of free will), including their experience as of being able to do otherwise:

> Taking a cue from recent empirical work on "folk intuitions", we think the best way to understand the phenomenology of free will—if there is one—is to find out what ordinary people's experiences are like. If this is not possible, philosophers' competing introspective descriptions will remain in yet another free-will stalemate. (164)

Nahmias and colleagues undertook this task in survey studies. Their studies appear to lend some support to the idea that the phenomenology of agency is compatibilist. However, we think the studies have significant shortcomings, so let us briefly describe one of those studies, and then identify what we find lacking.

In one study, Nahmias and colleagues pitted compatibilism and incompatibilism against each other directly. The study was based on "competing libertarian and compatibilist accounts of our experience of the ability to choose otherwise" (174). Their survey asked participants to imagine (or recall) an experience of making a difficult choice:

> Imagine you've made a tough decision between two alternatives. You've chosen one of them and you think to yourself, "I could have chosen otherwise" (it may help if you can remember a particular example of such a decision you've recently made). Which of these statements best describes what you have in mind when you think, "I could have chosen otherwise"?
>
> A. "I could have chosen to do otherwise even if everything at the moment of choice had been exactly the same."

B. "I could have chosen to do otherwise only if something had been different (for instance, different considerations had come to mind as I deliberated or I had experienced different desires at the time)."

C. Neither of the above describes what I mean. (2004: 175–6)

The majority of the participants gave the response that fits with compatibilism (i.e. B).

While this study is clearly focused on an issue that divides compatibilists and incompatibilists, there are a number of limitations to the study. First, participants are told to think of a decision and then told to think something else about the decision: that they *could have done otherwise*. It is thus unclear whether their initial recollection *actually* carried with it a sense of an ability to do otherwise. So if people make compatibilist judgments about these decisions, it might be because they are considering cases in which the phenomenology of the ability to do otherwise is absent.

Second, the key question is about experiences sometime in the past, rather than present-focused experiences where the phenomenology of agency is actually present and thus presumably more accessible.

Third, Nahmias and colleagues asked participants about *difficult* decisions, and this presents the opportunity to interpret "could have done otherwise" in confounding ways. Consider Martin Luther's decision to renounce his writings or be declared an outlaw and heretic. Legend has it that, after praying and consulting with advisors for a day, he said, "Here I stand. I can do no other," thereby reaffirming his writings. Luther might have chosen B in Nahmias's survey. But if he did, we should not conclude that there is no sense of "could have done otherwise" that captures some aspect of Luther's phenomenology and that is incompatible with determinism. For Luther could have responded as he did to express his commitment to his cause, a commitment that would only change if the considerations before him and his reasons for breaking with the Roman Catholic Church presented themselves differently. This commitment-expressive meaning of "could not have done otherwise" is consistent with other senses of "could have done otherwise"—consider whether Luther thought it was up to him whether to renounce his views—that might or might not be incompatible with determinism. Difficult decisions are subject to confounds like this, so the above survey does not cleanly address the question whether there is some aspect of the phenomenology of agency that is in tension with determinism.

Fourth and last, it is not clear whether the participants really understand the intended meaning of "even if everything at the moment of choice had been exactly the same" or "only if something had been different."

We wanted to run more comprehensive studies that fix these shortcomings. The result was the following three studies, which share a common structure. First, participants were asked whether they had an experience as of the ability to do otherwise when faced with a simple decision. Next, they were given a description of determinism. Of course, we did not use the term "determinism", since that might have conjured up unwanted associations in participants. Rather, we used a technical term—"causal completeness". To address concerns about comprehension of the materials, we used the familiar psychological technique of *training to criterion*, where we asked a series of questions that tested and, if necessary, corrected, the participant's understanding of determinism. Participants who passed the training were asked about the compatibility of their experience with determinism. In study 1, this question focused on both a first-person, present-focused experience in a hypothetical deliberative context and a past-focused judgment about such a situation. In study 2, we explored the phenomenology of actual rather than imagined choices. In study 3, we tested whether epistemic phenomenology—the phenomenology of uncertainty—feels incompatible with determinism.

2. STUDY 1

2.1 Overview

In our first study, we had participants imagine a decision about whether to go left or right on a sled. In one condition, the sledding scenario was set in the future; in the other condition, the scenario was set in the past. After reading the scenario, participants in condition 1 were asked whether they had a feeling of an ability to do otherwise; participants in condition 2 were asked for a retrospective judgment about whether they could have done otherwise. Participants who affirmed feeling (or having) an ability to do otherwise were directed to the training section in which causal completeness (i.e. determinism) was explained to them. Participants who passed the training were reminded of their affirmation regarding the ability to do otherwise and asked about consistency with causal completeness.

Our prediction was that when asked about the phenomenology of imagined decision-making, participants would tend to affirm a feeling of an ability to do otherwise and also regard this feeling as incompatible with determinism; but when asked for a retrospective judgment about the ability to do otherwise, we predicted that participants would be less likely to treat the ability to do otherwise in an incompatibilist way.[4]

[4] See e.g. van Inwagen 1983 (8–13) for an overview of other uses of "can."

Method:

Participants:

Eighty-four participants were initially recruited online through the Mechanical Turk (MTurk) website.[5] The survey itself was conducted using SurveyMonkey. Two participants did not complete the survey. They were excluded from the analysis.

Materials:

Each condition had three parts.

Part 1: The ability to do otherwise Participants were presented with a vignette and a question about the ability to do otherwise. For condition 1, this went as follows:

> Please read the following passage, and answer the questions that follow as best you can:
>
> Imagine that you are sledding down a snowy path on a mountainside. Your sled has a steering mechanism that allows you to control the direction of the sled. Below you is a fork in the path with snow built up in the middle, and you can tell that, if you don't direct your sled one way or the other, the contours of the mountain will channel you and your sled either to the left or to the right.
>
> **Ability Question**
>
> Consider how things seem to you as you approach the fork in the path. In particular, consider what it's like to decide which way the sled will go.
>
> Please indicate your level of agreement with the following statement:
>
> When deciding which way the sled will go, it feels like I can either go to the left or go to the right.

Participants were asked to indicate their agreement with this statement on a 7-point scale (1 = disagree completely; 7 = agree completely).

For condition 2, the vignette was the same except that participants were asked to imagine that the sledding episode occurred many years ago. And instead of getting a response regarding a phenomenology of the ability to do otherwise, we asked them to indicate agreement with a statement about a past ability to have done otherwise: "I could have gone right instead of left."

[5] MTurk is a website supported by Amazon.com <https://requester.mturk.com/mturk/welcome> that provides users the opportunity to fill out surveys for modest compensation. Recent work indicates that the data gathered through MTurk is at least as reliable as that gathered through standard psychology pools composed of undergraduates (see Buhrmester et al.).

Part 2: Training on determinism We wanted to focus on participants who had a phenomenology of the ability to do otherwise, so only participants who indicated a positive level of agreement to the first questions (5 or higher) were directed to the training section. Here, participants were given a detailed explanation of causal completeness, summed up as follows: "According to causal completeness, everything that happens is fully caused by what happened before it. This is true from the very beginning of the universe, so what happened in the beginning of the universe fully caused what happened next, and so on right up until the present. Causal completeness holds that everything is fully caused in this way, including people's decisions."

Participants were then given two kinds of cases to illustrate the phenomenon. In one case, they were told that an earthquake fully caused the volcanic eruption at Mt St Helens,[6] and they were then told, "According to causal completeness, if we could somehow replay the entire past right up until St Helens erupted on May 18, 1980, then St Helens would once again erupt at that time. Another way to put this is to say that all the events leading up to the eruption made it so that the eruption had to happen." In another case past events, feelings and beliefs led to Obama's decision to pick Joe Biden as his running mate, and participants were told "According to causal completeness, if we replayed the past right up until Obama's decision—including everything that was going through Obama's mind—then Obama would once again make exactly the same decision. That is, all the events leading up to Obama's decision (including everything that was going through Obama's mind), made it so that it had to happen that Obama would pick Biden."[7]

We then tested comprehension of causal completeness. We first asked participants to indicate whether the following was true or false:

> According to causal completeness, St Helens would have erupted on May 18, 1980 even if there had been no earthquake.

Participants who answered "True" (the incorrect answer) were corrected, and given an explanation of the right answer. These participants were then

[6] This is an oversimplification of the geological facts which we adopted to ease the load on participants.

[7] In defining determinism—our causal completeness (CC)—as meaning "everything that happens is fully caused by what happened before it," some might think this consistent with certain indeterminist conceptions of causation. But to say that events are *fully* caused is meant to avoid this reading—being fully caused suggests that nothing extra-causal is needed to help settle events. Our examples aid the preferred interpretation. We say, e.g. that under CC, "it had to happen that Obama would pick Biden."

given a similar question to see if they had absorbed the training. If they answered incorrectly yet again, they did not move on to answer the compatibility question, as they were deemed to have insufficient comprehension of causal completeness.

Participants who passed this first kind of question either on the first or second try were given another true/false question to test for comprehension:

> According to causal completeness, if a week from now Barack Obama decides to have soda with dinner, all the events leading up to that decision will make it the case that he has to decide to have a soda with dinner.

Our objective here was to test for and correct overly weak interpretations of causal completeness. Those who answered "False" (the incorrect answer) were corrected and given another chance at a similar question. If they answered incorrectly yet again they failed the training and did not answer the compatibility question. Participants who passed both kinds of questions either on the first or second try were deemed to have adequate comprehension of determinism, and these participants moved on to the third part of the study, the compatibility question.

Part 3: Consistency After successful completion of the training, in both conditions participants were told to recall their agreement with the statement regarding the ability to do otherwise (from Part 1 of the survey). For example, in condition 1, they were told:

> Now, recall that you previously agreed with the following statement:
>
> When deciding which way the sled will go, it feels like I can either go to the left or go to the right.

Following this, they were asked the compatibility question. In condition 1, this read as follows:

> **Compatibility Question**
> Considering this previous statement and your understanding of causal completeness, please indicate your level of agreement with the following:
>
> Even though it felt like I could either go to the left or go to the right, if causal completeness is true there is something mistaken about how that decision felt to me.

In condition 2, the compatibility statement was:

> Even though I said I could have gone right instead of left, if causal completeness is true there is something mistaken about what I said.

Agreement was indicated on the same 7-point Likert scale as was used for the Ability Question, and an answer above 4 was taken to be an incompatibilist answer.

Results:

Of the thirty-four participants who started condition 1, thirty-three completed it. Of these, thirty-one indicated a phenomenology of an ability to choose among possibilities and all but four of them passed the training section.[8] The remaining twenty-seven participants gave a mean response of 4.93 on the compatibility question, which differed significantly from the midpoint of the scale, $t(26) = 2.65$, $p = .014$. That is, participants tended to interpret their agentive experience as being incompatible with determinism.

In condition 2, of the fifty participants who started the survey, forty-nine completed it. Of these, forty-seven indicated an ability to choose among possibilities and all but two of them passed the training section.[9] The remaining forty-five participants gave a mean response of 5.24 on the compatibility question, which differed significantly from the midpoint of the scale, $t(44) = 5.05$, $p < .001$.

A *t*-test comparing conditions 1 and 2 showed no significant difference, $p = .448$. So participants tended to be just as incompatibilist about retrospective judgments of their ability to do otherwise as they are about their current experience as of being able to do otherwise. This first study provides evidence that people do indeed judge that their experience of deciding is inconsistent with determinism, in the sense that the experience is somehow mistaken or nonveridical if determinism is true. It also suggests that the effect is robust across retrospective and present-focused cases.

3. STUDY 2

3.1 Overview

One major limitation of study 1 is that it involved merely imagined choices. This inserts a distance between the actual phenomenology of decision-making and judgments about that phenomenology. As a result, in study 2 we introduce conditions in which agents actually make decisions. In addition, study 1 focused on decisions that have no moral weight. We

[8] Nine participants required correction, and ultimately passed the training section. The responses of those who required correction did not differ from those who answered correctly without training ($p > .2$).

[9] Seven participants required correction, and ultimately passed the training section. Again, there were no differences between those who required correction and those who didn't.

thus added a condition in study 2 in which the decision *does* have a moral element. So this study comprises three conditions to test for any effect from actual choices or from morally salient choices. We also introduced two innovations to the study's design.

First, we wanted to make our vignette more "choicey". It struck us that in situations such as sledding down a hill often we don't have a salient experience as of deciding which way to go. We just go one way or the other. Second, we wanted to address a potential worry about our use of the word "mistaken" in the compatibility question. For example, in condition 1 from our first study, we asked:

> Even though it felt like I could either go to the left or go to the right, if causal completeness is true there is something <u>mistaken</u> about how that decision felt to me. (Emphasis added.)

One worry about this wording was that participants might misinterpret it as asking whether they were mistaken in thinking that their experiences felt a certain way, rather than as asking whether there would be something mistaken about the content of their felt experiences.[10] Our solution was to replace the above wording with a wording of the following form:

> Even though it felt like I could either choose to X or choose to Y, if causal completeness is true then I couldn't really have chosen differently than I did.[11]

With these modifications, condition 1 presented participants with an *imagined* choice among two very similar charities, condition 2 presented participants with an *actual* choice among two similar charities, and condition 3 presented participants with an *actual morally salient* choice among two charities, one for endangered trees, another for children's cancer treatments.

[10] Thanks to Lucas Thorpe for this objection.

[11] Two reviewers worried that, with causal completeness in mind (earlier described in terms of events that "had to happen") participants fix on one reading of the modal "couldn't really have chosen differently" and on that reading they give an "incompatibilist" response, whereas the description of their phenomenology might invoke a different reading of the modal that would not merit an incompatibility response. Of course, the key issue for us is whether participants feel their phenomenology wouldn't be veridical if CC were true. The question in study 3 is designed to first refer to the participants' reports on their phenomenology—that it felt like they could choose X or choose Y—and we think this helps subjects focus on *that* modal content and whether *it* would be veridical if CC were true. Further, in study 1 we ask the compatibility question using different language that avoids this worry. We get the same incompatibilist results there.

Method:

Participants:

One hundred and fifty-five participants were initially recruited online through the Mechanical Turk (MTurk) website. The survey itself was conducted using SurveyMonkey. Twenty-one participants did not complete the survey, or indicated that they had recently taken a "very similar" survey.[12] They were excluded from the analysis.

Materials:

As in study 1, this study had three parts.

Part 1: The ability to do otherwise For condition 1 of this study, we asked participants to imagine deciding between two charities for endangered trees.

> Imagine that you have $0.50 to donate. You have two options:
>
> Donate to a foundation that protects the endangered tree *Castanea Dentata.*
>
> OR
>
> Donate to a foundation that protects the endangered tree *Ulmus Dentata.*[13]
>
> These are your only two options.

Condition 2 was similar except that participants were given an *actual* choice. Participants were told that they had $0.50 to donate to one of the two tree charities. We informed participants (truly) that we would actually donate this money to whichever charity they chose. Participants read:

> You have $0.50 to donate. We, the researchers, will actually donate this money for you whichever way you decide.

Participants were then presented with the same option language as in the imagined condition, and each option appeared as a radio button at the bottom of the page.

Finally, in condition 3 we presented participants with a morally salient choice between a foundation that protects the tree *Castanea Dentata* and The Childhood Cancer Foundation,[14] on the assumption that people tend to think that saving dying children has greater moral weight than saving endangered trees.

[12] Studies 2 and 3 were run after study 1, and some of the conditions in studies 2 and 3 were run serially, so we excluded participants who indicated they had taken a very similar survey to minimize the influence of having previously taken one of our surveys.

[13] *Castanea Dentata* and *Ulmus Dentata* are the names of the American Chestnut and the American Elm, respectively. They are endangered species in North America. The charities we used were The American Chestnut Foundation <http://www.acf.org/>, and Trees Winnipeg: Coalition to Save the Elms <http://www.savetheelms.mb.ca/>.

[14] <http://www.candlelighters.ca>.

In all conditions, after being given the imagined or actual choice, participants were asked a question about the ability to do otherwise. For instance, in condition 3, participants were asked to indicate their level of agreement (on a 7-point scale) with the following statement:

> When deciding which option to choose, it feels like I can either choose to donate to the endangered tree *Castanea Dentata* or choose to donate to the Childhood Cancer Foundation.[15]

(In conditions 2 and 3, participants were subsequently required to make a choice between the charities.) As in study 1, only participants who agreed with the ability-to-do-otherwise statement proceeded to the training.

Part 2: Training on determinism The training section was the same as that used in study 1, and once again only those who passed the training proceeded to the compatibility question.

Part 3: Consistency The compatibility question was adapted for the new cases. For example, in conditions 1 and 2, participants were asked to indicate agreement (on a 7-point scale) with this statement:

> Even though it felt like I could either choose to donate to *Castanea Dentata* or choose to donate to *Ulmus Dentata*, if causal completeness is true then I couldn't really have chosen differently than I did.

Results:

Of the fifty participants who started condition 1, forty-two completed it and had not recently taken a very similar survey (three had). Of these, thirty-eight indicated a phenomenology of an ability to choose among possibilities and all but three of them passed the training section.[16] The remaining thirty-five participants gave a mean response of 5.60 on the compatibility question, which differed significantly from the midpoint of

[15] We want to forestall a potential concern about the phrasing here. Suppose I am determined to choose p. It follows that I can choose p. And you might think it further follows that I can choose p or choose q, for this follows from the simple logical principle of disjunction introduction. In that case the ability to choose p or choose q is clearly *compatible* with determinism. However, participants report an incompatibilist phenomenology as of an ability to choose p or q, which suggests that they are not reading "can choose p or can choose q" in this compatibilist way.

[16] Twelve participants required correction and successfully passed the training section. There was a significant difference in responses between those who required correction and those who didn't. Those who required correction reported that their phenomenology was incompatible with causal completeness ($M = 4.82$) but to a lesser degree than those who answered these questions correctly the first time ($M = 5.96$), $p = .038$.

the scale, $t(34) = 6.08$, $p < .001$. The results of an imagined choice are consistent with the results of condition 1, study 1, if not stronger by virtue of the more "choicey" vignette.

In condition 2, of the forty-eight participants who started the survey, forty-two completed it and had not recently taken a very similar survey (four had). Of these, thirty-nine indicated a phenomenology of an ability to choose among possibilities and all but two of them passed the training section.[17] The remaining thirty-seven participants gave a mean response of 5.78 on the compatibility question, which differed significantly from the midpoint of the scale, $t(36) = 6.85$, $p < .001$. That is, participants were again incompatibilist about the phenomenology, this time of an actual choice.

In condition 3, of the fifty-seven participants who started the survey, fifty completed it and had not recently taken a very similar survey (three had). Of these, forty-three indicated a phenomenology of an ability to choose among possibilities and all but three of them passed the training section.[18] Most of the remaining forty participants (90 percent) opted to donate to the Childhood Cancer Foundation, as we expected on the assumption that the cancer charity would be regarded as more morally salient. The forty participants gave a mean response of 5.85 on the compatibility question, which differed significantly from the midpoint of the scale, $t(39) = 7.66$, $p < .001$. Once again we find incompatibilist phenomenology—this time with a morally salient choice.

ANOVA testing showed no overall effect of condition among conditions 1, 2, and 3, $F(2, 111) = .254$, $p = .776$. So there appears to be no effect produced by making the condition an actual choice, or by making the choice morally salient.

The results of study 2 show that people report incompatibilist phenomenology of agency for actual choices. Indeed, whether the decision is set up as an imagined one or an actual one does not affect the degree to which participants interpret their agentive experience as being incompatible with determinism. The results also show that whether or not the decision is morally salient doesn't affect the degree to which participants interpret their agentive experience as being incompatible with determinism. So the results of previous studies seem to extend to the moral domain, where issues of responsibility loom large.

[17] Six participants required correction and successfully passed the training section. There was no difference between the responses of those who required correction and those who didn't ($p > .2$).

[18] Six participants required correction and successfully passed the training section. Again, we found no difference between the responses of those who required correction and those who didn't ($p > .2$).

4. STUDY 3

4.1 Overview

One possible concern with our previous studies stems from the way we have phrased the key compatibility question. Notice our use of an "even though" locution in the following:

> Even though it felt like I could either choose to donate to *Castanea Dentata* or choose to donate to *Ulmus Dentata*, if causal completeness is true then I couldn't really have chosen differently than I did. (Emphasis added.)

Although we took this to be a natural phrasing of the question, one might think that "even though" primes the participant to agree with the statement, which in this case is an incompatibilist response. Our final study drops this potentially troublesome phrase and also tests whether the phenomenology of epistemic uncertainty differs from the phenomenology of being able to do otherwise in terms of compatibility with determinism. In condition 1, we once again presented participants with an actual choice among two options and tested whether they would continue to report having an incompatibilist phenomenology as of being able to do otherwise. In condition 2, we focused on epistemic phenomenology.

Method:

Participants:
One hundred and six participants were initially recruited online through the Mechanical Turk (MTurk) website. The survey itself was conducted using SurveyMonkey. Fifteen participants did not complete the survey, or indicated that they had recently taken a "very similar" survey. They were excluded from the analysis.

Materials:
The vignette and first question for condition 1 read as follows.

Part 1: The ability to do otherwise In both conditions, participants were told that they would have a chance to win 5 cents if they picked the right button. The text went as follows:

> At the bottom of this page, there are two buttons, labeled H and V. Each option is currently available for you to choose. In a moment, we'll ask you to choose just one of them. For this survey, only one of the buttons will give you an extra $0.05 (as bonus payment on MTurk) if you choose it. But we won't tell you which button it is—you'll have to make a choice and find out.
>
> But don't decide just yet.

> First, consider how things seem to you as you face your decision. In particular, consider what it's like to decide which option to choose.

In condition 1, participants were asked to indicate agreement (on a 1–7 scale) with the following statement:

> When deciding which option to choose, it feels like I can either choose H or choose V.

Condition 2 was the same except that we dropped the modal "can" and asked participants to "consider what it's like to wonder which option you'll choose." Participants were then asked to indicate their level of agreement with a statement about *epistemic* phenomenology:

> When wondering which option I'll choose, it feels like I don't know for sure before I select a button which button is the bonus button.

As in study 1, only participants who agreed with the initial statement proceeded to the training.

The two available options—H and V—appeared at the bottom of the screen, with a radio button representing each option. Participants were not told whether they had chosen the bonus button (H) until after they had answered the compatibility question.

Part 2: Training on determinism The training section was the same as that used in study 1, and again participants only proceeded to the compatibility question if they passed the training.

Part 3: Consistency The compatibility question was adjusted for the new cases. In condition 1, participants were told:

> Now, recall the button-choosing situation. You previously agreed with the following statement:
>
> When deciding which option to choose, it feels like I can either choose H or choose V.
>
> Considering this previous statement about how things felt to you before your choice and your understanding of causal completeness, please indicate your level of agreement with the following:
>
> If causal completeness is true, then I couldn't really have chosen differently than I did.

In condition 2, participants were reminded that they agreed with this statement:

> When wondering which option I'll choose, it feels like I don't know for sure before I select a button which button is the bonus button.

They were then asked to indicate their level of agreement with the following:

> If causal completeness is true, then I knew for sure before I selected a button which button was the bonus button.

Our aim was to test whether participants distinguish the sort of alternative possibilities they reported themselves as experiencing in other conditions from clearly compatibilist alternative possibilities, which have to do simply with our ignorance of the future.

Results:

Of the fifty-three participants who started condition 1, forty-seven completed it and had not recently taken a very similar survey (two had). Of these, forty-four indicated a phenomenology as of there being alternative possibilities in the situation and all but three of them passed the training section.[19] The remaining forty-one participants gave a mean response of 5.34 on the compatibility question, which differed significantly from the midpoint of the scale, $t(40) = 4.54$, $p < .001$. That is, participants once again demonstrated a strong tendency to interpret their agentive experience as being incompatible with determinism.

In condition 2, of the fifty-three participants who started the survey, forty-four completed it and had not recently taken a very similar survey (eight had). Of these, thirty-nine indicated a phenomenology of an ability to choose among possibilities and all but one of them passed the training section.[20] The remaining thirty-eight participants gave a mean response of 2.66 on the compatibility question, which differed significantly from the midpoint of the scale, $t(37) = -5.23$, $p < .001$. That is, participants tended to regard their phenomenology of uncertainty about the future as compatible with determinism. A *t*-test between conditions 1 and 2 showed that results differed significantly between these two conditions, $t(76) = 6.85$, $p < .001$.

This final study provides yet further evidence that people do indeed judge that their experience of deciding is inconsistent with determinism, in the sense that the experience is nonveridical if determinism is true. At the

[19] Eight participants required correction, and passed the training section. There was a significant difference in responses between those who required correction and those who didn't. Those who required correction reported that their phenomenology was incompatible with CC (M = 4.25) but to a lesser degree than those who answered these questions correctly the first time (M = 5.60), $p = .057$.

[20] Nine participants required correction, and passed the training section. There were no statistically significant differences in responses between those who required extra training and those who didn't ($p = .2$).

same time, people tend to think that the feeling of not knowing what will happen is perfectly consistent with determinism. This suggests an appropriate sensitivity to the fact that ignorance is not incompatible with determinism.

5. GENERAL DISCUSSION

5.1 Incompatibilism

Our results have implications for several issues concerning free will. Perhaps most importantly, our studies seem to vindicate the incompatibilist descriptions of our experience as of being able to do otherwise suggested by Campbell, O'Connor, and Searle. By the same token, our results run counter to the compatibilist descriptions of our experience suggested by Mill, Grünbaum, and Nahmias and colleagues. The design of our studies left it open for participants to describe their experience as involving the ability to do otherwise, while allowing them to interpret this ability however they wished. The results indicate that the people in the population we tested tended to judge that their experience was *incompatible* with determinism.

The results also address a concern that has plagued recent work on intuitions about free will. Nahmias and Murray (2011) contend that people give incompatibilist responses in previous experiments simply because people misunderstand determinism. This is an important concern. But rather than merely testing to see whether people misunderstand determinism, we attacked the comprehension issue directly by exploiting the familiar technique of training to criterion. And we did not find any widespread confusion of determinism and bypassing. Part 1 of our training controls for confusion between determinism and fatalism, and the majority of our participants reported that the accuracy of their experience as of being able to do otherwise is inconsistent with *determinism*, correctly understood. Across all our studies, the percentage of participants who didn't make it to the compatibility question due to failing the training section was small, at 6.15 percent. When we look at those participants who answered part 1 of the training incorrectly—that is, at those who *did* initially confuse determinism with fatalism, and who were directed to the follow-up training question—the percentage was small compared with Nahmias and Murray's results: only 20.68 percent of participants initially made this mistake. Of those who initially made the mistake, 85.71 percent answered the follow-up training question correctly. Thus, fewer than 3 percent of participants continued to confuse determinism and fatalism after training. And those who required correction did not respond in any significant way differently from those who didn't.[21]

5.2. The Ability to Do Otherwise

Much of the free-will debate, since at least Hobbes, has been about an *ability to do otherwise.* One influential compatibilist thought is that the notion of the ability to do otherwise should be understood in contrast to constraint or coercion. The idea is that an agent is able to do otherwise just in case, if she had chosen, or wanted, or tried to do otherwise, then she would have done so (cf. Moore 1912). There are also recent versions of such a "conditional analysis" of the ability to do otherwise. According to Kadri Vihvelin (2004), for instance, an ability to act (or not to act, which is simply to be able to act in another way) is analyzable along something like the following lines: an agent can Φ at t_1 (say, raise her hand at t_1) just in case were she to choose to Φ at t_2, and her body stayed working normally and nothing interfered with her, she would Φ at t_2.[22] In other words, Vihvelin holds that "persons have abilities by *having intrinsic properties that are the causal basis of the ability*" (2004: 438). So Vihvelin thinks that an ability to act is a disposition, or a bundle of dispositions. And, as she points out, "no one denies that dispositions are compatible with determinism" (2004: 429). After all, even if determinism is true, glass is still fragile—i.e. it has the disposition to break if struck.[23]

Other compatibilists embrace an epistemic reading of "can do otherwise." On this view, to maintain that I can go left or right is simply to note that it is epistemically open whether I will go left or right. J. J. C. Smart argues that this is a natural way to interpret the expression "could have done otherwise" even outside the sphere of action. When I say, "the plate fell, and it could have broken," I am not, says Smart, committing myself to any claim about determinism. Rather, what I am saying is that, before the plate completed its fall, for all I knew, the plate would break (1961: 298). Similarly, perhaps when I say that Oswald could have done otherwise, all

[21] Again, with the exception of study 2, condition 1, and study 3, condition 1. (See footnotes 16 and 19.)

[22] Vihvelin's exact formulation is as follows: "*S* has the ability at time *t* to do *X* iff, for some intrinsic property or set of properties *B* that *S* has at *t*, for some time *t′* after *t*, if *S* chose (decided, intended, or tried) at *t* to do *X*, and *S* were to retain *B* until *t′*, S's choosing (deciding, intending, or trying) to do *X* and *S*'s having of *B* would jointly be an *S*-complete cause of *S*'s doing *X*" (2004: 438).

[23] For similar accounts, see Smith (2004) and Fara (2008). Questions persist (see e.g. Clarke 2009) about whether any "dispositionalist" account is an adequate analysis of the ability to act, and thus of the ability to act otherwise.

I'm saying is that, before Oswald pulled the trigger, for all anyone knew, he wouldn't pull the trigger. If I'm merely making a claim about epistemic possibilities, then there is no conflict with determinism.

By contrast, incompatibilists think being able to do otherwise (in the relevant contexts) means being able to do something other than what one does, all prior conditions (including one's desires) remaining the same. This ability is presumed to be a matter of fact, not something about our epistemic access to facts.

At least insofar as the relevant notion of the ability to do otherwise is reflected in the experience as of being able to do otherwise, our results suggest that the compatibilist accounts fail. Across three studies, participants tended to interpret their agentive experience in terms of an ability to do otherwise, and they interpreted that ability incompatibilistically. Concerning the traditional compatibilist analysis, our results equally undercut old and new versions. After all, we allowed participants to describe their experiences as involving the ability to do otherwise or not, where they were free to interpret this ability however they wished. Participants then judged that *this* ability—the one they had been allowed to interpret however they wished—was incompatible with determinism. The epistemic compatibilist account is also undermined by these results. Participants gave compatibilist judgments about the case of ignorance about the future (study 3, condition 2), indicating that they do have an appreciation that the feeling of uncertainty is consistent with determinism.

We conclude that the notion of "can do otherwise," at least with respect to one's decisions, is naturally interpreted in ways that contravene the most familiar compatibilist approaches in the philosophical literature. When participants attend to their experience while they consider future events, their usage of "can" tends to reflect a sense of *metaphysical* openness that is incompatible with determinism.

5.3. Misinterpreting One's Agentive Experience

Obviously, the fact that people interpret their agentive experience as incompatibilist doesn't show that people actually have an incompatibilist ability to do otherwise. Terry Horgan argues that people might be mistaken in their interpretation of their own phenomenology. He allows that people might regard their agentive experience as incompatibilist:

> When one attends introspectively to one's agentive phenomenology, with its . . . [representational] . . . aspects of freedom . . . and when one simultaneously asks reflectively whether the veridicality of this phenomenology is compatible with causal

> determinism . . . , one feels *some* tendency to judge that the answer to such compatibility questions is No. (Forthcoming)[24]

But Horgan notes that we must distinguish between the content of our experience and the content of judgments. The former kind of content Horgan dubs "presentational content," and it

> . . . is the kind that accrues to phenomenology directly—apart from whether or not one has the capacity to articulate this content linguistically and understand what one is thus articulating, and apart from whether or not one has the kind of sophisticated conceptual repertoire that would be required to understand such an articulation. (forthcoming)

By contrast, "judgmental content" is the kind of content associated with linguistic articulations. Of course, we make judgments about our phenomenology, and so we can have judgmental content that aims to capture our presentational content. The key point here is that it is possible for our judgments about the (presentational) content of our experience to go awry.

That said, those judgments are at least prima facie evidence of the nature of the presentational-cum-phenomenal content, we maintain, so we would need some positive reason to think that participants have systematically misinterpreted the nature of their phenomenology. Further, even if we grant arguendo that the presentational content of agentive experience is (in the first instance) compatible with determinism, and that reports to the contrary count as mistaken interpretations, still, the fact that people judge the experience incompatibilist would be significant. For one thing, when considering how best to understand the notion of the "ability to do otherwise," in many cases what will be of primary importance is how people *think* about their ability to do otherwise, and that is clearly judgmental. Second and more interestingly, judgmental content can feed back into presentational content. It is well known that what one judges about a situation can affect one's perception of the situation. Horgan recognizes this, and he notes that the distinction between presentational and judgmental content isn't always sharp: "it may very well be that the two kinds of content can interpenetrate to a substantial extent" (forthcoming). As a result, even if the presentational content of agentive experience is, in the first instance, compatibilist, that doesn't mean that the presentational content *remains* compatibilist. It might be that the incompatibilist judgment shapes the presentational content.[25]

[24] For Horgan, this representational "aspect of freedom" is what we have been calling the experience as of being able to do otherwise.

[25] Note that Horgan's notion of "presentational content" is not simply "raw feels" with no propositional content. For everyone would concede that insofar as we have

5.4 Deliberation Compatibilism

A final issue that might be illuminated by our results is the debate over the presuppositions of deliberation. Some philosophers have maintained that deliberation carries with it a presumption of genuinely open possibilities of an incompatibilist variety. Richard Taylor writes, "I cannot deliberate about what to do, even though I may not know what I am going to do, unless I believe that it is up to me what I am going to do" (1983: 38–9). And this "up to me" is incompatible with determinism. Peter van Inwagen makes a similar point: "[I]f someone deliberates about whether to do A or to do B, it follows that his behavior manifests a belief that it is *possible* for him to do A—that he *can* do A, that he has it within his power to do A—and a belief that it is possible for him to do B" (1983: 155).

On the other side, we find "deliberation compatibilists," who maintain that deliberation contains no such presuppositions. Tomis Kapitan begins his paper (which would become the locus classicus for deliberation compatibilism) thus:

> By *deliberation* we understand practical reasoning with an end in view of choosing some course of action. Integral to it is the agent's sense of alternative possibilities, that is, of two or more courses of action he presumes are *open* for him to undertake or not. (1986: 230)

Kapitan goes on to argue that the presumption of openness does not require *metaphysical* openness, but only *epistemic* openness.[26] A number of philosophers have followed Kapitan in developing compatibilist accounts of the presuppositions behind deliberation (e.g. Nelkin 2004, Pereboom 2008).

Insofar as deliberation compatibilism claims that deliberation is not *as a matter of fact* experienced as having incompatibilist presuppositions, our

incompatibilist phenomenology, it must be presented at a level with greater conceptual sophistication than is provided by raw feels. Horgan is explicit about the possibility of rich conceptual resources being implicated in presentational content: "It is plausible . . . that humans can have presentational contents the possession of which require (at least causally) a fairly rich repertoire of background concepts that can figure in judgmental states." For instance, "One can have presentational experiences, for instance, as-of computers, automobiles, airplanes, train stations" (Horgan, forthcoming).

[26] According to Kapitan and other deliberation compatibilists, there are other conditions, too. In particular, Kapitan maintains that deliberation carries a presupposition of *efficacy*, which he characterizes roughly as follows: "an agent presumes that his Φ-ing is an open alternative for him *only if* he presumes that he would Φ if and only if he were to choose to Φ" (1986: 234). See also Pereboom (2008: 288). We leave this complication aside since it doesn't affect our point.

studies indicate this position is mistaken. This does not decide the dispute concerning deliberation compatibilism, but it does show that we need to distinguish three versions of deliberation-compatibilism:

(1) People's beliefs about their current deliberations are compatible with determinism;
(2) People's beliefs about their current deliberations are not compatible with determinism, but they can be adjusted to be compatible;
(3) People's beliefs about their current deliberations are not, and cannot be adjusted to be, compatible with determinism, but we can conceive of a rational being whose beliefs about deliberation are compatible with determinism.

Our results suggest that the first version of deliberation compatibilism is false. People's beliefs about their deliberations are incompatibilist. The second version—that our actual experiences are incompatibilist but revisable—is an interesting possibility, but it remains an open question whether it *is* possible to revise this aspect of our experience. Until we know more about what generates the incompatibilist experience, it is hard to know whether it can be modified. One possibility is that the incompatibilist experience is generated in a way that is not cognitively penetrable (see e.g. Bayne 2011). That is, it might be that even if we form the explicit high-level belief that deliberation is theoretically compatible with determinism, this will not eradicate our experience of our deliberation as incompatibilist. The third version of deliberation compatibilism—that we can conceive of rational creatures who deliberate as determinists—is not under any threat from our results. But if it turns out to be impossible for us to *be* such rational animals, that might undercut some of the interest of deliberation compatibilism.

6. CONCLUSION

The experience as of *being able to do otherwise* has long been central to debates about agency and free will. Libertarians appeal to this experience as evidence that determinism is false; compatibilists reject the libertarian accounts of the character of the experience. Despite the pivotal role of experience in these arguments, the experience itself has received scant attention. Our studies are an attempt to advance the issue. We find consistently incompatibilist judgments about the nature of the experience as of being able to do otherwise. This lends support to the phenomenological claim of libertarians, though we ourselves are not inclined to take the phenomenology of indeterminism as evidence that agency isn't determined. Our results also suggest that an incompatibilist interpretation of the notion of "ability to do otherwise" is the best interpretation of that notion, at least insofar as that notion is supposed to reflect our experience as of being able to do otherwise. Finally, our results also speak to the presuppositions of deliberation. What our studies indicate is that *as a matter of fact* our experience of deliberation features metaphysical openness (that is inconsistent with determinism). While this does not decide the dispute between deliberation compatibilists and deliberation incompatibilists, it does make salient the possibility that deliberation compatibilism requires an account of deliberation that is explicitly revisionist with respect to our actual experience of deliberation.

REFERENCES

Bayne, Tim (2011). "The Sense of Agency." In F. Macpherson (ed.) *The Senses.* (Oxford: Oxford University Press), 355–74.

Berofsky, Bernard (2002). "Ifs, Cans, and Free Will: The Issues." In R. Kane (ed.) *The Oxford Handbook of Free Will.* (New York: Oxford University Press), 181–201.

Buhrmester, M., T. Kwang, and S. Gosling. (2011). "Amazon's Mechanical Turk: A New Source of Inexpensive, Yet High-Quality, Data?" *Perspectives on Psychological Science* 6: 3–5.

Campbell, C. A. (1951). "Is 'Freewill' a Pseudo-problem?" *Mind* 60/240: 441–65.

Campbell, Joseph (2005). "Compatibilist Alternatives." *Canadian Journal of Philosophy* 35: 387–406.

Clarke, Randolph (2009). "Dispositions, Abilities to Act, and Free Will: The New Dispositionalism." *Mind* 118: 323–51.

Fara, Michael (2008). "Masked Abilities and Compatibilism." *Mind* 117/468: 843–65.

Fischer, John M. (1994). *The Metaphysics of Free Will.* (Oxford: Blackwell Publishing).

—— (2008). "Freedom, Foreknowledge, and Frankfurt: A Reply to Vihvelin." *Canadian Journal of Philosophy* 38/3: 327–42.

Frankfurt, Harry (2003/1969). "Alternate Possibilities and Moral Responsibility." In G. Watson (ed.) *Free Will.* (Oxford: Oxford University Press), 167–76.

Grünbaum, Adolf (1952). "Causality and the Science of Human Behavior." *American Scientist* 40(4): 665–76.

Horgan, Terry (Forthcoming). "Causal Compatibilism about Agentive Phenomenology." For a festschrift for J. Kim, eds., M. Sabates, D. Sosa, and T. Horgan.

Hume, David (1960/1739). *A Treatise on Human Nature*, L. A. Selby-Bigge (ed.) (Oxford: Clarendon Press).

Kapitan, Tomis (1986). "Deliberation and the Presumption of Open Alternatives." *The Philosophical Quarterly* 36/143: 230–51.

Lehrer, Keith (1960). "Can We Know that We have Free Will by Introspection?" *The Journal of Philosophy* 57/5: 145–57.

Mill, John Stuart (1865). *An Examination of William Hamilton's Philosophy*. (Boston: William V. Spencer).

Monroe, Andrew E. and Malle, Bertram F. (2010). "From Uncaused Will to Conscious Choice: The Need to Study, Not Speculate about People's Folk Concept of Free Will." *Review of Philosophy and Psychology* 9: 211–24.

Moore, G. E. (1912). "Free Will." From *Ethics*. (Oxford: Oxford University Press), 122–37.

Nahmias, Eddy, Stephen Morris, Thomas Nadelhoffer, and Jason Turner. (2004). "The Phenomenology of Free Will." *Journal of Consciousness Studies* 11/7–8: 162–79.

Nahmias, Eddy and Dylan Murray (2011). "Experimental Philosophy on Free Will: An Error Theory for Incompatibilist Intuitions." In *New Waves in Philosophy of Action*, Aguilar, Jesús, Andrei Buckareff and Keith Frankish (eds.) (Basingstoke, UK: Palgrave Macmillan), 189–216.

Nelkin, Dana. (2004). "Deliberative Alternatives." *Philosophical Topics* 32: 215–40.

O'Connor, Timothy (1995). "Agent Causation." In T. O'Connor (ed.), *Agents, Causes, and Events: Essays on Indeterminism and Free Will*. (New York: Oxford University Press), 173–200.

Pereboom, Derk (2008). "A Compatibilist Account of the Epistemic Conditions on Rational Deliberation," *The Journal of Ethics* 12/3: 287–306.

Perry, John (2004). "Compatibilist Options." In J. Campbell, M. O'Rourke, and D. Shier (eds.) *Freedom and Determinism*. (Cambridge, MA: MIT Press), 231–54.

Searle, John (1984). *Minds, Brains, and Science*. (Cambridge, MA: Harvard University Press).

Smart, J. J. C. (1961). "Free Will, Praise and Blame." *Mind* 70: 291–306.

Smith, Michael (2004). "Rational Capacities." In M. Smith, *Ethics and the A Priori: Selected Essays on Moral Psychology and Meta-Ethics*. (New York: Cambridge University Press), 114–35.

Taylor, Richard (1983). *Metaphysics*, 3rd edn. (Englewood Cliffs, NJ: Prentice Hall).

Van Inwagen, Peter (1983). *An Essay on Free Will*. (Oxford: Oxford University Press).

Vihvelin, Kadri (2004). "Free Will Demystified: A Dispositional Account." *Philosophical Topics* 32: 427–50.

6

Reasons-Responsiveness, Agents, and Mechanisms[1]

Michael McKenna

Many philosophers are convinced by Harry Frankfurt's (1969) controversial argument that moral responsibility does not require the ability to do otherwise. According to Frankfurt, it is possible to construct an example in which an agent acts freely and is morally responsible for what she does but is unable to do otherwise. Frankfurt's example featured Jones, who shot Smith for his own reasons. As it happened, Black wanted Jones to shoot Smith on his own, but covertly arranged things so that if Jones were about to fail to do so, Black would manipulate Jones so that he (Jones) would shoot Smith. Since Jones shot Smith on his own, Black remained dormant; he played no role in Jones's act. Jones, it seems, acted freely and was morally responsible for doing so. Yet because of Black's presence, Jones could not have done otherwise. Hence, as the example involving Jones illustrates, moral responsibility does not require the ability to do otherwise.[2]

[1] This paper was written for the 2011 New Orleans Workshop on Agency and Responsibility (NOWAR), organized by David Shoemaker. I would like to thank him for arranging this excellent event. I delivered an earlier version of this paper to the University of Texas at Austin Philosophy Department. I am indebted to various members of the audience for their instructive comments and criticisms. Notable among them were Sinan Dogramaci, Jeremy Evans, Alex Grzankowski, Bob Kane, Chris Simpson, Nicole Smith, David Sosa, and Michael Tye. At NOWAR, I profited from discussion with many in attendance, especially Michael Bratman, David Brink, Sarah Buss, Stephen Kearns, Rahul Kumar, Dana Nelkin, Paul Russell, Tim Scanlon, Josh Sheppard, David Shoemaker, (yet again) Nicole Smith, Jada Strabbing, and Gary Watson. For further helpful advice, I would also like to thank John Martin Fischer, Ish Haji, Terry Horgan, Al Mele, Shaun Nichols, Derk Pereboom, Carolina Sartorio, and Brandon Warmke. Finally, I would like to thank two anonymous referees for Oxford University Press.

[2] For an examination of the further details offered in support of Frankfurt's argument, see the introduction to Widerker and McKenna (2003), and also my (2008). Among those who defend some variation of Frankfurt's argument are Fischer (1994),

Granting that moral responsibility requires some kind of freedom, the lesson that those convinced by Frankfurt's argument should take from it is that the kind of freedom exercised by Jones, whatever it comes to, is not to be accounted for in terms of the ability to do otherwise. Hence, it is not to be understood in terms of *alternatives* to the way an agent acts. It is, instead, to be understood in terms of the *source* of her actions. It requires, as John Martin Fischer (1994) has put it, attention to the actual-sequence of events leading to action. Call this kind of freedom *source freedom*, and the kind that is understood in terms of alternatives *leeway freedom*.[3]

Despite what many assume, Frankfurt's conclusion is not the exclusive domain of compatibilists. *Source incompatibilists* have accepted the conclusion to Frankfurt's argument and set out to develop an actual-sequence account of free action, one that requires the falsity of determinism.[4] Nevertheless, compatibilists stand to gain more should Frankfurt's argument turn out to be sound. This is because incompatibilists have a powerful argument for the conclusion that determinism is incompatible with the ability to do otherwise. This argument, the consequence argument, strongly suggests that if free will and moral responsibility require the ability to do otherwise, then determinism is incompatible with both.[5] Armed with Frankfurt's argument, *source compatibilists* can grant that determinism is incompatible with the ability to do otherwise. But then they can argue that, even if it is, this is not relevant to whether moral responsibility and (all of) the freedom required for it are compatible with determinism. As a source compatibilist, this is how I proceed.[6]

The project for the source compatibilist who relies upon the soundness of Frankfurt's argument is to offer a plausible actual-sequence account of source freedom that satisfies two desiderata. First, it is of a sort that can be

Fischer and Ravizza (1998), Haji (1998), Hunt (2000), McKenna (2003), Mele (2006), Mele and Robb (1998), Pereboom (2001), Stump (1996), Widerker (2006), and Zagzebski (2000).

[3] I write in terms of "source freedom," while Fischer (1994), and Fischer and Ravizza (1998) use the term "guidance control." I regard these terms as amounting to the same thing. Likewise, I prefer the expression "freedom to do otherwise" or "leeway freedom," whereas Fischer and Ravizza write in terms of "regulative control."

[4] For example, see Hunt (2000), Pereboom (2001), Stump (1996), Widerker (2006), and Zagzebski (2000).

[5] The consequence argument was first developed by Ginet (1966), and then subsequently refined by van Inwagen (1975), and Wiggins (1973).

[6] The person who deserves the credit for this way of framing the dialectic is Fischer (1994). Fischer coins a very similar view "semicompatibilism." Where I differ from him is that, on the basis of (some version of) the consequence argument, I am prepared to grant to the incompatibilist the incompatibility of the ability to do otherwise. Fischer instead wishes to remain uncommitted on this point.

present in a Frankfurt example; second, it can be shown to be compatible with determinism. Although many source compatibilists are attracted to a mesh or hierarchical theory of the sort that Frankfurt himself endorses, in my estimation, such theories face insurmountable problems.[7] In what follows, I shall instead explore the prospects of accounting for source freedom in terms of responsiveness to reasons. In so doing, I will pay special attention to a distinctive approach thoughtfully developed by Fischer and Mark Ravizza (1998) which focuses on the mechanisms of an agent's actions. As I shall explain, a mechanism-based approach appears to be necessary in order to fit a reasons-responsive theory for Frankfurt examples. Unfortunately, this approach comes at a high cost, since there are serious problems with analyzing exercises of source freedom in terms of an agent's mechanisms of action. Ultimately, I shall seek an account of reasons-responsiveness that avoids these difficulties.

1. THE APPEAL OF A REASONS-RESPONSIVE THEORY

Reasons-responsive theories have an ancient lineage that arguably can be found in Aristotle.[8] They are appealing irrespective of the compatibilism-incompatibilism debate for at least two reasons. First, they target a feature of agency distinctive of persons, a feature that marks persons as rational animals. Second, they offer elegant explanations of the conditions in which various excuses and exemptions apply—the persons in question, it is contended, are not suitably responsive to a proper range of reasons. For instance, a plausible way to distinguish between the person who, weak willed, freely drinks the beer from the person who is compelled by her addiction to do so is that the former is responsive to a wider range of reasons for not drinking than is the latter.

Compatibilists have found powerful reasons to develop a reasons-responsive theory quite apart from the unique dialectical burdens of source compatibilism. Reasons-responsiveness seems especially well suited for a compatibilist analysis since in explaining sensitivity to reasons, we must consider how an agent would respond to a range of reasons and not just those present to an agent when she performs an action in the actual circumstances that she is in. This fits naturally with attending to a range of dispositional or modal properties of agency, and there is little reason to

[7] See Appendix I for my source compatibilist objection to these theories.

[8] See Terrence Irwin's (1980).

think that determinism is incompatible with the possession of these sorts of properties generally.

To appreciate how a reasons-responsive theory might be developed, begin with the following two preliminary points:

First, when an agent acts who is suitably reasons-responsive, the most important factor for the source compatibilist in accounting for her freedom is that the etiology of the act which she actually performed involved springs that were sensitive to reasons. For now, let us think of those springs in terms of the agent herself as the cause of her acts, where this is *not* meant to commit to thinking of agents as distinct, irreducible substance causes. Different reasons, understood as different inputs, would have yielded different outputs, understood as alterations in modes of conduct. And what this shows is that the agent's response to the actual "inputs" played a role that was *itself* sensitive to, or responsive to, the actual conditions in which the agent acted. To illustrate, consider a simple example of the sensitivity of a primitive gizmo, a thermostat. Suppose a thermostat is set at 76 degrees (Fahrenheit) and the room the thermostat is in turns out to be 76 degrees. One might wonder if the thermostat's setting accounts for the temperature of the room. After all, it might be disconnected and so merely a fluke that its setting and the room's temperature are 76 degrees. When we learn that the room would come up to 78 were the thermostat set to 78, or would come down to 74 were the thermostat set to 74, and so on for numerous other values high and low of 76, we do not just learn something about the way the thermostat *would* behave; we also learn about how, in the actual scenario when it is set to 76, it *does* behave; it plays a certain causal role from reliable and suitably sensitive resources.

The same can be claimed about a person who is responsive to reasons. Consider by contrast a clear case of reasons-responsive failure, a case in which a person fails to take reasons as inputs in a manner that yields proper results (like a defunct thermostat). Imagine a compulsive hand-washer washing her hands after handling some trash. Not knowing her condition, we might initially think that her freedom consists in part in her responding rationally to conditions that would warrant hand washing. But now suppose that we learn that she'll wash her hands when they brush up against a bottle of bleach, or when she touches her own shirt, or if the sun shines on them, or if the wind blows, or if Don Knots is on TV, or there is a Christmas carol playing on the radio. This casts doubt on whether the reason she is washing her hands in the actual world is because she was handling some trash. Or, even if her washing her hands *is* to be accounted for in terms of her handling some trash, her insensitivity to other reasons suggests that the *manner* in which she caused her action was not suitably sensitive to the significance of trash-handling, as in contrast with, say, the wind blowing.

Second, responsiveness to reasons comes in degrees. This raises questions about what amounts to an adequate degree of responsiveness and what sorts of failures are tolerable. Consider again the thermostat. Suppose the thermostat functioned as described for uses in a normal home environment. It would not impugn the adequacy of the thermostat to learn that it is not super-responsive, and so not responsive to temperatures of 2 degrees at one end and 150 at the other. Alternatively, the thermostat might be only very minimally responsive, functioning properly between settings of 75 through 77, but yielding whacky results at any other settings. Furthermore, failures of responsiveness can be localized. Imagine that the thermostat that is set at 76 would not have responded differently to any settings of between 74 and 78, but would have responded properly to settings of lower than 74 or higher than 78. Here, we could still not determine what role setting the thermostat at 76 played in the room's being 76 degrees, since the room would have remained at 76 degrees with settings of as low as 74 or as high as 78. Even if we were to conclude that setting the thermostat at 76 caused the room to become 76 degrees, we would be right to think that the mode of its causing the change in temperature was too fluky, so much so that all we could feel confident in claiming as a credible *explanation* is that it reliably caused the room's temperature to be some temperature anywhere between 74 and 78, and as chance would have it, it happened to be 76.

The same points apply to persons and their responsiveness to reasons. A person who is suitably reasons-responsive in a way that is sufficient for free action need not be able to respond to a range of reasons that meets exceedingly high standards, otherwise almost no one would turn out to act freely. And a person who is only responsive to a very slim range of reasons will fall short of acting freely while nevertheless being responsive to some reasons. Take our compulsive hand-washer. It seems that she will wash her hands come what may. But it might be that a very limited range of incentives would result in her not washing her hands: if her child were on fire, or if her favorite Mozart concerto were playing and she instead wanted to dance, and so on. While such cases might show her to be responsive to reasons, she would only be minimally so, and this would not be enough to regard her as acting freely in washing her hands as she did. Furthermore, as with simple gizmos like thermostats, so too with persons, localized failures are possible. Someone who is as competent as anyone might fall to pieces in the presence of an abusive spouse or a domineering partner. Localized foibles are part of the package deal for most of us imperfect humans.

2. FISCHER AND RAVIZZA'S ACCOUNT OF MODERATE REASONS-RESPONSIVENESS

Now, in light of the two preceding points, let us examine Fischer and Ravizza's carefully developed reasons-responsive theory.[9] According to them, at one end of the spectrum is strong reasons-responsiveness (SRR). An agent who is SRR with respect to an act is such that, if there were sufficient reason for her to do otherwise, she would do otherwise (Fischer and Ravizza, 1998: 41). At the other end of the spectrum is weak reasons-responsiveness (WRR). An agent who is weakly reasons-responsive with respect to an act is such that there is at least one (nonactual) scenario in which there is sufficient reason to do otherwise and the agent does otherwise (44). As should be clear, and in keeping with the second point developed above (in Section 1), SRR is too strong; if a condition of free agency required SRR, it would turn out that almost no one acts freely and is morally responsible for what she does. On the other hand, WRR is too weak; if a condition of free agency required no more than WRR, it would turn out that many severely impaired agents would fully satisfy the freedom conditions for moral responsibility. Our compulsive hand-washer was WRR in washing her hands since she'd have not washed them if her children were on fire. So, what is needed, and what Fischer and Ravizza develop in admirable detail, is a middle-of-the-spectrum degree of responsiveness, moderate reasons-responsiveness (MRR), that is suited for sane, competent, albeit imperfect, moral agency—the sort of agency of which most psychologically healthy adults are capable while functioning in normal practical contexts of deliberation and action (69–84).

So how do Fischer and Ravizza develop MRR? Understanding their proposal requires careful attention to what is involved in being responsive to reasons. Fischer and Ravizza distinguish between a *receptivity* component and a *reactivity* component (1998: 41). Being reasons-receptive is a matter of being able to recognize what reasons there are, and in particular, being able to recognize what reasons count as sufficient for action. Being reasons-reactive is a matter of being able to react to the reasons one recognizes as sufficient by choosing and acting as needed. According to Fischer and Ravizza, MRR requires *moderate reasons-receptivity.* For a person to be morally responsible for what she does, she must be receptive to reasons in

[9] In examining Fischer and Ravizza's view in this section, I shall depart from their doing so in terms of mechanisms that are reasons-responsive and that are owned by the agent. I shall instead just write in terms of agents being reasons-responsive. In the next section, I'll recalibrate to fit their mechanism-based formulation.

a manner that displays a rich pattern of recognition whereby reasons can be placed on a scale with a continuum involving stronger and weaker incentives. For instance, if an agent regarded an incentive of $100 as sufficient for acting in a certain way, she would take higher values as also sufficient, while being open to taking lower values as insufficient. Furthermore, the agent's pattern of reasons recognition must display a sane appreciation for reasons that are grounded in reality, and the agent must also be able to recognize moral reasons and be able to grasp that sometimes they are sufficient for acting in ways that morality prescribes. All of this, of course, is consistent with being an agent who is not receptive to the full spectrum of reasons for how she ought to act, including the full spectrum of moral reasons.

As regards reasons-reactivity, Fischer and Ravizza allow for a striking asymmetry. An agent who is MRR need only be *weakly reasons-reactive*. It is enough, they contend, that among the worlds in which an MRR agent is receptive to sufficient reasons to do otherwise, she react to only one of those reasons. Here, one might wonder why weak reactivity would be enough. Wouldn't this show that in the wide spectrum of cases in which the agent recognizes sufficient reasons but does not act upon them that the agent *cannot* act upon them and so is impaired for morally responsible agency? Fischer and Ravizza do not think so. They contend that "reactivity is all of a piece" (1998: 73). As they see it, the fact that there exists just one world in which an agent reacts differently to a sufficient reason to do otherwise is sufficient to establish that in *each* world in which an agent recognizes sufficient reasons to do otherwise that she is able to react to those reasons, even if at those worlds she does not.

Critics have pressed Fischer and Ravizza on various details of their proposal regarding the spectrum of MRR. One has to do with their requirement of weak reasons-reactivity. As I have argued elsewhere (2005), weak reactivity looks to be too weak. The fact that an agent would react to only one reason to do otherwise but not a constellation of other similarly related reasons calls into question her agency. Hence, I have proposed a variation on MRR in terms of *weaker* reactivity, not weak reactivity.[10] The spectrum of reasons to which an agent must be reactive can be weaker than the spectrum of reasons to which she is receptive[11], but

[10] Fischer (2006: 328) has granted this point in response to Mele (2006b: 290), who presses a similar worry.

[11] This is needed in order to allow for cases in which an agent freely and knowingly acts contrary to what she has sufficient reason to do. Note, however, that I have only claimed here that the spectrum of reactivity *can* be weaker than that of receptivity. Elsewhere (2005), I have made the stronger claim—that their view requires this asymmetry. Dana Nelkin and David Brink have convinced me that this is a mistake. The requirements of morally responsible agency do not, strictly speaking, *require* any asymmetry. Nevertheless I suspect for nearly all actual persons who are morally responsible

the spectrum of reactivity still needs to display a sane, stable pattern along the lines of moderate receptivity. There are further details that we might add here to supplement Fischer and Ravizza's proposal, or instead various details of their thesis with which we might take issue, but for present purposes the preceding discussion will do.

While it is important to make clear the delicate effort Fischer and Ravizza make to get right the degree of responsiveness needed to account for MRR, the most important factor in their account of freedom is that the etiology of an agent's act involved springs that were themselves sensitive to reasons. This shows Fischer and Ravizza's deference to the first of the two preliminary points set out above (in Section 1). I pause here to emphasize this since it is easy to lose sight of. The salient point to note is that by demonstrating that when an agent acted she was MRR, one does not *just* show how that agent would respond to other reasons if those reasons were present to her. One also shows that in acting as she did, the agent *was* responsive to the conditions she was actually in. Her actually-operative springs of action were functioning in ways that *themselves* were responses to the conditions in which she actually found herself.

To elaborate on the preceding point: It is easy when reflecting upon Fischer and Ravizza's proposal to misunderstand the theoretical purpose to which certain counterfactuals are put. Consider a MRR agent who in deciding to remain home and work on an article would be receptive and reactive to the reason that the legendary blues guitarist Buddy Guy is putting on a concert that night and she could easily go. Were she presented with this reason, she would not work on her article. Instead, she'd go to the show. Don't be snookered here into thinking that what this counterfactual is meant to establish is that, in acting as the agent did, she could have done otherwise. That is, the point of this counterfactual, as well as other counterfactuals involving other reasons to which this agent might be both receptive and reactive, is not meant to establish that the agent had a freedom that concerned alternatives to what she did do—leeway freedom. It is instead meant to establish that in acting as she did, she was sensitive to reasons in such a way that her actual bringing about of her act of remaining home and working on her article was *itself* suitably sensitive to rational considerations. Some compatibilists, such as those classical conditionalists like Ayer (1954), Davidson (1973), Hobart (1934), and Moore (1912), attempted to rely upon similar counterfactuals in order to account for a compatibilist theory of the freedom to do otherwise. But Fischer and Ravizza do not

agents, there will be some asymmetry of this sort; most persons are not moved to act by the full range of sufficient reasons for action to which they are receptive.

understand the worlds in which an agent does otherwise in response to differing reasons as worlds that, as they put it, are *accessible* to the agent in the actual world in which she does act (1998: 53). So, as they see it, it cannot be that by virtue of such worlds an agent in the actual world is able to do otherwise. Hence, Fischer and Ravizza do not mean for these counterfactuals to aid in underwriting an account of the freedom to do otherwise.

3. SHIFTING TO A MECHANISM-BASED THEORY

Given the preceding treatment, can we fit a reasons-responsive theory into the actual-sequence constraints of Frankfurt's argument? Regrettably, on first appearance, these two compatibilist-friendly theses do not appear to make comfortable bedfellows. Recall Jones from the Frankfurt example presented earlier. Jones, says the source compatibilist, acted freely. But due to Black's presence, whatever reasons might have been put to Jones, it is not the case that he would have done other than as he did. This is because Black was playing the role of a "counterfactual intervener" and stood prepared to ensure that Jones would shoot Smith no matter what. Thus, if an agent acts freely in a Frankfurt example, it seems that her freedom cannot be explained in terms of her responsiveness to reasons. Due to the presence of a counterfactual intervener, the agent would not act otherwise even if different reasons were present (1998: 38). So is reasons-responsiveness ill-suited for a source compatibilist theory?

According to Fischer and Ravizza, we can fit a reasons-responsive theory to the kind of freedom present in a Frankfurt example by shifting from an *agent-based* to a *mechanism-based* theory. The *agent* in a Frankfurt example, they contend, is *not* responsive to reasons, but *the mechanism* from which she acts, which is her own mechanism of action, *is* reasons-responsive. To appreciate their intriguing proposal we need to answer two questions. First, what is a mechanism of action as they understand it? Second, how is it that focus upon an agent's own mechanism of action ensures immunity to the problem that seems to arise for an agent's reasons-responsiveness in a Frankfurt example? Why does the shift to mechanisms help?

As regards the first question, the mechanism of action, as Fischer and Ravizza understand it, is meant to pick out "nothing over and above the process that leads to the relevant upshot" (1998: 38). To make clear that they do not mean to reify the notion of mechanism or to suggest that it is something like a natural kind, they remark that rather than use the term, we could instead just speak in terms of the "way the action came about" (38).

Clearly, in this sense, the mechanism of action at work in the Frankfurt example involving Jones's acting on his own is very different from the mechanism that would have operated were Black to have intervened and caused Jones to shoot Smith. That much, it can be safely granted, is clear and uncontroversial. Furthermore, one can also infer that among the full constellation of events, states and processes constituting Jones, or Jones's complete psychic condition at the time, not all of the elements of that constellation were causally implicated in Jones's shooting Smith. Jones, for instance, might have a deep love of his mother, a fondness for kittens, and be especially skilled at solving calculus problems. But these, we can safely assume, can be "filtered out" of the complex of Jones's psychic events, states, and processes that were implicated in the process leading to has act of shooting Smith. Thus, an agent's mechanism of action, however it is to be construed, carves out a subset of the full set of psychological elements constituting the agent's psychology.

As regards the second question, Fischer and Ravizza argue that to test whether in a Frankfurt example an agent's own actually-operative mechanism of action was reasons-responsive, we need to test it for sensitivity to various reasons in conditions in which it operates uninhibited, and to do this we have a license to "go to worlds", as the expression goes, in which Black's presence is subtracted. So, were Black not present, and were Jones presented with pertinent reasons for not shooting Smith, Jones, by way of his mechanism of action, would react otherwise and not shoot Smith. According to Fischer and Ravizza, in order for it to be the case that Jones exercises adequate source freedom in shooting Smith, his mechanism of action must be such that the pattern of reasons to which it is sensitive would be MRR.

4. PROBLEMS WITH MECHANISMS

Various critics have objected to Fischer and Ravizza's mechanism-based approach. R. Jay Wallace has remarked that by shifting from an agent-based to a mechanism-based view, the "intuitive locus of responsibility, the person, drops out of view" (1997: 159). He claims that instead we should tend to the normative competence of persons. Gary Watson has expressed his concern that characterizing a mechanism of action in terms that are no more specific than " 'a process resulting in behavior' is too amorphous to do the refined work needed by the theory" (2001 as appearing in Watson, 2004: 299). And Carl Ginet has claimed that the postulation of mechanisms is not doing any useful work in their theory (2006: 235–6).

According to him, Fischer and Ravizza want it to be that an agent would respond differently to certain reasons by virtue of a mechanism. Their theory, Ginet contends, could be preserved by cutting out the middle term and just theorizing in terms of an agent being reasons-responsive.[12] I too have registered some concerns with their appeal to mechanisms. However, my criticisms have been more limited. I have only argued that Fischer and Ravizza need to provide further support for their account of mechanisms in order to avoid various objections.

In my estimation (2001), Fischer and Ravizza face difficulty with individuating the mechanisms operative in exercises of free action. They resist pressure to provide any principled basis for mechanism individuation and contend that they can make do with an intuitive sense of what counts as the same mechanism and when, in shifting between contexts, we move to different mechanisms (1998: 40).[13] Clearly, there is a different mechanism of action at work when an intervener such as Black causes Jones to act as in contrast to when Jones acts on his own. Thus, on their view, just as we are able to reidentify the same face or smile, or the same house or car, without a particular theory or general principle, so too we can recognize same mechanism of action operative in different contexts. The problem, however, is that there are some contexts crucial to the development of their theory where, absent some principled bases for mechanism individuation and differentiation, it is hard to assess their broader claims. Obviously they are in no danger when distinguishing between actual and alternative mechanisms in Frankfurt examples. But it is an indispensible part of their theory that we are to test how *the same mechanism* behaves in response to different reasons. How are we to settle whether, when different reasons are put to an agent, the same mechanism is operative? We need some purchase on what it is that we are holding fixed when we hold fixed "the same mechanism" while testing its degree of responsiveness.

The problem stems from Fischer and Ravizza's lean characterization of a mechanism of action. "The process that leads to action" or "the way the action came about" allows for further specification by way of extremely narrow or instead extremely wide descriptions. Picking out the process that

[12] For Fischer's replies, see his (2004: 169–71) reply to Watson, and his (2006: 333–4) reply to Ginet. He replies to Wallace in his (2011, ch. 9: 144–62).

[13] They do offer one theoretical constraint. The descriptive content by which the mechanism is identified must be "temporally intrinsic" (1998: 46–7). That is, it must pick out the mechanism in such a way that it does not involve reference to later times. Otherwise, the mechanism would require certain outcomes, and thus would not permit variability in the face of different reasons. Hence, the temporally extrinsic mechanism "deliberation prior to donating to a charity" is ruled out since it entails acting in a certain way.

leads to action seems most naturally to involve the entire complex of proximal antecedent states and events figuring into a correct explanation of the pertinent action.[14] Suppose that we limit the items in this complex just to the agent-involving ones—an obviously permissible restriction. Still, is every agent-involving feature of the complex to be included in the description? What salient features might one chose to whittle the description down? At one level of description, sameness of mechanism right down to microdetails would be required. But, as I have argued elsewhere (2001: 97), this would likely give rise to serious problems. Suppose that the blameworthy reason upon which an agent acts in the actual world—or rather the set of psychological ingredients involved in recognition of this reason—is identical with or supervenes upon some neurological state of the brain. If sameness of mechanism requires sameness of microdetails, when *that* mechanism is tested to learn how it would respond to different reasons, it would have to respond to different reason by resources that shared the same microdetails with the brain states involved in recognition of the blameworthy reason in the actual world. This is nomically unlikely, albeit, admittedly not metaphysically impossible. It thus exposes Fischer and Ravizza's theory to the rather likely prospect of empirical refutation. So it seems that a very narrow description needs to be ruled out.

Of course, Fischer and Ravizza would be on solid ground in ruling out such narrow descriptions.[15] Typically, when identifying causal processes or ways events are brought about, we do not mean to focus upon microdetails. For instance, identifying the event of the flood as the cause of the erosion needn't require that we focus on the microdetails of just how it was that the water molecules were arranged. In such contexts it would be highly unreasonable to demand such a level of specificity. Fischer and Ravizza are certainly entitled to a similar reply here as regards the sorts of processes leading to (putatively) free acts.

But then, how wide, and so permissive, should the needed description be? Because Fischer and Ravizza offer no principled basis for mechanism individuation, we are left to settle the matter exclusively by appeal to our intuitive reactions to varying cases. Unfortunately, intuitions in these kinds of highly theoretically charged contexts can vary widely. The neurophysiologists' modes of parsimony, the cognitive scientists', or instead the cognitive therapists', are going to be very different from those employed in normal folk-psychological discourse in which typical claims

[14] I am indebted to Carl Ginet for this point. See also his discussion of it (2006: 233).

[15] I am indebted to Derk Pereboom, who emphasized this point.

of blameworthiness and praiseworthiness have their natural homes. And among philosophers differentially committed as regards the free will dispute, it is easy to see that some will be inclined to a hyperrestricted description that will narrow in on requiring same past and laws. Whereas others will be open to descriptions that are far more permissive. Of course, this would just be to redraw old lines in new sand. So that won't help. What is needed, it seems, is some rationale with theoretically independent appeal that will allow us to assess claims of same or different mechanism.[16]

5. A DEEP PROBLEM FOR ANY MECHANISM-BASED APPROACH

The preceding considerations strongly suggest that if a mechanism-based version of a reasons-responsive theory is to survive, there must be some principled way to get an independent purchase on what counts as the same mechanism across relatively diverse contexts. In my earlier work (e.g. 2001, 2011), I have often expressed confidence that there is a way to supplement Fischer and Ravizza's theory so as to offer the right kind of rationale for mechanism individuation. Unfortunately, my considered opinion now is that there is a structural problem which stands in the way of offering any more specific content to the individuation conditions for mechanisms of action. To explain, it will be useful to begin by considering an objection developed by Ginet (2006: 234–5).[17] Fischer and Ravizza take up the example of a person driving a car from a mechanism of unreflective habit (1998: 86). This person takes an exit she often takes, thinking nothing of it. Were it blocked, she'd respond properly to the pertinent reason without any deliberation at all. She'd just head to the next exit. But as Ginet plainly points out, while the case Fischer and Ravizza consider is readily handled by their treatment, there is a slight variation on it which is problematic. Imagine instead a person who, upon seeing the exit blocked, is unsure of what other exit she ought to use. Here, what would be most natural would be for her to deliberate, and it would seem that a reasons-responsive theory should seek to accommodate such a case. But Fischer and Ravizza hold fixed the mechanism of *unreflective habit*. Hence, they cannot permit that

[16] For an illustration of one particular point in which, in responding to an incompatibilist concern, Fischer and Ravizza cannot make do without a principled basis for mechanism individuation, see Appendix II. There, I also discuss Fischer's (2004: 166–71) thoughtful reply to my criticism.

[17] Watson has raised a similar worry (2001 as appearing in 2004: 298).

the mechanism of unreflective habit from which this agent acts is plastic enough to allow reflection when conditions call for aborting reliance on habit alone.

Ginet's insightful criticism is a specific instance of a more general problem that is bound to plague any attempt to give more content to what might count as the conditions for individuating mechanisms. The general problem, as I see it, is that any complex system will have "subsystems" that are designed to function precisely by shutting down or by permitting other systems to override in some contexts but not others. Assuming a person functioning as a practical agent can be understood as a complex system, or at least as relevantly analogous to one, any attempt to whittle down which agential elements are the ones implicated in "the" mechanism of her action is bound to restrict the agent's flexibility as regards reasons-responsiveness.

The postulation of an agent's mechanism of action as narrower than the full person herself looks innocent enough when one is invited to consider a case like Jones and note that Jones's ability to do calculus is irrelevant to the agent-involving factors causally contributing to the shooting of Smith. There are likely lots of these ingredients in Jones's complete psychic constitution that could be whittled away: his love of his mother, his fondness for kittens, and so on. But once we seek to put more meat on a restriction of a pertinent mechanism of action, we will face problems like the one Ginet has identified. If we back off and allow for the description of the mechanism to be "fatter" so as to allow for shifts between processes like unreflective and then reflective habit, we pretty much have what is the functional equivalent of the agent herself as the mechanism, minus various outlier traits like the ability to do calculus.

Reflection on complex mechanisms like automobiles or computers, things that are built up out of smaller mechanisms, helps to illustrate the point. Suppose that one of a computer's programs runs unimpeded, but the computer is designed to divert that program and prioritize other operations if the system is under stress, or uploading new software, or about to lose power, or what have you. The same applies to an automobile. A car will allow things like unimpeded gas flow through the fuel injection system unless another part of the system recognizes problems with the fuel mix or something of the sort. If we were to evaluate the degree of sensitivity of such a system by "holding fixed the actually operative sub-mechanism" we'd hamstring the system for a variety of conditions to which, without holding these fixed, the larger system as a whole would be able to respond quite easily.[18]

[18] An anonymous referee for OUP has astutely asked about the following way of defending Fischer and Ravizza's appeal to mechanisms. To illustrate, consider Ginet's example of the driver acting from the mechanism of unreflective habit. Why is it a

To give just one action-theoretic example of this sort of thing, consider Alfred Mele's (1995, 2006) work on free agency (sometimes he writes in terms of autonomy). One element in Mele's approach is his account of both weakness and strength of will. To account for these phenomena, Mele distinguishes between the processes by which an agent judges what it is best to do from the motivational ingredients involved in directing her actions. Often, when a person acts, her evaluations of the objects of her desires are aligned with the strength of the desires themselves. In these cases, when she acts, the resources by which she might be able to exercise strength of will and avoid temptation are, so to speak, off line. They're causally inert, and so could not be counted as anything like part of the actual causal process issuing in action. But the agent might be such that she has a kind of monitoring system, and were it that conditions would present themselves (were certain reasons to arise) that would involve the agent desiring more something incompatible with what she judges it best to do, the agent would be able to rely upon these agential resources in such a way as to act with strength of will and thereby act contrary to what she most desires (Mele, 1995: 27). Mele's model here is a perfect example of the kind of dynamic relation between "subsystems" within the overall complex of an agent as a whole. If we want to include in "the mechanism of action" this entire complex, we will have to open the descriptions for "same mechanism" to include more than just what is, narrowly construed, causally implicated in the process leading to action. Here, it seems, "the mechanism" will become so inclusive that we might as well simply identify the entire person with the mechanism.[19]

problem for Fischer and Ravizza that in some scenarios the agent would have to "shift" to a different mechanism by beginning to deliberate? Isn't it enough that the mechanism of unreflective habit is responsive enough to certain reasons to allow for it to, so to speak, shut down and allow another to come on line? Why isn't this enough to say that the mechanism is suitably reasons-responsive? In response, I acknowledge that this is after all a thoughtful way to attempt to defend Fischer and Ravizza. But the upshot is to do so by claiming that the full spectrum of reasons to which an agent (by way of mechanisms or otherwise) ought to be responsive for (something like MRR) requires the postulation of a *plurality* of mechanisms. At this point, it appears that it is *not* a matter of holding fixed *a* mechanism, and now one wonders how close we are getting to just straightforwardly talking of the responsiveness of *persons.*

[19] In correspondence, David Shoemaker has thoughtfully expressed some skepticism about this last point—that a more inclusive notion of a mechanism of action will be so inclusive that we might as well identify it with the person, or with the functional equivalent of the person. In conversation, Shaun Nichols has as well, citing the work by Fodor on the modularity of the mind. Here, I have no problem leaving it as an open question whether there is some richer notion of mechanisms that would do the work that needs to be done for a mechanism-based view like Fischer and Ravizza's. But I strongly suspect that even if such a kind of mechanism were identified, and even if it turned out to

6. RETURNING TO AN AGENT-BASED THEORY AND THE REJECTION OF SOURCE COMPATIBILISM?

So, are we back to the incompatibility of a reasons-responsive theory and source compatibilism? According to Fischer and Ravizza, in order for an *agent* to be reasons-responsive, it must be that if different reasons were presented to the agent, then for at least some range of reasons, the agent would react otherwise. In a Frankfurt example, they contend, due to Black's presence, *the agent is not able to react otherwise*.

This result—the incompatibility of reasons-responsiveness and source compatibilism—would come as welcome news to those contemporary compatibilists who reject Frankfurt's argument and defend the traditional association between free action, moral responsibility, and the ability to do otherwise. Call these traditionalists *leeway compatibilists*. Amongst these leeway compatibilists Michael Fara (2008), Michael Smith (2003), and Kadri Vihvelin (2004) have each independently built upon recent developments regarding the nature of dispositions to account for the ability to do otherwise. In a Frankfurt example, each argues, an agent retains the ability to do otherwise. Randolph Clarke (2009) has labeled the thesis these philosophers advance *the new dispositionalism*. It is open to the new dispositionalists to argue that in a Frankfurt example, an agent *is* reasons-responsive, because, in being able to do otherwise, she is able to react to different reasons.

While there are differences in how they execute their arguments, the new dispositionalists are united in proposing that the free-will ability is somehow to be accounted for in terms of dispositions. Following the efforts of philosophers like David Lewis (1997) and C. B. Martin (1994), these leeway compatibilists all pay heed to accounts of dispositions that incorporate lessons learned from the way dispositional properties can be masked or finked. A disposition is masked when its manifestation is concealed in some way. A piece of salt placed in water remains soluble even if, when placed in water, it does not dissolve because it is encased in wax. A disposition is finked when, in just those conditions that would otherwise trigger manifestation, it is altered so as not to have the disposition. A glass vase sitting on a shelf undisturbed possesses the disposition of fragility even if, were it knocked over, a wizard would turn it to stone before striking the ground, rendering it not fragile.

be narrower than the full person qua agent, it would be *much* richer in content than anything like something restricted just to what is actually involved in the causal generation of action.

Cases of masking and finking show that dispositions cannot be analyzed in terms of simple conditionals such as, "if the vase were toppled, it would break" or "if the salt were placed in water, it would dissolve." The vase mentioned above was fragile, but because it was finked, it would not break when toppled. The salt was soluble, but because it was masked, it would not dissolve if placed in water. In slightly different ways, each of the new dispositionalists propose more complex counterfactuals the truth of which confirms the possession of pertinent dispositions. Roughly, their strategy is of the following variety:

> SC: If this vase were toppled, and if it retained during the relevant duration of time its intrinsic properties P_1-P_n, and if it were not interfered with in a way that would impede the causal efficacy of those intrinsic properties, then it would break.

Now, SC is no more than a crude illustration of what a carefully worked out formulation should look like, the sort that, for instance, both Fara and Vihvelin consider.[20] But for present purposes it will suffice. It is by virtue of counterfactuals like SC that the new dispositionalists propose a more sophisticated treatment of dispositions. If abilities are then to be explained (Fara, 2008), or, more vigorously, fully analyzed (Vihvelin, 2004), in terms of dispositions, then counterfactuals like SC provide a template for an explication or analysis of the free-will ability. An especially important factor in any such proposal is that the deeper explanation of the possession of the dispositional properties by an object concerns the relevant intrinsic properties possessed by the object (e.g. P_1-P_n as mentioned in SC). It is these that give the object the kinds of causal resources to have the effects characteristic of a disposition's typical manifestation(s). The same, then, can be said on such an account for a person, qua agent, and her abilities; the deeper explanation of her abilities, and most notably the ones constituting her free-will ability, will be by virtue of certain intrinsic properties possessed by the agent.

It is open to the new dispositionalists to develop their respective accounts in terms of responsiveness to reason, as both Vihvelin and Smith appear to do (though neither explicitly use the label "reasons-responsive theory"). For a relevant range of reasons, counterfactuals like SC could be constructed to help underwrite claims about the spectrum of responsiveness that Fischer and Ravizza attempted with their MRR—in particular, the spectrum of reasons to which an agent is both receptive and reactive. So, take Jones, and

[20] Smith's (2003) essay is written in a more general, programmatic style. He does not develop the kind of detailed proposal that either Fara or Vihvelin does, but it is clear from his exposition that he is open to the development of his view in ways are amenable to Fara's or Vihvelin's efforts.

for the moment set aside his acting in a Frankfurt example. Suppose for now he is not in a Frankfurt example and that he shot Smith on his own and for his own reasons. Suppose also that one of the reasons to which Jones would have been both receptive and reactive in a way that would have resulted in his not shooting Smith is if he were to have learned that Smith's child was with him at the time. Were Jones to have learned of this, he would not have shot Smith. But as it happens in the actual world, he learns of no such reason and proceeds to shoot Smith. Here is a counterfactual, built by modeling it on SC, that shows how it is that Jones is responsive to this reason, call it R_1:

> RR_1: If Jones were to become aware of R_1, and if Jones retained during the relevant duration of time intrinsic agential properties P_1-P_n, and if Jones were not interfered with in a way that would impede the casual efficacy of those properties, then Jones would not shoot Smith.

By compiling a collection of counterfactuals like RR_1, say RR_1 through RR_n, for a spectrum of reasons R_1 through R_n, the new dispositionalists could establish that an agent like Jones is MRR in just the fashion that Fischer and Ravizza have been careful to work out.

Thus far, I have tried to show, in admittedly only broad brushstrokes, how it is that by relying upon counterfactuals like RR_1 through RR_n, the new dispositionalists could make use of their account of abilities to develop a theory of reasons-responsiveness. As I have explained, it is open to them to build the theory in such a way that it is very much like Fischer and Ravizza's proposal for MRR but for the fact that their account would be agent-based, *not* mechanism-based. But how is it that on their view the truth of counterfactuals RR_1 through RR_n also underwrites the ability to do otherwise in a way that shows how an agent retains this ability even in a Frankfurt example? This is what is needed for them to put the nail in the coffin of source compatibilists. Here their rationale is simple. Take Jones whom we pulled from the context of a Frankfurt example, and take RR_1 through RR_n. Now place Jones back in a Frankfurt example. All of RR_1 through $RR_{n,}$ the new dispositionalists will tell us, remain true. Like Fischer and Ravizza, to assess the truth of RR_1 through RR_n they too must go to worlds in which P_1-P_n are allowed to operate unfettered, and this will also require factoring out the presence of a counterfactual intervener like Black. But according to them, this is perfectly consistent with Jones's possession of the ability to do otherwise when he acts on his own. How so? When Jones acts in a Frankfurt example, according to the new dispositionalists, Black plays the role of a fink. When Black remains inactive, because he in no way alters or impedes the efficacy of the pertinent agential properties such as P_1-P_n, Jones retains the ability to do otherwise—just like the vase sitting on the shelf remains fragile despite that fact that, if toppled, a sorcerer would turn it to stone. Of course, were Black to intervene, Jones would lose that ability. But Black remains passive (like the sorcerer). So, when Jones acts freely, according to the new dispositionalists, and contra Fischer and Ravizza, he *retains* the ability to do otherwise.

Is the new dispositionalists' effort to undermine Frankfurt's argument successful? No. As Clarke (2009: 339–42) is careful to point out, the problem with the new dispositionalists' reply to source theorists is not that they fail to identify *an* ability that an agent like Jones retains in a Frankfurt example. Indeed, because when Jones acts on his own, none of his relevant dispositions are disturbed, there clearly is a kind of ability that he retains. But this ability can be described more carefully as a general capacity to do otherwise, or instead as a general ability to do otherwise that can, for instance, be exercised outside the context of Frankfurt examples, or when the agent is not asleep, or when she is not tied up, and so on.[21] Nevertheless, as Clarke remarks:

> [T]here apparently are abilities that Jones lacks, because of Black's readiness to intervene. Though Jones might have the capacity to act otherwise, the circumstances are not friendly to his exercising that capacity, and it may fairly be said that it is not up to him whether he exercises it, or that he does not have a choice about whether he does so. (2009: 340)

Furthermore, Clarke notes (341), Fischer, in articulating Frankfurt's argument, is careful to point out that what is in dispute is whether *under the circumstances* an agent can chose and do otherwise (Fischer, 2002: 304). The salient point is that the new dispositionalists are only able to claim victory in refuting Frankfurt's argument if the ability that they contend remains for an agent in a Frankfurt example is the same ability that is in dispute in debates about free will and moral responsibility—and it is this ability at which Frankfurt's argument is aimed.

Precisely the same point applies regarding the question of whether an agent can be reasons-responsive while acting within the context of a Frankfurt example. *If*, as Fischer and Ravizza suppose, reasons-responsiveness requires that an agent be able to react otherwise in response to sufficient reasons to do otherwise, then what is in question is whether *under the circumstances* of a Frankfurt example, an agent is able to react as required. Recall (from Section 3) Fischer and Ravizza's qualification regarding the counterfactuals in virtue of which MRR is confirmed; they do not involve worlds that are *accessible* to the agent (1998: 53). In the context of a Frankfurt example, it is not accessible to an agent like Jones to react otherwise in response to different reasons and, while in the presence of Black, not shoot Smith.

[21] Ann Whittle makes a similar point in her insightful (2010).

7. A NEW PROPOSAL: AN AGENT-BASED, REASONS-RESPONSIVE, SOURCE COMPATIBILIST THEORY

I have argued that a mechanism-based account of reasons-responsiveness is not a viable option because it requires a principled basis for mechanism individuation. Regrettably, this leads to insuperable problems fitting mechanisms for the degree of plasticity called for by proposals like MRR. But if the reasons-responsive theorist is forced to return to an agent-based theory, is it possible for her to fit such a thesis into the constraints of Frankfurt examples, and thereby account for source freedom? According to Fischer and Ravizza it is not; Frankfurt examples rule out the ability to do otherwise, and agent-based reasons-responsive theories require that agents who are suitably reasons-responsive be able to react, and so do, otherwise. As I have explained, in light of this assumed conflict, the new dispositionalists are positioned to claim that the winner is reasons-responsiveness. According to them, what must go is a commitment to Frankfurt's argument and the presumption that source freedom is freedom enough. But, the new dispositionalists can reason, this is all to the good, since their analysis of the free-will ability in terms of dispositions is, quite independently of any commitments to a reasons-responsive theory, sufficient to show that in a Frankfurt example an agent *is* able to do otherwise. The problem with this rejoinder is that the new dispositionalists have it wrong; they have not refuted Frankfurt's argument. What is at stake in a Frankfurt example is whether—*while in the context of a Frankfurt example*—an agent is able to exercise an ability to do other than as she does. And nothing in the new dispositionalists' playbook speaks to that issue.

So, assuming a commitment to Frankfurt's argument and a focus on source freedom, is it possible to develop a reasons-responsive theory without relying upon the problematic notion of mechanisms? I think it is. The single proposition standing in the way of fitting an agent-based reasons-responsive theory to the contours of source freedom is that an agent in a Frankfurt example is not reasons-responsive because, being unable to do otherwise, she is thereby unable to react otherwise when give sufficient reason to do so. I think this proposition should be rejected. I now wish to propose a simple solution to this puzzle. An agent who acts freely in a Frankfurt example, I shall argue, is suitably reasons-responsive despite being unable to react otherwise when given sufficient reason so to do. I offer two points in support of my claim.

First, the fact that an agent in a Frankfurt example is not able to react otherwise in response to sufficient reasons does not exhaust the resources available to show that, when she acts unimpeded, *she*, the agent, is reactive to sufficient reasons to do otherwise. Being reactive to sufficient reasons to do otherwise, I now propose, is not the same as being able to react otherwise in response to sufficient reason to do otherwise. In a Frankfurt example, an agent who is suitably reasons-responsive *is* reactive to sufficient reasons to do otherwise even if, due to the presence of a character like Black, she is unable to react otherwise.

To illustrate this point, consider again Jones outside the context of a Frankfurt example. Again suppose he shoots Smith for his own reasons. In this case, suppose that his actual reason for shooting Smith is revenge. Smith harmed Jones's family, and now, as Jones sees it, it's payback time. Let us also assign Jones a pattern of receptivity and reactivity that, we can stipulate, satisfies the conditions of MRR. Jones is receptive to reasons R_1 through $R_{z,}$ and is reactive to reasons R_1 through $R_{n.}$ Assume that the latter is a subset of the former in such a way that Jones's degree of reactivity is weaker than his degree of receptivity. Recall that reason R_1 as mentioned above is that Smith's child is with him, and were Jones to learn of this, he'd be receptive to this as a sufficient reason not to shoot Smith, despite his actual reason of revenge, and he would react accordingly, not shooting Smith. And so it would be for the entire spectrum of reasons R_1 through R_n, but not for the reasons ranging R_{n+1} through R_z. For this latter range, Jones would be receptive to these reasons as sufficient for not shooting Smith, but he'd shoot Smith regardless. Imagine, for instance, that reason R_{n+1} is that shooting Smith would cause Jones's mother to be disappointed with her son. While Jones would recognize this as a sufficient reason for not shooting Smith, sorry son that he is, he'd not be reactive to it; he'd shoot Smith for his own reasons of revenge, despite the grief he'd knowingly cause his dear mum. Jones, given this set up, for the range of reasons R_1 through R_n is both reactive to sufficient reasons to do otherwise, *and* would react otherwise in light of such reasons.

Now reinsert Black. Note that Black need not interfere with Jones were Jones to face up to any of the reasons within the spectrum of R_{n+1} through R_z. Jones is not reactive to these reasons in the sense that they would not deter him from acting on his own reasons for shooting Smith. So Black is able to leave Jones to function on his own. Should Jones consider the reason R_{n+1}, that his mother would be disappointed with him if he shot Smith, he'd shoot Smith on his own anyway. Hence, Black can leave well enough alone. But as for the reasons R_1 through R_n, were any of these reasons to become relevant to Jones's practical context, Black would intervene. In being prepared to do so, clearly Black makes it the case that Jones is unable to react otherwise in response to these reasons. But it is nevertheless true that Jones *is reactive* to this spectrum of reasons R_1 through R_n in just this sense: for each reason, were it to become salient for Jones, it is not the case that, given his own intrinsic agential condition, he would act on his own reasons of revenge for shooting Jones. He is at least reactive to reasons in this manner. In reacting to this spectrum of reasons (at relevant possible worlds), he, by virtue of his own agency and his own reasons of revenge, is not the cause of shooting Smith. This is so despite that fact that he is unable to act otherwise and thereby avoid shooting Smith. While acting within the context of a Frankfurt example, Jones *is reactive* to this spectrum of sufficient reasons to do otherwise, R_1 through R_n, *despite the fact that he is not able to react otherwise.* His being reactive in this context consists in his not persisting in acting on his own reasons of revenge to shoot Smith in light of reasons such an R_1 (Smith's child is with him).

Drawing upon the metaphysics of causation, Carolina Sartorio has made important contributions to our understanding of source compatibilist accounts of freedom, one of which is relevant to my current proposal. In threat-cancelation scenarios, Sartorio (forthcoming) points out, causation is not transitive along the path of a single causal chain. And the contrary to fact scenarios in Frankfurt examples—that is, the scenarios in which the intervener becomes activated—are set up as threat-cancelation scenarios. To explain: Jones, given his own reasons for shooting Smith, is such that, in the presence of certain reasons, such as R_1 (Smith's child is with him), he would not shoot Smith. Suppose that, indeed, Jones *is* in the presence of this reason, R_1. This creates a threat to the event of his shooting Smith; it's the sort of thing that's liable to lead to Jones *not* shooting Smith. The presence of this threat, as a link in a causal chain, *then* causes the intervener to bring it about that Jones shoots Smith. Hence, there *is* a causal chain from Jones and his relation to R_1 through to his shooting Smith. But because of the way this causal chain unfolds, Jones, given the way he is and his own reasons prior to the intervention, does not cause his shooting of Smith. Why? Because, the way he is disposed to respond to reasons like R_1 creates a threat to his shooting Smith. So, despite the fact that there is a causal chain from Jones to the event of his shooting Smith via Black's manipulation, Jones—just as he is as an agent, given his intrinsic properties—is not the cause of shooting Smith. Thus, as Sartorio makes clear, the difference between actual and counterfactual cases in Frankfurt examples is a difference in the causal role played by agents such as Jones. This, I maintain, helps cast light on the point I am at pains to emphasize here. The point has to do with the way Jones is reactive to the different kind of reasons that might bear on exercises of his agency. Some reasons, such as R_{n+1}, are such that, in their presence, Jones persists in being the right kind of agential cause of his action (of shooting Smith). Some, such as R_1, are such that, in their presence, he does not persist in being the right kind of agential cause of his action.

The first point I am offering here bears some similarity to earlier failed efforts to defend PAP against Frankfurt's argument by distinguishing acting on one's own and not acting on one's own, which seems to be a kind of alternative that cannot be expunged from Frankfurt examples (e.g. see Naylor, 1984). This strategy for defending PAP was handily rejected because the alternative of acting or not acting on one's own in suitably structured Frankfurt examples were not alternatives that were themselves within the voluntary control of the agent (Fischer, 1994). Why doesn't this apply to my current proposal?[22] Here's why: My focus on the distinction between an agent's acting upon her own reasons in an actual scenario and her not acting upon her own reasons in alternative scenarios is not being used to underwrite any claim about an agent's freedom to do otherwise. It is only being used to call attention to a fact about the mode of an agent's acting in the actual-world scenario in which she acts on her own. The only reason that in the alternative scenario the counterfactual intervener causes Jones to act as he does is because, given Jones's state, just as he is as an agent, he would not act on his own reasons of revenge for shooting Smith were these other reasons also salient. Granted, in the alternative scenario in which the intervener forces him to act, Jones does not do anything intentional to make it so that he does not act on his own reasons to shoot Smith; this scenario is not within the scope of his voluntary control. But there is no claim here that the degree of reactivity displayed by Jones in this range of counterfactuals affords Jones anything like a freedom to do otherwise. These counterfactual scenarios are not alleged to be accessible to Jones from the actual world. But, I say, they still display that, given the way he is in the actual world, he is in some manner reactive to pertinent reasons to do otherwise.

Perhaps some will find my argument here too thin. They might demand that being reactive to sufficient reasons to do otherwise requires that, in the presence of the relevant range of reasons, the agent's own responsiveness-grounding resources have to be directly causally involved in the mode of differential reactivity. And when Black the intervener is in the driver's seat—that is, when Black has taken over and is generating the action—that's just not what is going on. In fairness, I grant that this is a reasonable source of resistance. However, in my own estimation, those inclined to press it in this context are fixed upon an alternative possibilities model of

[22] Thanks to both Derk Pereboom and Brandon Warmke for raising this worry.

freedom, which the current proposal is meant to reject. Nevertheless, there is a distinct way of arguing that an agent in a Frankfurt example is suitably reactive to reasons, which I shall now explore.

Second, return to the complex counterfactuals which the new dispositionalists exploit for the purpose of advancing a compatibilist theory of the ability to do otherwise. Recall this one, where "R_1" names the reason that Smith's child is with him:

> RR_1: If Jones were to become aware of R_1, and if Jones retained during the relevant duration of time intrinsic agential properties P_1-P_n, and if Jones were not interfered with in a way that would impede the casual efficacy of those properties, then Jones would not shoot Smith.

As noted, to test for the truth of RR_1, we have license to go to worlds in which a counterfactual intervener like Black is missing. This is similar to the way that Fischer and Ravizza test counterfactuals regarding the dispositional properties of the mechanisms of action. The reason that factoring out Black in these contexts is felicitous is because it is what is needed to test the dispositional properties up for consideration. Now, the new dispositionalists might be wrong to think that counterfactuals such as RR_1 can be used in the service of giving an adequate account of the dialectically relevant ability to do otherwise. And Fischer and Ravizza might be wrong to think that we should draw on similar counterfactuals to account for the dispositional properties of mechanism rather than agents. But the counterfactuals themselves, such as RR_1, still tell us something important about the *agents* in question. They call to our attention dispositional or modal properties that these agents really do have—even in the context of Frankfurt examples. Jones really is disposed to respond to a reason such as R_1 in certain ways. Were he left unfettered to act as he wished, in relevant contexts, he would act accordingly.[23]

Why is the preceding point so important? The challenge presently before us is to account for an agent's reactivity to reasons in contexts in which she is not able to react otherwise, due to the presence of a counterfactual intervener like Black. The point currently on offer is that counterfactuals such as RR_1, and in particular a sequence of them, say in a patterned form of RR_1 through RR_n help to make what is in essence the same basic point I was at pains to emphasize early on (Section 2). These counterfactuals aid in demonstrating that when an agent like Jones acts as he does on his own,

[23] Ginet (2006: 235–6) makes a similar point when advising Fischer and Ravizza to give up the mechanism component of their theory and instead just consider the reasons-responsiveness of agents.

the act he actually does perform is *itself* a reaction or a response to the actual conditions in which he acts. Furthermore, what these counterfactuals collectively help to underscore is not merely that Jones was indeed the cause of his so acting, but that the *manner* of his causing his act was sufficiently sensitive to reasons.[24] In this way, when an agent in a Frankfurt example acts, she is, so to speak, *being* suitably reactive in acting as she does, and so is *exercising* a kind of source freedom, a kind that can be characterized in terms of responsiveness to reasons.

One residual worry about the second point I offer here concerns cases in which, intuitively, an agent acts freely, but in which there are conditions intrinsic to her own agency such that these conditions play the role of a fink or a mask analogous to the role played by an extrinsic counterfactual intervener like Black. For instance, suppose that if Jones were not about to shoot Smith for the very particular sane reasons he had, then rather than Black intervening, some psychotic episode would unfold, leading to his shooting Smith anyway. Or we could instead run the case in terms of some latent phobia that would arise in the case of certain reasons but not others. If we hold fixed *all* of the intrinsic properties constituting Jones, we'd have to include these psychotic-constituting or phobia-constituting properties as well. In such cases, the relevant counterfactuals would come out false as applied to the agents on the proposal currently on offer. But if one were able to identify just certain features of the agent, as mechanisms, and hold these fixed, one might still be able to account for responsiveness. Clearly, these kinds of cases push back in the direction of a mechanism-based view.[25]

[24] An anonymous referee for OUP has raised the following insightful objection: Here I advocate an account of an agent's reactivity by claiming that the pertinent counterfactuals help to reveal the manner of the *agent's* causing her action. But, as I myself have noted above, Fischer and Ravizza themselves contend that they do not mean to reify the notion of a mechanism, and all they mean by appeal to an *agent's mechanism* is the "manner that an action was caused." Why, the referee asks, is my proposal any different from Fischer and Ravizza's? This is an especially keen question, but the answer, to my mind, is telling in a way that speaks against Fischer and Ravizza and on behalf of my agent-based proposal. Even if Fischer and Ravizza mean to pick out no more than "the way an action is brought about," they are at least committed to the existential claim of there being such a way, and to holding *that* fixed when testing it for responsiveness to different reasons. My proposal is that what gets held fixed for the pertinent testing is *the agent*. Now, *either* Fischer and Ravizza will grant that they mean the same thing, in which case their appeal to mechanisms really is the functional equivalent of the notion of agency itself and they might as well drop talk of mechanisms altogether, *or* they will insist that they mean to fix on something narrower than the agent. In that case we have just the very difference I have been at pains to bring out in this essay.

[25] Stephen Kearns, Nicole Smith, David Sosa, and Jada Strabbing each pressed this challenging concern in slightly different ways.

Previously, it was precisely because of cases like these that I had been convinced that reasons-responsive theorists could not recover an agent-based theory while remaining committed to source compatibilism. But upon reflection, I do not think that these sorts of cases should carry the day. Admittedly, they are problematic. However, there are two quite different strategies for resisting their threat to the current agent-based proposal for explaining reasons-reactivity. One is to resist what seems intuitive in such cases. It's not entirely clear that in such cases the agents really do act freely even when the pertinent psychosis or the phobia is inert. The agent's "success" in acting uninfluenced by the troubling latent condition seems too fluky. The agent, after all, seems highly constrained in responding to just the reasons she does, so much so that her own agential resources impede her from responding to relevant patterns of reasons. A very different strategy would be to consider whether, when specifying the agent-constituting intrinsic properties that are to be held fixed in counterfactuals like R_1, there is a principled way to rule out those constituting the psychosis or the phobia. The rough idea would be to treat these conditions as in some way alien or distant from those ingredients constituting the agent's identity, or her real self.[26] This is a promising avenue worth pursuing, though I'll not do so here. It is enough, it seems to me, just to make clear that cases such as these pose a threat to the current proposal, but that there are avenues for addressing them which allow preservation of this basic claim. When a range of counterfactuals like RR_1 through RR_n *are* true of an agent, these *do* help to show how it is that when she acts as she does, she is reacting to reasons, even if she cannot react otherwise.

For a long while now, at least since the appearance of Fischer and Ravizza's excellent and highly influential *Responsibility and Control* (1998), it has been assumed that the only way to marry a reasons-responsive theory of freedom to the lesson learned from Frankfurt's argument is to forgo an agent-based theory. The way to account for source freedom in terms of reasons-responsiveness, I had previously thought, is by shifting to a mechanism-based theory. In light of the two preceding points, combined with the difficulties involved in accounting for mechanism individuation, I believe that this is mistaken. I propose an agent-based, reasons-responsive compatibilist theory of source freedom.

[26] I am indebted to Michael Bratman for this point. He drew upon it to point out that this gives us reason to consider joining the resources of a reasons-responsive strategy with those of mesh theories like Frankfurt's. The latter seek to make sense of something like the boundaries of the "real self" as narrower than the full spectrum on an agent's psychological profile. (Though Bratman then raised the question of whether such a joint venture should be regarded as a friendly or instead a hostile takeover.)

APPENDIX I: WHY NOT A MESH THEORY RATHER THAN A REASONS-RESPONSIVE THEORY?

In the wider philosophical community, perhaps the best known source compatibilist account of freedom is *not* a reasons-responsive theory. It is Frankfurt's own, which he expresses in terms of acting freely and of one's own free will (1971). Frankfurt developed his position by attending to a fitting relation between higher-order and lower-order desires. An agent acts of her own free will when the first-order desire moving her all the way to action is the one that, at a higher order, she identifies with and so most wants to act upon. Here, the further details do not interest me. What does is the general strategy. Frankfurt's is one version of a mesh theory because it accounts for freedom in terms of a harmonious mesh between different sub-systems within an agent's overall mental economy.[27] On Frankfurt's approach, an agent's freedom consists in the well-functioning relation between different orders of desire. When these different elements "line up," the agent acts unencumbered, and so she acts freely. But when, for instance, what Sally most wants is for her desire not to drink the beer to win out, and when, despite that, her first-order desire to drink the beer leads her all the way to action, Sally acts from an unharmonious mesh. She is, in a sense, impeded or encumbered by her own psychological constitution and so, according to Frankfurt, is not free. The system issuing in her actions is out of alignment with the system that at a higher order constitutes what she most wants.

When reflecting upon cases in which an agent acts from a harmonious mesh, it is easy to see how that agent's freedom, so construed, could easily be slipped into a Frankfurt example. So long as her action does actually issue from the mesh, there is no reason to be concerned about alternatives to what an agent does do. All that matters in such a case is that her well-functioning mesh operated unfettered. Thus, mesh theories such as Frankfurt's fit seamlessly into the actual-sequence demands of Frankfurt examples. Why then explore the prospects of a reasons-responsive theory at all? Why not opt for some variation on a mesh theory?

In my estimation, all mesh theories, not just Frankfurt's, face a deep problem for those committed to source compatibilism, one that I suspect is insurmountable. Mesh theories might comfortably capture conditions sufficient for acting freely: When an agent acts from a harmonious mesh, she acts freely. But in many cases in which an agent acts from an unharmonious mesh, mesh theories either generate the wrong results, or are instead unsatisfactorily silent—at least this is so for those committed to source compatibilism. Mesh theories generate the wrong results when they treat acting from a harmonious mesh as also necessary for

[27] For other efforts to develop a mesh theory, see Dworkin (1970), and Watson (1975). Bratman (2007) has explored similar ideas. For a more detailed critical discussion of mesh theories, see McKenna (2011).

acting freely; they are unsatisfactorily silent when they refrain from accounting for an agent's freedom in such cases.

To explain, consider the case of Sally above. Grant that on Frankfurt's view, acting from a harmonious mesh is not only sufficient but also necessary for free action. Maybe Sally acts from a freedom-compromising compulsion in drinking that beer; suppose she is a full-blown, hopeless alcoholic. If so, there is no threat to Frankfurt's mesh theory; it yields the result that Sally does not act freely insofar as she fails to satisfy the necessary condition at issue. But maybe instead Sally *freely* acts upon her lower-order desire as in opposition to the higher-order preferences with which she identifies. Suppose she also judges it best not to drink the beer. This is a classic case of weakness of will, understood in terms of *freely* acting contrary to one's better judgment. A mesh theory such as Frankfurt's generates the wrong result in a case like this. It has it that Sally does not act freely when in fact she does.

A viable option for the mesh theorist is to claim that the necessary condition which she proposes is merely an *ability* to act from a harmonious mesh. This would allow the mesh theorist to handle the distinction between exercises of free, weak-willed agency and failures that are due to freedom-undermining compulsion. The weak-willed agent, it can be argued, did not act from a harmonious mesh, but it remains true that she had the ability to do so; she could have done otherwise. Obviously, this reply will not work for any mesh theorist who, like Frankfurt, is also committed to source compatibilism, since it relies on claims about leeway freedom.

It is open to the mesh theorist who *is* committed to source compatibilism to retreat by weakening the reach of her theory. She could give up any proffered necessity condition and claim that she is only offering sufficient conditions for free action in those cases involving acting from a harmonious mesh. These conditions can be shown to be compatible with determinism without in any way relying upon any assumptions about leeway freedom. After all, this is all that is strictly required for her to answer the metaphysical challenge of the incompatibilist: the incompatibilist contends that determinism is incompatible with free action; she has produced cases of free action from a harmonious mesh that are compatible with determinism. Case closed. The objection to mesh theories currently under consideration concerns a refutation of the necessary conditions they propose. Why can't the mesh theorist simply retreat by relinquishing this necessary condition?

Such a retreat strikes me as an unpromising dialectical tactic. A mesh theory that remained silent with respect to acting freely from an unharmonious mesh would leave unaccounted for a wide swath of (putatively) free actions. A credible theory of freedom should be able to account for these sorts of mundane cases of agency. And so a theoretical lacuna would invite further attempts to account for freedom in these cases. But then, once these kinds of cases are accounted for in ways that would be satisfying to compatibilists, it will only be natural to consider whether the proper account can *also* explain the freedom present in the cases involving a harmonious mesh. If so, then the work of accounting for free action will be done more systematically with these alternative resources. Hence, the work done by the mesh theory will be rendered otiose. This is, in fact, how I see the explanatory resources of a reasons-responsive theory of freedom.[28]

APPENDIX II: ONE POINT WHERE FISCHER AND RAVIZZA REQUIRE AN ACCOUNT OF MECHANISMS

To illustrate a place where, without an independent rationale for doing so, Fischer and Ravizza are too quick to allow sameness of mechanism to fall their way, consider their contention that reactivity is all of a piece (1998: 73). They offer this proposal in response to the incompatibilist worry that in the actual world, assuming a deterministic context, when a blameworthy agent fails to act on the moral reason to do otherwise, the agent cannot react to this reason. After all, holding fixed the past and the laws, when that reason is present in the nearest possible world (the actual one), the agent does not react to it. Fischer and Ravizza attempt to ward off this objection by arguing that an agent's reacting differently to a different reason to do otherwise in another possible world is sufficient to establish that she is able to react differently to *any* reason to do otherwise. Of course, if this were so, she would be able to react differently to the moral reasons that were present in the actual scenario, even though she did not. It is in this context that Fischer and Ravizza (1998: 73–4) consider the following incompatibilist-friendly challenge to their 'reactivity is all of a piece' thesis (which I paraphrase here):

> Is it not possible for the same mechanism to get more energy or focus from different incentives? If it were, then the fact that an agent would react to *some* reasons to do otherwise, does not mean that she is able to respond to *all* reasons to do otherwise, and so it does not mean that she is able to react differently to the actual moral reason to do otherwise that was present to her at the time of action.

If mechanisms were to behave this way—that is, in a way that disconfirms that reactivity is all of a piece—then Fischer and Ravizza would be forced into a difficult dialectical corner. They would be forced to take head on the incompatibilist worry that, in a deterministic context, a blameworthy agent in the actual situation in which she acts is not able to react to the moral reasons that were in fact present at the time.

[28] I do, however, think that there is a way for mesh theories and reasons-responsive theories to form a mutually beneficial alliance. The basic idea would be that a mesh theory could be used to account for the nature of agency more generally (rather than free agency). In particular, on the proposal I am considering, only agents able to adopt higher-order attitudes towards their motivations can even be regarded as candidates for having the kind of complexity for free agency as specified by something like MRR. Note this might very well afford reasons-responsive theorists the resources to handle the sorts of problems alluded to above. (See especially note 26).

They would have to show directly that, contrary to what the incompatibilist claims, such an agent is able to react differently to the particular moral reasons that were present, or they would have to commit to the claim that an agent could be blameworthy for failing to act on moral reasons that she was not able to act upon.[29]

To this challenge, they counter that in those cases in which an *agent* reacts differently to some incentives only because she acquires more energy or focus, it is natural to say that it is because a *different* mechanism of action is operative (74). This is one particular point where, I contend, Fischer and Ravizza are too quick to allow sameness of mechanism—or, rather, in this case, difference of mechanism—to fall their way (McKenna, 2001: 98–9). They claim that it is natural to think of the pertinent mechanisms as they propose. But it is hard to see how it is natural. It's not clear in the first place what a mechanism of action is. It's not a term of folk psychology that has accrued enough use to allow for much in the way of intuitive baggage. Why can't a mechanism of action be reactive under, say, extreme pressures, but fail in normal contexts? After all, literal mechanisms, like thermostats, behave this way all the time, especially when they begin to fail. Indeed, what is the exception is to find an actual mechanism—a manmade gizmo of some sort—that does not respond differentially under extreme pressures, that does not have foibles and limits, and so on.

Fischer has responded to my objections by likening the problem I raise here to generality problems in other areas of philosophy. A certain level of generality, and thus lack of specification, is acceptable in other domains of inquiry, such as ethical theory. So too, Fischer argues, is a degree of generality acceptable when theorizing in terms of mechanisms of agency (Fischer, 2004: 169). Fischer also points out that it is unreasonable to demand of a successful theory that all of its elements are fully analyzed (168). Thus, he resists the burden of offering any "purely 'principled' account of mechanism individuation—an account that did not at some level appeal to intuition" (167–8). He thus remains committed to relying exclusively on appeal to intuitions in response to difference cases.

While I sympathize with Fischer's thoughtful reply, it is not enough to put the objection to rest. Requesting *some* principled basis of mechanism individuation is not the same as demanding a full analysis of mechanisms—one that is "purely" principled. And a degree of generality is of course acceptable in theorizing in areas such as this one. But when the degree of generality and an *exclusive* reliance on intuition will not help adjudicate differences between cases crucial to assessing the theory, then demanding *some* principled basis for settling the dispute is only dialectically fair and reasonable. This is fully consistent with accepting Fischer's contention that appeal to intuitions ought to enter the scene at some level.

[29] I have defended the former of these two options (2005), while Pereboom (2006) has advised Fischer and Ravizza to opt for the latter. Watson has also commented on this issue (2001, as appearing in 2004: 300–1). He suggests that on pain of consistency, Fischer and Ravizza should opt for the latter option.

REFERENCES

Ayer, A. J. (1954). "Freedom and Necessity." In his *Philosophical Essays.* (New York: St Martin's Press), 3–20.

Bratman, Michael (2007). *Structures of Agency.* (New York: Oxford University Press).

Buss, Sarah and Lee Overton (2002). *Contours of Agency: Essays on Themes from Harry Frankfurt.* (Cambridge, MA: MIT Press).

Clarke, Randolph (2009). "Dispositions, Abilities to Act, and Free Will: The New Dispositionalism." *Mind* 118: 323–51.

Davidson, Donald (1973). "Freedom to Act." In Honderich (ed.), 1973: 67–86.

Dworkin, Gerald (1970). "Acting Freely." *Noûs* 4: 367–83.

Fara, Michael (2008). "Masked Abilities and Compatibilism." *Mind* 117 (468): 843–65.

Fischer, John Martin (1994). *The Metaphysics of Free Will.* (Oxford: Blackwell Publishers).

—— (2002). "Frankfurt Style Compatibilism." In Buss and Overton (eds.), 2002: 1–26.

—— (2004). "Responsibility and Manipulation." *Journal of Ethics* 8. 2: 145–77.

—— (2006). "The Free Will Revolution (Continued)." *Journal of Ethics* 10, No. 3: 315–45.

—— (2011). *Deep Control.* (New York: Oxford University Press).

—— and Mark Ravizza (1998). *Responsibility and Control: An Essay on Moral Responsibility.* (Cambridge: Cambridge University Press).

Frankfurt, Harry (1969). "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66: 829–39.

—— (1971). "Freedom of the Will and the Concept of a Person." *Journal of Philosophy* 68: 5–20.

Ginet, Carl (1966). "Might We Have No Choice?" In Lehrer, 1966: 87–104.

—— (2006). "Working With Fischer and Ravizza's Account of Moral Responsibility." *Journal of Ethics* 10: 229–53.

Haji, Ishtiyaque (1998). *Moral Appraisability.* (New York: Oxford University Press).

Hobart, R. E. (1934). "Free Will as Involving Indeterminism and Inconceivable Without It." *Mind* 43: 1–27.

Honderich, Ted (ed.) (1973). *Essays on Freedom and Action.* (London: Routledge & Kegan Paul).

Howard-Snyder, Daniel and Jeff Jordan (eds.) (1996). *Faith, Freedom, and Rationality.* (Lanham, MD: Rowman and Littlefield).

Hunt, David P. (2000). "Moral Responsibility and Unavoidable Action." *Philosophical Studies* 97: 195–227.

Irwin, T. H. (1980). "Reason and Responsibility in Aristotle." In A.O. Rorty (ed.), 1980: 117–55.

Lewis, David (1997). "Finkish Dispositions." *Philosophical Quarterly* 47: 143–58.

McKenna, Michael (2001). "Review of John Martin Fischer and Mark Ravizza's *Responsibility & Control.*" *Journal of Philosophy*. XCVIII, no. 2: 93–100.

—— (2003). "Robustness, Control, and the Demand for Morally Significant Alternatives." In *Moral Responsibility and Alternative Possibilities* (eds.), Widerker and McKenna. 2003.

—— (2005). "Reasons Reactivity & Incompatibilist Intuitions." *Philosophical Explorations* 8. 2: 131–43.

—— (2008). "Frankfurt's Argument against Alternative Possibilities: Looking Beyond the Examples," *Noûs* 42: 770–93.

—— (2011). "Contemporary Compatibilism: Mesh Theories and Reasons-Responsive Theories." In R. Kane (ed.), 2011, *Oxford Handbook of Free Will*, 2nd edn. (New York: Oxford University Press), 175–98.

Martin, C. B. (1994). "Dispositions and Conditionals." *Philosophical Quarterly* 44: 1–8.

Mele, Alfred (1995). *Autonomous Agents.* (New York: Oxford University Press).

—— (2006a). *Free Will and Luck.* (New York: Oxford University Press).

—— (2006b). "Fischer and Ravizza on Moral Responsibility." *Journal of Ethics* 10. 3: 283–94.

—— and David Robb (1998). "Rescuing Frankfurt-Style Cases." *Philosophical Review* 107: 97–112.

Moore, G. E. (1912). *Ethics.* (Oxford: Oxford University Press).

Naylor, Marjory (1984). "Frankfurt on the Principle of Alternate Possibilities." *Philosophical Studies* 46: 249–58.

Pereboom, Derek (2001). *Living Without Free Will.* (Cambridge: Cambridge University Press).

—— (2006). "Reasons Responsiveness, Alternative Possibilities, and Manipulation Arguments Against Compatibilism: Reflections on John Martin Fischer's *My Way*." *Philosophical Books* 47: 198–212.

Sartorio, Carolina (forthcoming). "Actuality and Responsibility." *Mind.*

Smith, Michael (2003). "Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion." In Stroud and Tappolet (eds.), 2003: 17–38.

Stroud, Sarah and Christine Tappolet (eds.) (2003). *Weakness of Will and Practical Irrationality.* (New York: Oxford University Press).

Stump, Eleonore (1996). "Libertarian Freedom and the Principle of Alternative Possibilities." In Howard-Snyder and Jordan, 1996: 73–88.

Van Inwagen, Peter (1975). "The Incompatibility of Free Will and Determinism." *Philosophical Studies.* 27:185–99.

Vihvelin, Kadri (2004). "Free Will Demystified: A Dispositional Account." In *Philosophical Topics* 32: 427–50.

Wallace, R. Jay (1997). "Review of John Martin Fischer's *The Metaphysics of Free Will.*" *Journal of Philosophy* 94. 3: 156–59.

Watson, Gary (1975). "Free Agency." *Journal of Philosophy* 72: 205–20.

—— (2001). "Reason and Responsibility." *Ethics* 111: 374–94.

—— (2004). *Agency and Answerability.* (New York: Oxford University Press).

Whittle, Ann (2010). "Dispositional Abilities." *Philosophers' Imprint* 10. 12: 1–23.

Widerker, David (2006). “Libertarianism and the Philosophical Significance of Frankfurt Scenarios.” *Journal of Philosophy* 103: 163–87.

—— and Michael McKenna (eds.) (2003). *Moral Responsibility and Alternative Possibilities*. (Aldershot, UK: Ashgate Press).

Wiggins, David (1973). “Towards a Reasonable Libertarianism.” In Honderich 1973: 31–62.

Zagzebski, Linda (2000). “Does Libertarian Freedom Require Alternate Possibilities?” *Philosophical Perspectives* 14: 231–48.

7

Responsibility, Naturalism, and "The Morality System"*

Paul Russell

> Theory typically uses the assumption that we probably have too many ethical ideas . . . Our major problem now is actually that we have not too many but too few, and we need to cherish as many as we can.
>
> Bernard Williams, *Ethics and the Limits of Philosophy*

Lying at the heart of P. F. Strawson's core strategy in "Freedom and Resentment" is his effort to direct his naturalist claims and observations against not only the philosophical extravagance of a general skepticism about moral responsibility but also against all nonskeptical attempts to provide responsibility with some form of external rational justification (Strawson 1962: 23). According to Strawson, efforts of this kind are not only misguided and unconvincing in themselves, when they fail they encourage a general skepticism about moral responsibility. The alternative strategy that Strawson pursues is one that places the foundations of responsibility on our natural, universal emotional propensities and dispositions relating to moral sentiments or reactive attitudes. This naturalistic turn invites us to focus our attention on familiar facts about human moral psychology, and to drop our focus on the analysis of the concept of "freedom" as a way of dealing with the threat of determinism. Beyond these general features, however, the details of Strawson's strategy become both complex and layered. As a consequence of this, interpretations and assessments of his arguments differ greatly. There

* I am grateful to David Shoemaker and other participants at the New Orleans Workshop on Agency and Responsibility (November 2011) for their very helpful comments and criticism. Versions of this paper were also presented at a Workshop on Moral Agency, Deliberative Awareness, and Conscious Control (Erasmus University, Rotterdam, October 2010) and at the University of Arizona (February 2012). Thanks also to those who were present on those occasions for their comments and suggestions, and especially to Michael McKenna, Maureen Sie, and Andras Szigeti.

is, nevertheless, a general consensus among both followers and critics alike that there are significant strands in Strawson's specific naturalistic arguments that are implausible and unconvincing and that some "retreat" from the original strong naturalist program that he advanced is required.

In this paper I take up two closely related issues arising out of this overall problem. First, I want to consider if the right way to amend and modify the naturalistic program is to adopt a "narrow" construal of our moral reactive attitudes along the lines proposed by R. Jay Wallace, one of Strawson's most prominent followers on this subject. The narrower approach, as I will explain, involves a substantial retreat from Strawson's original naturalistic program and has significant implications for Strawson's core claim that our commitment to responsibility requires no external rational justification. The second issue that will be addressed is whether or not we should interpret moral responsibility entirely within the confines of what Bernard Williams has described as "the morality system." (Williams 1985: ch.1). The narrow construal of responsibility requires that we understand moral responsibility within the conceptual resources provided by the morality system, making notions of obligation, wrongness, and blame essential to the analysis of moral responsibility. There is, therefore, an intimate connection between these two issues. With respect to both these issues I will argue that the narrow approach, while it has legitimate criticisms to make of Strawson's original strategy, nevertheless takes us in the wrong direction and involves an unacceptable distortion and truncation of moral responsibility.

1. TWO MODES OF NATURALISM

Strawson's naturalistic account of moral responsibility insists that a proper understanding of this matter must begin with a description of what is involved in *holding* a person responsible. The key to his analysis is the role that reactive attitudes play in this sphere, where this is understood in terms of our natural emotional responses to the attitudes and intentions that we manifest to each other. Strawson's naturalistic strategy, as based on this general observation, has two aspects or dimensions that need to be carefully distinguished—although this is not done in his own presentation. There is, in the first place, a strong form of naturalistic argument that involves the claim that even if we had some theoretical reason to abandon entirely or altogether suspend our reactive attitudes (e.g. as required by skeptical arguments based on considerations of determinism), it would be psychologically impossible for us to do this (Strawson 1962: 9–12,18). A systematic repudiation of all reactive attitudes of this kind, he argues, would result in "a thoroughgoing objectivity of attitude" to others, with the resulting loss of all our interpersonal personal relations (1962: 12–3). While the objective attitude may be appropriate when we are dealing with the abnormal and incapacitated, and may even on occasion be available to us when dealing with normal adults, we cannot "do this for long, or altogether" (1962: 9–10). Armed with these claims, Strawson suggests an easy way to deal with the skeptic. No matter what theoretical arguments may be presented, our human nature is such that we will continue to feel or entertain *tokens* of reactive attitude. According to this view, skeptical arguments will inevitably fail to dislodge or wholly eradicate these attitudes and responses. Let us call this strong line of reply token-naturalism, since it turns on the claim that our tokens of reactive attitudes cannot be systematically eliminated or abandoned whatever (philosophical) arguments may be advanced against them.

Most philosophers have found Strawson's token-naturalism too strong and unconvincing. There are two basic objections to be made against it.[1] First, the psychological claim that it makes is doubtful in point of fact. It is not at all obvious that we are constitutionally incapable of entirely ceasing to entertain tokens of reactive attitude should we be persuaded that they are systematically unjustified. Moreover, even if the psychological claims were true, this does nothing to remove or discredit the objections put forward by the skeptic, since these claims do not address the justificatory issue that concerns us. It would, in fact, be disturbing to discover that we will naturally continue to entertain tokens of reactive attitude in face of well-founded grounds for discrediting these emotional responses to others.

While Strawson's token-naturalism may be judged too strong, there is a weaker form of naturalism that is more convincing and has attracted greater interest. What is crucial to naturalism, on this account, is that we are all *liable* or *disposed* to reactive attitudes. It is from within the framework of this weaker form of naturalism that Strawson develops his general rationale of excuses and exemptions. While we may all have a natural liability to reactive attitudes, particular tokens can be discredited by reference to excuses and exemptions. Excuses operate by way of showing that the agent's conduct was consistent with the underlying general demand for "some degree of goodwill or regard" (1962: 7). In cases of this kind, involving an accident, ignorance and other such factors, the conduct in question does not manifest any ill-will or malicious intent, even if some injury has

[1] For a more detailed development of this analysis of Strawson's naturalism see Russell (1992).

occurred. With respect to exemptions, however, we are asked to view the individual as an altogether inappropriate target of our reactive attitudes on the general ground that we are dealing with someone for whom we cannot make the moral demand due to an abnormality or incapacity of some relevant kind. Strawson employs this two-level account of excuses and exemptions to show that the thesis of determinism fails to engage any of these recognized considerations and cannot, therefore, constitute a basis for systematically discrediting all our reactive attitudes associated with moral responsibility.

This weaker naturalist approach may be described as a form of *type*-naturalism, where this is understood in terms of our natural disposition to reactive attitudes (just as we have a natural liability to love, fear, joy and other basic emotions). What is crucial, however, is that unlike token-naturalism, type-naturalism offers no easy way of discrediting the skeptic. On the contrary, it is essential, on this approach, that a plausible account of excusing and exempting conditions is provided consistent with compatibilist commitments. At this level, concerning our natural liability to reactive attitudes, we must still engage in arguments that counter the skeptical challenge. At the same time, type-naturalism does insist on the "internal" nature of these replies to the skeptic (1962: 23). While it is possible that our token reactive attitudes could be systematically discredited from within, there remains no need or possibility of providing an external, rational justification of a more general kind for our (natural) propensity to these emotions (any more than we need to do this for our similar liability to love, fear, etc.). Although justificatory issues remain with us, and cannot be evaded by way of token-naturalist claims, these justificatory requirements do not take the form of a demand for general or external rational justifications.

2. REACTIVE ATTITUDES AND NARROW RESPONSIBILITY

Having established a distinction between token- and type-naturalism, let us now consider Wallace's amended account of the Strawsonian project, as developed on the basis of his narrow construal of the reactive attitudes. Wallace's compatibilist account weaves together two distinct strands of thought. The first is a broadly Strawsonian description of holding people responsible, interpreted in terms of our reactive attitudes (Wallace 1994: 8–12). The other strand is his Kantian theory of reflective self-control or moral agency (1994: 12–15). Taken together, these two strands constitute what Wallace calls his "normative interpretation" of responsibility, which maintains that the correct way to understand what it is to be a morally responsible agent is by way of describing those conditions under which it is *fair* to hold an agent responsible (1994: 5,15, 64). Our stance of holding a person responsible must itself be understood in terms of the mutual dependence between expectations and reactive attitudes (1994: 20–5). To hold someone to an expectation, or a demand of some kind, is to be susceptible to reactive attitudes when the expectation is violated (1994: 21). We are susceptible to a reactive attitude if we either feel this emotion or believe that it would be appropriate to feel it in these circumstances (1994: 23, 62). While these moves are generally consistent with Strawson's original approach, Wallace aims to substantially modify and amend this approach by providing a narrower and more fine-grained account of moral reactive attitudes.

According to Wallace, Strawson's account of reactive attitudes is too "inclusive" and needs to be refined into several different categories relating the various emotions we are concerned with (1994: 25–40). On Wallace's analysis we need to draw two overlapping distinctions (1994: 33). The first is between moral and nonmoral reactive attitudes. Both forms of reactive attitudes depend on their reciprocal relationship with a system of expectations. In the case of moral reactive attitudes the relevant set of expectations are justified with reference to moral reasons, and the expectations they justify are obligations (1994: 35–6). However, not all expectations are backed by moral reasons, as we find in the case of etiquette, where a breach may generate resentment even though no distinctive moral claim has been violated (1994: 36–7). According to Wallace, the central cases of reactive attitudes are the emotions of resentment, indignation, and guilt (1994: 29–30), which are all negative in character. This is a feature of our reactive attitudes, he suggests, that explains their "special connection to the negative responses of blame and moral sanction" (1994: 62, 71).

Wallace draws out several significant points from this narrow account of the reactive attitudes. The first is that reactive attitudes are not coextensive with the emotions we feel towards people with whom we have interpersonal relationships, since reactive attitudes must be identified with reference to their "constitutive connections with expectations" (1994: 31). There are many interpersonal emotions we may experience—such as attachment, friendship, sympathy, love, and so on—that cannot be counted as reactive attitudes since they do not have any relevant connection with expectations. It follows from this that we must reject Strawson's claim that a life without reactive attitudes would commit us to an (impossible) universal adoption of the "objective attitude." People may continue to entertain various other forms of interpersonal emotions and relations even in the *complete absence* of reactive attitudes narrowly conceived. Moreover, against Strawson, Wallace argues that even if some form of interpersonal relations are an inescapable feature of human life, it does not follow that the reactive attitudes are "similarly inevitable" (1994: 31). This point leads Wallace to his second distinction relating to his account of the reactive attitudes.

We also need to draw a distinction, Wallace says, between moral reactive attitudes and other kinds of moral sentiment. Not all moral sentiments take the form of moral reactive attitudes, with some identifiable tie to expectations and obligations. Among the examples of nonreactive moral sentiments Wallace cites are shame, gratitude, and admiration; all of which involve different kinds of "modalities of moral value" (1994: 37–8). This distinction allows us to acknowledge that there are other cultures with forms of ethical life that do not have any commitment to reactive attitudes but may have other kinds of moral sentiment, as we find in shame cultures (1994: 31, 37–40). Evidently, then, on Wallace's narrow interpretation of reactive attitudes, Strawson's naturalism is excessive not only at the token level, but also at the type level, since it is entirely conceivable that there are cultures where members are not subject or liable to reactive attitudes at all. For Wallace, abandoning naturalism at both the token and type levels is a price well worth paying, as it is the only way to avoid an overly inclusive account of reactive attitudes and a false dichotomy between the reactive attitudes and the objective attitude. The narrow construal, Wallace maintains, is more faithful to the relevant psychological and sociological facts and also permits us to identify more accurately the justificatory issues that arise with respect to issues of moral responsibility.

Wallace is well aware that his narrow construal of the reactive attitudes commits him to an interpretation of moral responsibility understood entirely in terms of the conceptual resources of "the morality system."[2] The morality system is understood as a particular form of ethical life, associated with our modern, Western, Christian culture. Its central normative concepts are obligation and blame, along with related notions of wrongness and voluntariness. These are all the same key elements that feature in Wallace's narrow construal of moral responsibility. The narrow account renders moral responsibility, so interpreted, as a local and contingent cultural achievement, involving a legalistic, neo-Kantian view of morality. Understanding moral responsibility in narrow terms presents its adherents with their own set of difficulties. Some of these difficulties are anticipated in Wallace's discussion.

[2] Wallace 1994: 39–40, 64–66; and Williams 1985: esp. ch. 10.

3. THE COSTS OF GOING NARROW

One of the most obvious costs of analyzing moral responsibility in terms of a narrow interpretation of our reactive attitudes is that it commits us to an "asymmetrical" account with respect to "worthy and unworthy actions" (Wallace 1994: 71). Since our moral reactive attitudes are, on the narrow interpretation, aroused only when expectations that we endorse are violated, it follows that any moral sentiments we experience that are positive responses to other "modalities of moral value" are not strictly reactive attitudes—and so no part of moral responsibility. While we may feel gratitude or admiration for a morally worthy act, the moral emotions involved are not to be accounted for in terms of the specific structure of beliefs about the violation of moral obligations (1994: 37–8). Wallace is unapologetic about this asymmetrical feature of the narrow view on the ground that it accounts for the "special connection" that exists between holding people responsible and our retributive practices involving blame and punishment (1994: 71). Beyond this, Wallace also claims that holding a person responsible for a worthy action "does not seem presumptively connected to any positive emotion in particular" (1994: 71).

Wallace's defense of the asymmetrical features of the narrow account of moral responsibility is unconvincing for several reasons. First, while we may agree that there is a close connection between moral reactive attitudes and our retributive practices, this does not imply that we need to interpret our reactive attitudes as *exclusively* negative in character, and as *always* connected with blame and moral sanctions. If we allow that there are reactive attitudes of a positive kind, based on beliefs concerning worthy actions and admirable character traits, then these too may be connected with positive retributive dispositions such as praise and rewards. Second, it is not obvious that our emotional resources with respect to our responses to morally worthy actions and traits are any more impoverished or limited than in the case of our negative reactive attitudes. As Wallace's own observations suggest, we not only have "thin" ethical concepts such as approval and praise (correlates to disapproval and blame), we also have many "thicker" concepts, including gratitude and admiration. Finally, and most importantly, any asymmetrical account of the kind advanced by the narrow view inevitably truncates and distorts our experience of moral life and the various ways in which our reactive attitudes are grounded and directed. A one-sided view that is exclusively concerned with negative reactive attitudes, focusing entirely on their connection with blame and retribution, offers us an impoverished and unbalanced interpretation of responsibility and fails to properly accommodate the constructive role of reactive attitudes in endorsing and supporting morally worthy or admirable actions and traits.

Another objection to the narrow interpretation, which Wallace also anticipates and tries to fend off, is that it presents an account of moral responsibility that has a "local" bias toward (our own) modern, Western Christianized culture—i.e. toward "the morality system." It follows from the narrow account that the ancients Greeks and shame cultures, among others, have practices that are at best "analogous" to ours, based as they are on different moral beliefs with distinct patterns of moral response (1994: 65–6) Considered in terms of the narrow interpretation, these alien forms of ethical life do not share *our* understanding of moral responsibility, not simply in the sense that they have a *different* understanding but rather that they have *no commitment* to moral responsibility (since their ethical responses cannot be understood in terms of the narrow account of reactive attitudes). This view of things renders moral reactive attitudes and moral responsibility as both local and contingent, and thereby places a conceptual barrier between ourselves (modern, Western, etc.) and alien forms of ethical life that are removed from us in historical time and geographical space. From one point of view this narrow account *oversimplifies* our own (modern) attitudes and practices, which are not perfectly or purely represented by "the morality system" and evidently involve dimensions of holding people responsible that cannot be compressed into the narrow and rigid framework provided. From another point of view, it denies us the *critical* apparatus and resources to question and challenge the way we (moderns) have (locally) arranged and structured our own attitudes and practices relating to responsibility. When we are confronted with other cultures and forms of ethical life outside "the morality system" they must, according to the narrow account, be set aside as—by definition—no longer possessing *any* conception of responsibility. Even if it is granted that we moderns are straightforwardly committed to "the morality system" and its narrow construal of responsibility, this still puts unnecessary and excessive (conceptual) distance between ourselves and these alternative forms of ethical life. More specifically, it erects a conceptual barrier to any genuine critical exchange and confrontation on the subject or question of responsibility itself—since there is, on this account, no shared or common ethical life with respect to the attitudes and practices that are actually constitutive of moral responsibility.[3]

[3] Clearly confrontation between ethical cultures that are removed from each other in historical time or geographical space may be, as Williams observes, either "notional" or "real" (Williams 1985: ch. 9). Be this as it may, historical sensitivity about the contingency of our own "local" commitments naturally puts pressure on reflective confidence in our own attitudes and practices. To this extent our *awareness* of other modes and forms of ethical life, less attached to the rigidities of "the morality system,"

Finally, there is a further objection to the narrow construal of moral responsibility, which raises difficulties that Wallace does not anticipate or directly address. Granted that moral responsibility narrowly interpreted in terms of the concepts provided by the morality system is both local and contingent, it follows that we must reject Strawson's original claims concerning our *type*-naturalist commitments to the reactive attitudes. If this is the case, then a fundamental plank of Strawson's original naturalistic strategy must be abandoned: namely, the claim that we do not need and cannot provide any external rational justification for moral responsibility. If we accept the narrow construal, then there is no natural, universal liability to reactive attitudes and they do not serve as a natural foundation for all recognizably human forms of ethical life. Since the framework of moral responsibility is erected around culturally local forms of moral emotion, confrontations with other cultures and forms of ethical life will place us in the position of needing external rational justification for the entire framework of moral responsibility so conceived. Internal justifications, provided in terms of a rationale of excuses and exemptions, will not serve this purpose even if it is successful in fending off the (internal) skeptical challenge based on worries about determinism. To this extent, the skeptical threat remains with us not just at the token level but also at the type level. Nor is it an option to retreat back to other "analogous" forms of moral emotion since, if the narrow account is correct, this will not secure or preserve responsibility properly understood. The crucial point remains that if we embrace the narrow construal then, contrary to Strawson's core original view, we are faced with the task of providing the whole edifice of moral responsibility with an external rational justification.

4. TYPE-NATURALISM AND BROAD RESPONSIBILITY

Whatever difficulties may be found for the narrow construal of moral responsibility it is important to begin with a full appreciation of Wallace's critique of the original Strawsonian strategy, much of which is justified. There are four particular features of Wallace's critique that should be endorsed as clearly justified.

(1) Token-naturalism is, as we have noted, psychologically implausible and fails, in any case, to discredit the justificatory issues advanced by

may bring us to question whether our confidence in "our *modern* concept of responsibility" is altogether well-founded.

the skeptical challenge (e.g. as based on worries about the implications of determinism).

(2) We do require a more fine-grained and less inclusive account of the reactive attitudes. In particular, it is essential that we exclude interpersonal emotions that lack any relevant cognitive element containing ethical content, as in the case of emotions such as love, sympathy, friendly feeling, and so on.

(3) It is also essential that any plausible naturalistic strategy is one that is historically or genealogically sensitive, displaying an awareness of the considerable variation in human ethical life and the range of moral emotions this may involve. In particular, we must avoid any crude form of naturalism that projects *our* (local) sentiments onto (alien) others.

(4) Finally, it is entirely correct to argue, as Wallace does, that we need a theory of moral capacity to provide the basis for an account of exemptions, in order to answer the (internal) skeptical challenge that determinism would somehow render us all morally incapacitated and thus inappropriate targets of reactive attitudes. This last issue is, however, not itself part of the revised Strawsonian analysis of *holding* people responsible and is not, therefore, a matter for our present concern.[4]

Our concerns rest with the issues arising out of the first three items on the list above. In the case of all three of these items, I will argue, Wallace's narrow view construal of responsibility involves a series of unnecessary and misleading oppositions. It is possible for us to avoid the weaknesses identified in Strawson's original strategy without collapsing into the excessively narrow view of moral responsibility that reflects the meager resources of "the morality system."

We may begin by noting that we can readily reject token-naturalism without rejecting type-naturalism. If we adopt this approach then, it is true, we will be denied the sort of easy way with skepticism that token-naturalism encourages. Moreover, as already explained, the type-naturalist approach, building on our natural liability to reactive attitudes, still leaves us needing a theory of excuses and exemptions that is consistent with compatibilist commitments, if we are to defeat the skeptical challenge. Although the skeptical effort to discredit all tokens of reactive attitudes is one that we must take seriously, if we accept type-naturalism, we do not

[4] I have argued elsewhere that there is a more intimate relationship between our capacity for holding agents responsible and our capacity for reflective self-control than (Kantian) theories such as Wallace's acknowledge. See, in particular, Russell 2004 and Russell 2011.

need any external rational justification for reactive attitudes (i.e. justification at the foundational level). Whether this further claim is acceptable or not will depend on whether we accept the narrow construal of our reactive attitudes. Clearly if we go narrow, then type-naturalism and the associated claim regarding the dispensability of external rational justifications must be dropped. The resolution of this issue depends, therefore, on our interpretation of reactive attitudes.

As I have argued, although we do need a narrower account of our reactive attitudes, we need to make sure we do not go too narrow, as otherwise we will generate some of the difficulties that have already been noted. Wallace's narrow view places considerable and appropriate emphasis on the "propositional content" involved in the beliefs that serve to delineate our reactive attitudes (1994: 11,19,74).[5] The narrow view would restrict the contents in question to the limited range provided by the conceptual resources of "the morality system." It is these limits that result in the problems of asymmetry and localism, as we have described them. We need to find, therefore, a *middle path* that avoids the inclusiveness of Strawson, on one side, and the excessively narrow approach of Wallace on the other. To put ourselves back on the right track we may turn again to Bernard Williams's critique of "the morality system."

The narrow view, as we have seen, presents ethical considerations in highly restricted terms, specifically with reference to obligation and blame, which is appropriate when obligations are violated. Williams identifies these features as central to the morality system (1985: ch. 10). This tendency to reduce and simplify is also manifest in ethical theory, a philosophical project which is itself intimately linked to the assumptions and prejudices of the morality system (1985: ch. 1). One aim of ethical theory is to provide an account of morality that will provide an exact boundary between ethical and nonethical considerations. This is done primarily by reducing the diversity of ethical (and nonethical) considerations, with a view to identifying a narrow and strict range of ethical considerations that may serve as moral reasons available to all rational agents—the universal constituency. The most notable features of moral theory are its simplicity, reductionism, and systemization of our ethical concepts and claims. Williams's critique of the morality system involves challenging and rejecting these assumptions. In the first place, while our ethical considerations certainly include obligations, under some

[5] Wallace claims that Strawson's account of reactive attitudes does not manage to clearly connect them with any propositional content (1994: 39). This charge seems unfair to Strawson since he is careful to ground reactive attitudes in our beliefs about the attitudes and intentions of other human beings and "the very great importance" that we attach to them (Strawson 1962: 5).

interpretation, they extend well beyond this. The scope of the ethical relates more broadly to "the demands, needs, claims, desires, and, generally the lives of other people, and it is helpful to preserve this conception in what we are prepared to call an ethical consideration" (1985: 12; see also 1985: 153, where Williams mentions our need to share a social world in relation to these various ethical considerations). What is required, from this perspective, is an account of ethical considerations that also includes forward-looking concerns relating to welfarism and utilitarianism, as well as ethical considerations that relate to our ideals and self-conceptions that mark out actions that fail the standards and boundaries that we may set for ourselves (e.g. considerations of what we regard as demeaning, base, dishonorable, etc.). When we interpret ethical considerations in this *broader* manner we find that these interests are plural and varied in their nature and secure no sharp boundary between ethical and nonethical considerations. Vagueness, conflict, and diversity—contrary to the demands of "theory" and the prejudices of "the morality system"—are of the essence of human ethical life.

Our own ethical reactive attitudes must be understood in these broader and vaguer terms. One of the implications that Williams draws from this is that we should be skeptical of the effort to understand reactive attitudes in the reductive, thin language of binary judgments; approval and disapproval, guilt and innocence, and so on (1985: 37, 177, 192). If we reconfigure our ethical reactive attitudes in terms of a broader construal of the ethical considerations that ground them and serve as their propositional content, then we may acquire a very different understanding of the scope and content of moral responsibility, as based upon these emotions. Our ethical qualities are manifest in the "deliberative priority" and "importance" that we give to ethical considerations as expressed in our conduct and character. So interpreted, ethical reactive attitudes may be construed as *reactive ethical value*, where this is understood as emotional responses to the weight and value given by an agent or person to ethical considerations widely conceived (i.e. in terms of our human needs, interests, welfare, claims, and the requirements of social cooperation). Ethical reactive attitudes involve coming to see a person in a certain ethical "light" based on these lower-order evaluations of their ethical qualities. Clearly we have varied and diverse ethical norms and standards that serve as the relevant basis for evaluating an agent's ethical qualities understood in these terms. These evaluations of agents based on their ethical qualities serve to generate or arouse a myriad of ethical reactive attitudes which may be either "positive" or "negative" in nature. As Williams argues, our ethical and emotional language here is not at all "thin" (e.g. praise and blame etc.) but is "thick" and varied, involving notions such as "being creepy," or a "cad," and so on—all of which are responses loaded with ethical significance.

Different cultures and different forms of ethical life will not only have different lower-order ethical norms, they also deploy a different or variable set of ethical reactive attitudes (reflecting their variable propositional content).

The significance of this criticism of "the morality system," along with the style of ethical theory associated with it, for our understanding of ethical reactive attitudes should be clear. The revised account is broad enough to accommodate positive ethical reactive attitudes (e.g. gratitude, admiration, etc.) as well as "alien" reactive attitudes (e.g. shame) all under the umbrella of those ethical considerations that serve to ground or justify them. This avoids the costs of going too narrow, by way of relying on the limited and restricted resources of the morality system. The broader construal is, nevertheless, controlled and focused enough to exclude elements that do not relate at all to ethical considerations and ethical qualities (e.g. friendly feeling, sympathy, romantic love, etc. do not count as ethical reactive attitudes because they are not reactive to *ethical* qualities as such). On this analysis, we should not be surprised or disappointed to find that there is no sharp or clear boundary between ethical reactive attitudes and other emotional responses to qualities and features of those with whom we are dealing. No such sharp boundary should be expected if we want an accurate understanding of the nature of ethical life and the way in which it "bleeds" into human life in general.[6]

Taking this broader approach to ethical reactive attitudes has other significant advantages as well. It avoids, for example, the "legalism" of the narrow account, which turns moral responsibility into a model of legal responsibility—eliminating the more nuanced and complex set of responses we have outside legal contexts. We also avoid the failings of what Williams refers to as "progressivism," the assumption that we moderns alone have access to a full and complete concept of moral responsibility and are "better off" than those who lack our own understanding.[7] It is Williams's view that not only should we be open to the possibility that we might learn from the ancients, this is in fact our situation. Learning from the ancients is possible—and desirable—precisely because we *share* a concept of moral responsibility with them, however differently we may interpret various key elements associated with it (1993: 55).

For our present purposes, our concern is not to present a worked-out alternative to the narrow model of reactive attitudes—not the least because,

[6] This is, of course, a recurrent theme throughout Williams's writings.

[7] Williams 1985: 32 n. 2; and, more generally, Williams 1993: esp. ch. 1. Williams's view is, of course, the opposite of this, since he holds that we would be better off without the morality system (1985: 174), just as we don't need ethical theory and should abandon its aims (1985: 17, 74).

for reasons given, this may itself be a problematic ambition driven by the aims of "ethical theory." What is important, however, is to insist on finding some middle ground that can accommodate ethical reactive attitudes broadly conceived without expanding this set to include interpersonal emotions that have no relevant ethical content (i.e. which do not involve our emotional reactions to a person's ethical qualities). Wallace's own observations suggest that this can readily be done, since he allows "analogous" forms of responsibility and also speaks of "responsibility for worthy acts" (Wallace 1994: 38–40, 64–6, 71). To see, in a particular case, how this middle ground between an excessively narrow and overly inclusive view may be found let us consider *shame*. Shame may be based on standards and norms that have no ethical content, as in the case of concern about one's physical appearance (e.g. my frail constitution) or economic status (e.g. my family's poverty). In other cases, however, the relevant standards and norms may move into the territory of our ethical qualities and characteristics, such as feeling shame about being lazy or being vulgar. Whether a response is an ethical reactive attitude or not will depend on the *nature* of the quality or consideration it is a *reaction to*. There are, moreover, clear cases of ethical shame that cannot be analyzed or understood in terms of the apparatus of obligation and doing wrong. We may, for example, feel ashamed of failing to live up to our own ethical ideals and standards, even when we are well aware that we have not failed to comply with any obligations and cannot be blamed for our conduct. An example of this is provided in Joseph Conrad's *Lord Jim*, where Jim is ashamed of himself because he fails to act heroically and is, therefore, disappointed in himself in these ethical terms but not in terms of any recognizable requirements of "the morality system."[8] The general point here is that there is a wide range of ethical reactive attitudes lying outside the theoretical schema of the narrow interpretation that, nevertheless, do not collapse into an overly "inclusive" set of interpersonal emotions lacking any relevant ethical content. Some cases of shame will be cases of ethical reactive attitudes and others will not. What will settle this issue will be the specific content and target of what we are ashamed of. Moreover, the fact that there are no sharp or precise boundaries to draw here is a failing only if we assume the prejudices of the morality system and the forms of "theorizing" associated with it.

We have noted that it is essential that the naturalist approach to moral responsibility should be sensitive to historical and cultural variations with regard to our understanding of moral responsibility and the specific and

[8] For an illuminating discussion of this example see Doris 2002: 160–4.

various forms which our ethical reactive attitudes may take. It should be pointed out that Strawson is himself alive to these concerns. Speaking of our increased "historical and anthropological awareness of the great variety of forms which these human attitudes may take at different times and in different cultures" Strawson says:

> This makes one rightly chary of claiming as essential features of the concept of morality in general, forms of these attitudes which may have a local and temporary prominence. No doubt to some extent my own descriptions of human attitudes have reflected local and temporary features of our own culture. But an awareness of variety of forms should not prevent us from acknowledging also that in the absence of *any* forms of these attitudes it is doubtful whether *we* should have anything that we could find intelligible as a system of human relationships, as human society. (1962: 24–5. Strawson's emphasis)

It may be argued, along the lines of Wallace's criticisms of Strawson's claims about the objective attitude, that Strawson's remarks in this passage run together two distinct issues. One claim is that we could not recognize a society as truly human without any ethical reactive attitudes. Clearly this need not be the case, so long as we do not overly expand the class of ethical reactive attitudes to include all interpersonal emotions. Nevertheless, the general point that Strawson is primarily concerned to make in this context still stands: namely, without *some form* of ethical reactive attitudes we could not recognize or find intelligible a system of human relationships that would qualify as a *human ethical life*. In other words, an ethical life devoid of all forms of ethical reactive attitudes is not recognizable or intelligible *to us* as a form of *human* ethical life.

It should be clear, in light of Strawson's observations, that we do not need to choose between naturalism and genealogy, where this is understood in terms of sensitivity to historical and cultural variation and diversity. On the broad construal, ethical norms and the ethical considerations to which they give weight, may vary greatly from one culture and historical period to another. With these variations we will also find variations in the particular forms of ethical reactive attitudes that are adopted and endorsed. One form these ethical reactive attitudes may take is the narrow form encouraged by the morality system—which makes obligation and blame its central features. While this form of ethical reactive attitude may be local and contingent, it does not follow that ethical reactive attitudes *broadly construed* are local and contingent, unless we take the local form to be the sole legitimate representative form of moral responsibility. Since the broad construal neither interprets ethical reactive attitudes nor moral responsibility in these restrictive terms, it is able to acknowledge that the ancient Greeks, among others, have ethical reactive attitudes that are recognizably continuous with our own conception of moral responsibility. This can be done without in any way denying that there are significant differences between their culture and our own with regard to the attitudes and practices involved (e.g. with respect to issues of voluntariness and intention). The irony about this situation is that it is the narrow construal, which rejects (type) naturalism, which cannot accommodate genealogical sensitivity to cultural and historical variation. Given that the narrow construal insists that moral responsibility be understood in terms of the concepts of the morality system, and that this interpretation alone constitutes genuine or real moral responsibility, it is compelled to exclude all other understandings (i.e. as based on a broader construal of ethical reactive attitudes) as falling outside the parameters of moral responsibility. It is, therefore, the narrow construal that is insensitive to genealogical variation and diversity as it arises within the framework of moral responsibility. For reasons that have been explained, this is not simply a verbal issue, as it involves and encourages a truncated and distorted understanding of moral responsibility and the way in which it is naturally rooted in human ethical life.

It has been shown that "asymmetry" and "localism" are unnecessary and unacceptable costs of the narrow approach. What, then, about the further issue relating to type-naturalism and external rational justifications? If we adopt the broad construal then no external rational justification of our liability to ethical reactive attitudes is required, where ambitions of this kind involve what Strawson describes as the tendency to "over-intellectualize the facts" concerning the natural foundations of moral responsibility (Strawson 1962: 23). In contrast with this, the narrow construal needs to provide some relevant external rational justification since, *per hypothesis*, it is a local (modern, Western) achievement. Clearly the broad view can accept that the *local forms* of ethical reactive attitudes (e.g. as based on the requirements of the morality system) may come and go. Considered from this perspective, the morality system and its associated narrow interpretation of moral responsibility may be judged vulnerable to extinction for several related reasons. First, as already argued, it suggests a truncated and distorted account of our own ethical concerns (i.e. even from a modern, Western perspective). Second, it generates (unnecessary) problems of asymmetry and localism as we have described them. Third, the narrow account is also especially vulnerable to internal skeptical collapse due to worries about determinism. (Although Wallace believes these internal skeptical objections can be defeated, not all those who endorse the morality system believe this can be done, much less done on the basis of compatibilist commitments.) It follows from all this that there is a real prospect of this *local* form of ethical reactive attitudes collapsing under both internal and external skeptical pressure. What does not follow from this, however, on the broad construal, is the total collapse of moral responsibility in any recognizable form. We may still have available other forms of ethical reactive attitudes that are not similarly vulnerable in any of these dimensions. Clearly it would be a mistake, therefore, to present the collapse of the local form of moral responsibility associated with the morality system as putting us in the predicament of having to adopt some form of utilitarian, forward-looking approach that has no place for ethical reactive attitudes *of any kind*. Alternative forms of reactive attitudes remain available and viable within the structure of ethical life that is still recognizably *human* and intelligible to us.[9]

The question we must now turn to is what is the relationship between skepticism and type-naturalism as understood on the broad approach? Skepticism may take the form of aiming to discredit *local* understandings of our ethical reactive attitudes. This does not, as has been argued, show that all forms of ethical reactive attitude are thereby discredited or that moral responsibility, as such, cannot be vindicated. It remains true, nevertheless, that whatever local forms our ethical reactive attitudes may take, they (all) remain vulnerable, in principle, to internal, global skeptical challenge at the *token* level. That is to say, our commitment at the type level, even on a broad construal, does not secure any general immunity from potential global skepticism with respect to *all* tokens of our ethical reactive attitudes.[10] The crucial point that needs to be emphasized, however, is that even in these circumstances, the skeptic remains committed to ethical reactive attitudes at the *type* level (unless, of course, the capacity to feel and experience ethical reactive attitudes is itself damaged). In other words, although the skeptic may systematically disengage from all *tokens* of ethical reactive attitude from "the inside," she cannot abandon the propensity or liability to these attitudes. That is a project that, from one point of view, would require radical intervention with her own nature (e.g. by way of genetic engineering or medical surgery)

[9] Wallace refers to the utilitarian approach as "the economy of threats" model (1994: 54–61). It is crucial to his critique of this model that it lacks "depth," where depth is provided by the "attitudinal" features of blame and retribution (1994: 56, 75). On a broader construal, however, "depth" can be found in other forms of ethical reactive attitude, such as shame and anger—a point that Wallace comes close to endorsing in some passages. See, e.g. 1994: 89.

[10] This may well be regarded as highly unlikely or even incredible—but it is not inconceivable. Imagine, for example, the spread of some terrible disease or genetic mutation that affected us all by damaging our most basic and universal moral capacities in such a manner that exemptions applied universally (however broadly interpreted).

and, from another point of view, would place her outside the recognizable human ethical community.[11]

We have already noted that whereas the narrow construal would place the ancient Greeks (and other shame cultures) outside the framework and fabric of moral responsibility, the broad construal does not. What, then, about the skeptic? In contrast with individuals who are engaged participants in forms of ethical life involving ethical reactive attitudes, the skeptic has systematically disengaged from all such participation or involvement. Disengagement of this kind requires (internal) doubts about the justification for *any* proposed tokens of ethical reactive attitude. So described, the skeptic cannot evade the challenge of providing some account of the excusing and exempting considerations that apply universally in such a manner that all *tokens* of ethical reactive attitude are discredited. If the skeptic fails or refuses to provide any such rationale for her (disengaged) stance then her skeptical stance has not been vindicated or justified. Contrast the skeptic with another distinct character, who we may call the "Vulcan."[12] Vulcans are understood to be entirely rational but incapable of human emotion. As such, Vulcans may rationally understand (human) ethical norms but are incapable of feeling or entertaining ethical reactive attitudes (or similar moral emotions with an attitudinal aspect). Vulcans have, in other words, no type-naturalist commitments with regard to ethical reactive attitudes. It is, for this reason, a mistake to assimilate the skeptic to a Vulcan, as plainly the skeptic is not a Vulcan. The Vulcan faces no skeptical problem with respect to ethical reactive attitudes. They have no token commitment because they have no type commitment to this range of (ethical) emotion. For the (human) skeptic, however, the skeptical challenge is *real* because their type propensities require something to be said with respect to disengaging all tokens of these reactions to the ethical qualities of others in their community. In this way, both the skeptic and the anti-skeptic, in contrast with the Vulcan, can accept Strawson's type-naturalism and dispense with the search for external rational justifications. What divides them is the issue of whether or not a theory of excuses and exemptions can handle relevant internal skeptical worries (e.g. as based on the implications of determinism).

[11] An individual who lacks any *type* commitment to ethical reactive attitudes would not be recognizably human, not because she is a systematic skeptic with respect to these attitudes but because the skeptical issue *does not arise for her* with respect to these attitudes, since she is constitutionally incapable of experiencing or entertaining such attitudes.

[12] Vulcans are aliens from the planet Vulcan, as described in *Star Trek*. The character of Mr Spock was half Vulcan and half human. Wallace refers to this example in a related context at Wallace 1994: 78 n.41. See also Russell 2011: esp. 212–14.

5. AGAINST THE NARROW CONSTRUAL OF MORAL RESPONSIBILITY

It has been argued that the narrow construal of reactive attitudes and its associated account of moral responsibility has unacceptable costs. While it is true that there are significant failings in Strawson's original naturalistic project that need to be addressed and corrected (e.g. we should reject token-naturalism), we should retain the core feature of type-naturalism. In order to do this we need to provide a *broader* account of ethical reactive attitudes that extends beyond the constraints and limits of "the morality system" and its conceptual structures. It is only by taking this route that the difficulties we have described relating to "asymmetry" and "localism," as well as the fruitless and misguided search for external rational justifications, can be avoided. We may summarize the significance of these observations in the following points.

(1) The narrow construal of moral responsibility, as developed on the basis of the morality system, both distorts and truncates our understanding of human ethical life as it relates to moral responsibility. In particular, it makes it impossible to accommodate both positive ethical reactive attitudes and alien ethical reactive attitudes as they may arise from outside our (modern, Western) culture. Even our own *local* understanding of moral responsibility is not fully or adequately captured by this narrow construal.

(2) It is the broad construal, along with its commitment to type-naturalism, which is able to accommodate genealogical sensitivity to historical and cultural variation in relation to our understanding of moral responsibility. The narrow construal excludes all alternative forms that do not fall into the constraints imposed by "the morality system" as mere analogues or prototypes of moral responsibility. As such, the narrow construal constitutes a form of conceptual imperialism with regard to (real, true) moral responsibility and also commits us to an implausible "progressivism" concerning our own (modern, Western) views. In contrast with this, the broad approach recognizes the variation in modes and forms of ethical reactive attitude within a wider understanding and appreciation of the emotional fabric of moral responsibility.

(3) Type-naturalism, as understood on the broad construal, provides no easy way of dealing with a potential internal skeptical challenge (i.e. in contrast with the aims of token-naturalism). Even allowing for our natural liability to ethical reactive attitudes, on a broad construal, we must still formulate some relevant schema of excuses and exemptions. From this perspective it is always conceivable that a systematic or global skepticism

could be generated from "the inside" (i.e. extending to all our token ethical reactive attitudes). This possibility does not, however, license a search for external rational justifications, since our liability or propensity to such emotions is natural and not rationally grounded. The skeptic remains committed to ethical reactive attitudes at this level, even if she has entirely abandoned or disengaged any commitment to tokens of these attitudes (in light of internal skeptical pressures of some kind).

(4) Much of the motivation behind Wallace's narrow construal of the reactive attitudes is to find a satisfactory compatibilist account of moral responsibility consistent with the core requirements and constraints of the morality system. From this perspective the internal skeptical challenge is especially acute, since it is targeted on the notions of wrongness, blame, desert, and retribution that are central to moral responsibility as the morality system interprets it. It is evident, however, that the broad construal of ethical reactive attitudes, along the lines that has been sketched, significantly *deflates* these (internal) skeptical pressures. The reason for this is that a broader and more liberal conception of ethical reactive attitudes does not place such heavy weight or emphasis on the very elements of the morality system that have proved especially vulnerable to skeptical criticism (i.e. desert, blame, etc., along with their apparent dependence on ultimate or absolute agency). Even if—contrary to what Wallace argues—it proves impossible to vindicate this local interpretation of moral responsibility, as understood on the narrow construal, it does not follow, given a broad interpretation of ethical reactive attitudes, that global skepticism results. All that follows from the success of the skeptical challenge, so described, is that the *local* understanding of moral responsibility encouraged by the morality system cannot survive critical reflection.[13]

[13] It is true, of course, that many skeptics about moral responsibility are concerned to discredit the local conceptions of moral responsibility associated with the morality system. However, for reasons that have been discussed, skepticism of this kind does not in itself constitute global skepticism—since it does not discredit, and may not even aim to discredit, alternative forms of ethical reactive attitudes. Having said this, it is important to note that many skeptical projects of this kind either explicitly or tacitly endorse the narrow construal and its assumption that alternative accounts of ethical reactive attitudes somehow fail the standard of *real or genuine* forms of moral responsibility. When this assumption is made, the critique of our local conception of moral responsibility framed in terms of the requirements of the morality system is (mistakenly) *inflated* into a form of *global* skepticism about moral responsibility. Suffice it to say that much of the contemporary free-will debate, along with its associated worries about the skeptical threat, proceeds on this assumption of the narrow construal and the morality system.

In sum, we may contrast the relative strengths and weaknesses of the broad and narrow accounts in these terms. The narrow construal not only generates a partial and incomplete account of moral responsibility, it also leaves the entire edifice of moral responsibility, so understood, vulnerable to both internal and external skeptical threat. The broad construal not only avoids the significant difficulties that the narrow construal encounters (e.g. asymmetry), it provides for the complexity, variation and nuance that we find in this sphere. Moreover, the broad construal, by moving away from the rigidities and (peculiar) demands of the morality system, deflates the internal skeptical threat and eliminates all worries relating to the misguided ambition of providing a satisfactory external rational justification. These are fundamental points relating to moral responsibility and the defects of the morality system that the discussions of both Strawson and Williams converge on.

REFERENCES

Doris, John (2002). *Lack of Character: Personality and Moral Behavior.* (Cambridge & New York: Cambridge University Press).

Russell, Paul (1992). “Strawson’s Way of Naturalizing Responsibility.” *Ethics* 102. 2.

—— (2004). “Responsibility and the Condition of Moral Sense.” *Philosophical Topics*, 32, 1; 2, 287–305.

—— (2011). “Moral Sense and the Foundations of Responsibility.” In *Free Will*, 2nd edn., ed. Robert Kane (New York: Oxford University Press).

Strawson, P. F. (1962). “Freedom and Resentment,” reprinted in P. F. Strawson, *Freedom and Resentment and other essays.* (London, New York and Oxford: Methuen. 1974).

Wallace, R. Jay (1994). *Responsibility and the Moral Sentiments.* (Cambridge, MA: Harvard University Press).

Williams, Bernard (1985). *Ethics and the Limits of Philosophy.* (London: Fontana).

—— (1993). *Shame and Necessity.* (Berkeley, CA: University of California Press.).

8

The Three-Fold Significance of the Blaming Emotions

Zac Cogley

1. INTRODUCTION

Many philosophers working on moral responsibility follow P. F. Strawson (1982) in understanding claims about someone's moral responsibility or the phenomenon of holding people morally responsible in terms of the appropriateness of a certain class of emotions (Bennett 1980; Watson 1993; Wallace 1994; Fischer and Ravizza 1998; McKenna 1998; Macnamara 2009). But even those who would not follow Strawson in identifying moral responsibility attributions with the appropriateness of emotions hold that emotions do play a role in our moral responsibility practices (Scanlon 2008, 143). In spite of this, the significance of the blaming emotions for moral responsibility has been under-theorized. (I am concerned here with people's moral responsibility for their actions or omissions, rather than, for example, whether in general someone is a morally responsible person.)

In order to fully appreciate the import of the blaming emotions for moral responsibility we need a more adequate moral psychology. As an initial step, in this paper I appeal to recent work in psychology of emotion to argue that the blaming emotions—anger, resentment, and indignation—are significant for our moral responsibility practices in three different ways.[1] They are important to moral responsibility in *appraising* people as acting wrongfully, in *communicating* the appraisal to perceived wrongdoers, and in *sanctioning* people who are appraised as wrongful.[2] I also investigate the conditions of appropriateness of the blaming emotions. My methodology is inspired by

[1] While there are positive emotions that are connected to moral responsibility, I focus on the blaming emotions as they have received much more philosophical and psychological discussion than have candidate positive emotions like gratitude.

[2] An anonymous reviewer points out that these also correspond to three broad categories of response to wrongdoing. I agree—in fact, I think we categorize responses

recent philosophical attention to reasons for attitudes: for example, the reasons in favor of believing a proposition (Shah 2003) or blaming another person (Hieronymi 2004). There has also been some attention—though not nearly as much—to the reasons that bear on emotions (D'Arms and Jacobson 2000). As I will demonstrate, the three ways in which the blaming emotions are significant for our moral responsibility practices are associated with very different kinds of appropriateness considerations.

My work is also inspired by the fact that although there has been significant recent attention to the concept of moral responsibility, there is little agreement about it. Indeed, in one recent attempt to clear the conceptual territory, John Martin Fischer and Neal Tognazzini argue that there are up to *thirteen* different analytical or conceptual "stages" of moral responsibility attributions, organized (roughly) into two broad categories: attributibility and accountability (2010). Here they are inspired by Gary Watson's (1996) distinction between these two concepts, but urge that conceptual clarity about moral responsibility requires far more distinctions.[3]

I am deeply sympathetic to the project of achieving clarity about our conception of moral responsibility as it is central to making progress on some of our most vexing issues about moral responsibility, including whether moral responsibility is compatible with determinism. However, I fear that some recent attempts to introduce clarity risk further confusion because they have not paid sufficient attention to the moral psychology of the blaming emotions. Not only, then, do I try to enrich our moral psychological picture of the blaming emotions, but I also link appraisal, communication, and sanction to representative accounts of moral responsibility. I suggest that each kind of account is inspired by a different way in which the blaming emotions are significant, and thus each account implicitly emphasizes a different consideration of emotional appropriateness. Fittingness accounts of moral responsibility are linked to appraisal, moral address accounts correspond to the communicative dimension of the blaming emotions, and desert accounts of moral responsibility are inspired by the blaming emotions' sanctioning role. If I am right, part of the reason debates about moral responsibility have been so intractable is that many theorists share the assumption that appropriate blaming emotions are a reliable indicator of a person's moral responsibility, while inappropriate blaming emotions are evidence of a lack of moral responsibility. This makes it appear as if all parties to the debate are operating with the same

to wrongdoing as appraisals, communications, and sanctions in virtue of their connection with the blaming emotions. Space precludes making that argument here.

[3] Fischer and Tognazzini's analysis places consideration of the blaming emotions squarely into the accountability category. My analysis here complicates that categorization.

conception of moral responsibility in mind. However, because different accounts are implicitly linked to different kinds of appropriateness, the wide agreement that the appropriateness of the blaming emotions is revealing of moral responsibility obscures significant disagreements about the concept and the conditions for its application that emerge with a more refined focus.[4]

While discussion of all of the blaming emotions is common, theorists often emphasize one or two to the exclusion of others. For example, R. J. Wallace speaks of indignation and resentment (Wallace 1994) as does Tamler Sommers (2007), while Derk Pereboom has remarked that "of all the attitudes associated with moral responsibility, it is anger that seems most closely connected with it" (Pereboom 2001, 208).[5] In what follows, I assume that from a psychological standpoint, resentment and indignation are ways of being angry.

2. APPRAISAL

An important strand of contemporary psychological research on emotion seeks to determine characteristic appraisals that are assumed to elicit distinctive emotions.[6] "Appraisal" refers to a person's evaluation or interpretation of a situation. According to this research, different emotions are caused by distinct appraisals. For example, Richard Lazarus claims that anger is produced by a person's appraisal of a "personal slight or demeaning offense" (1991, 223), while in a later collaboration with Craig Smith (1993), both believe that anger is caused by an appraisal of "other-blame," which they claim can be broken into three separate components: motivational relevance (the situation is personally relevant), motivational incongruence (the situation is inconsistent with what is desired), and other-accountability (the emotion is directed at someone else). Philosophers have roughly concurred.

[4] In his recent paper, "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility" David Shoemaker offers a similar argument that distinguishes among moral responsibility concepts, though without a focus on the blaming emotions (2011).

[5] Most commonly, it appears that these different terms mark a difference in whether the object of the emotion is second or third-personal. For example, see (P. F. Strawson 1982; Wallace 1994; Sommers 2007; Pereboom 2009).

[6] While there has been much debate over whether or not the relevant appraisals are cognitive, beginning with (Zajonc 1980; Lazarus 1982) and continued in (Zajonc 1984; Lazarus 1984), that debate is orthogonal to my concerns. For an excellent recent discussion of this issue, see (Prinz 2004, 21–51).

For example, Jesse Prinz and Shaun Nichols claim that "Anger arises when people violate *autonomy* norms, which are norms prohibiting harms against persons" (2010, 122). If we make the plausible assumption that slights and offenses both involve the violation of norms, we can see all these authors offering a roughly similar account of the appraisal involved with the blaming emotions, though they do disagree about how best to capture it.

While I agree that the blaming emotions have a characteristic appraisal, the above accounts make two errors regarding it. First, these accounts fail to pinpoint the characteristic appraisal of the blaming emotions. The early Lazarus, as well as Prinz and Nichols, construe the appraisal too narrowly. For example, blaming emotions are commonly elicited by harms against nonhuman animals, violations of religious commandments, the nonharmful breaking of promises and many other situations that go beyond slights and harms against persons.[7] On the other hand, the account from Lazarus and Smith is too broad; adding up their three appraisal components (an emotion directed toward a personally relevant situation that is inconsistent with what is desired) does not give us the characteristic appraisal of the *blaming emotions*. Such an appraisal is also compatible with sadness. We do better if we follow James Averill, who argues that "the typical instigation to anger is a value judgment. More than anything else, anger is an attribution of blame" (1983, 1150) or Shaver et al., who hold that the eliciting appraisal is that "the situation is illegitimate, wrong, unfair, contrary to what ought to be" (1987, 1078).

I propose, then, that the way a person feeling a blaming emotion appraises her situation is best captured as an appraisal of wrongful conduct. This is the core appraisal of the blaming emotions, but we can break it into constituent parts as follows:

If a person, A, feels a blaming emotion, she evaluates her situation as containing:

(i) a person, B,[8] whose
(ii) action or omission
(iii) transgresses a norm on proper conduct (including, but not limited to, moral norms, though the norm need not be codifiable by a rule)
(iv) because B is motivated by ill will or has shown insufficient concern,
(v) and A glosses B's action as bad.[9]

[7] Surprisingly, Prinz and Nichols themselves note the connection between blaming emotions and harms against nonhuman animals (2010, 130).

[8] In some situations A and B will be the same person.

[9] I have in mind here the fact that anger has a distinctive unpleasant phenomenology that might be glossed as "feeling ready to explode" (Roseman, Wiest, and Swartz 1994).

This treatment of the appraisal dimension of the blaming emotions handles the fact that we often feel the blaming emotions in response to violations that don't harm persons, as well as the variety of situations where we feel blaming emotions because a person's action violates an autonomy norm, or is a demeaning offense or personal slight.

This is an improvement, but there is another error in the above treatments of the appraisal involved in the blaming emotions. All of the above treatments construe the relation between the appraisal and the blaming emotions as a causal relation. That is, the appraisal is what *brings about* the blaming emotion. This is the second mistake in the literature about the relation between the blaming emotions and their characteristic appraisal. Not all psychologists believe that appraisals always precede blaming emotions or are necessary for them; indeed there is not clear evidence that appraisals always cause episodes of the blaming emotions, though there is no question they often do (Berkowitz and Harmon-Jones 2004a; Berkowitz and Harmon-Jones 2004b; Parkinson 1999).

I think we better understand the psychology of a person feeling a blaming emotion if we hold that the blaming emotions need not be *caused* by their characteristic appraisal (though they often are). However they are caused, the blaming emotions *are* an appraisal of conduct as wrongful. Consider, by analogy, a particular belief: my belief that it is sunny outside. While my belief that it is sunny outside might be caused by present sun outdoors (if I was just outside and noticed the weather), that belief might be caused in a number of other ways. I might come to believe it is sunny outside based on your testimony or by inferring today's weather based on what the weather was yesterday. In these cases my belief that it is sunny outside isn't caused by occurrent sun. A similar point applies to the blaming emotions. While in many cases they are caused by their characteristic appraisal, their link to appraisal is better understood as conceptual (Parkinson 1997).

This analogy between beliefs and the blaming emotions is also relevant because it relates to our practice of taking our blaming emotions to be appropriate or inappropriate, depending on their aptness for the situation. The blaming emotions are not unique in this respect. As Justin D'Arms and Dan Jacobson have pointed out, we commonly argue about whether or not things are sad, enviable, shameful, or worthy of pride or resentment. Our practice of considering these issues of emotional appropriateness presupposes that we can make sense of whether or not an emotion's characteristic appraisal is accurate, or to use their terminology, *fitting* (2000). When a blaming emotion is fitting, it accurately presents its object as having the features contained by its appraisal; the fittingness of a blaming emotion is analogous to the epistemic relation that obtains between the world and a true belief. Anger, resentment, and indignation are fitting to feel when, for example, someone intentionally wrongs you out of ill will.

Thus, the blaming emotions are fitting when they are felt in response to a person who satisfies conditions (i)-(v), above. The lack of any one of the five conditions means that a blaming emotion is unfitting.[10] We can also distinguish between "degrees" of fit between a blaming emotion and the situation it appraises. I should be angrier with someone who tries to ruin my career than a neighbor who thoughtlessly mows his lawn at 8 a.m. on a Sunday morning. And you should be more upset with the driver who intentionally tries to run you over while you are out for a walk than you should be with a person who somewhat carelessly backs his car into your path. Thus, the seriousness of the wrong in question and the person's relation to the wrong both help to determine the amount of anger fitting for the situation.

3. COMMUNICATION

There is much psychological evidence to suggest that the blaming emotions not only appraise the conduct of others, but also play a role in communication. Specific speech patterns (including rate of articulation, intensity, and frequency of vocal fold vibrations) appear to be associated with different emotions, particularly anger (Scherer 1986; Scherer et al. 1991). Psychologists have also found that different bodily movements and postures are associated with different emotions (Wallbott 1998). Perhaps most probatively, the blaming emotions, just like many other emotions, are associated with characteristic facial expressions (Ekman 1999). Relevantly, while people commonly interpret the emotional facial expressions of others as signifying a person's appraisal of her situation, anger expressions are more likely than the expression of other emotions to be interpreted as conveying intentions or requests (Horstmann 2003). Also notably, the characteristic facial expressions of different emotions appear to be highly associated with interpersonal interaction. For example, winners on the medal stand at the Olympic games are more likely to smile during interactions with other people than during the rest of the ceremony

[10] By distinguishing between the five conditions on the fittingness of a blaming emotion, I call our attention to conceptual distinctions. However, I allow that these different aspects of a blaming emotion's appraisal may often, or even always, affect each other in interesting ways. For example, it may be that someone's act motivated by ill will—even if she does something that I know has no chance of actually harming anyone—itself transgresses a norm on proper conduct and is therefore bad.

(Fernández-Dols and Ruiz-Belda 1995) and bowlers are less likely to smile when they first roll a strike than when they turn to face others watching at the end of the alley (Kraut and Johnston 1979).

In a typical interpersonal episode of a blaming emotion, various bodily, vocal, and facial responses communicate that the person feeling the emotion is angry and thereby give the person who is the target of the blaming emotion information about the way her conduct is being appraised. In many situations, this information is not contained in spoken words but is transmitted instead by the overall emotional demeanor of the person feeling the blaming emotion. These communicative aspects of a person's emotional demeanor are observed, responded to, or ignored by others. The responses of others—or the lack of a response—are then an opportunity for continued emotional engagement and transformation. Thus, in most interpersonal interactions, a person who feels a blaming emotion not only appraises the conduct of another as wrongful, she also communicates to the target of the blaming emotion that she construes his behavior as wrongful.[11]

In human psychological response, these communicative aspects are very closely connected to the having of the blaming emotion itself. It turns out to be almost impossible not to register your emotional state on your face in some way, even if briefly, and it is hardest to mask negative emotions (Porter and Brinke 2008). Since others are extremely attentive to such displays, it is often better to try not to have the emotion than to allow the emotion to run its course while attempting to mask what you feel. Thus, I want to suggest that the communicative significance of the blaming emotions can sometimes result in their being inappropriate to feel even when they fittingly appraise a target. That is, there are additional considerations that bear on whether to have a blaming emotion than merely considerations of fit. In characterizing communicative considerations that bear on emotional appropriateness, I am inclined to follow Angela Smith in distinguishing between considerations of standing, the degree of fault displayed in the wrongful action, and the response that the blamed person takes (or will, or could take) to the person who feels the blaming emotion.[12]

Whether or not someone has proper standing, or authority, to feel what would be a fitting blaming emotion toward a wrongdoer has much to do

[11] To be clear, I am not arguing that the communication of this information is intentional on the part of the person feeling the blaming emotion, only that the information is "there for the taking." Along similar lines, the signals that indicate which play is being called by a team may transmit the same information to the opposing team if the opposing team has attended to which signals are reliable signs of particular plays.

[12] My discussion of these three issues is indebted to (A. M. Smith 2007, 478–83).

with her relationship to the wrongdoer and those who are wronged. So, for example, even though I may observe what I regard as condescending and obnoxious behavior from one of two parties having a public argument, I may be inappropriately angry with the offending party if I have no social connection to either one of them. We can see this more clearly when we consider that it would not be uncommon for the victim of the obnoxious behavior to become angry with me when my indignation on his behalf becomes known.

My own fault and hypocrisy can also affect my standing to react to another's conduct with a blaming emotion. Thus, if my awareness of its health effects has not moderated my long-time smoking habit, my friend with a drinking problem will regard as out-of-line my fitting anger at him for neglecting his health when he succumbs to the temptation to drink. My friend's failing, just as my own, may be a legitimately moral one, but the fact that I am unable or unwilling to similarly guide my own behavior makes it inappropriate to feel angry toward him even though anger is fitting for what he does.

Beside the fact that the communicative appropriateness of the blaming emotions can be undermined by my standing to have such responses, it can also be affected by the relative significance of the fault a person displays in her conduct. Thus, while a blaming emotion would fittingly appraise a student who fails to keep his scheduled appointment with me, the degree of fault (simple forgetfulness) and the degree of harm to me (a very mild inconvenience) prohibit me from feeling any resentment toward him. We can see this interact with another communicatively salient factor—the agent's own response—if we suppose that the student rushes over to my office, apologizing profusely even as he walks in the door. The student's self-reproach indicates that he understands he did wrong and is committed to doing what he can to prevent it from occurring in the future. While being indignant would fittingly appraise his faulty conduct, I should not feel indignant toward him because it would be communicatively inappropriate for me to target him with a blaming emotion given his indication that he understands his error and the importance of keeping appointments.[13] (Of course, things might be different if this same student has routinely missed appointments even while protesting that it will never happen again.)

In such a case, we again have a communicative reason against feeling a blaming emotion, even though the blaming emotions fittingly appraise the actions of the person in the situation.

[13] For an analysis of the communicative dimension of emotions supporting this claim, see (Macnamara forthcoming).

4. SANCTION

Another way in which the blaming emotions are significant for our moral responsibility practices relates to communication but is ultimately distinct, namely, affecting the behavior of others by imposing costs.[14] While this is sometimes accomplished via the communication of a message, at other times it is accomplished simply through changing the costs and benefits of another person's possible actions. Though the relationship between the blaming emotions, deliberation, and action-aimed-at-sanction is complex, in this section I will highlight some of the relevant psychological findings to demonstrate the connection of the blaming emotions to sanction.

Anger has important effects on deliberation and social perception that play a role in determining the behavior of someone feeling a blaming emotion (though the effects vary across individuals, depending on their level of awareness and cognitive skills). Angry people tend to have a sense of significant control (Lerner and Keltner 2000) that leads them to be optimistic about the success of their probable actions (Lerner and Keltner 2001). They are also "eager to make decisions and are unlikely to stop and ponder or carefully analyze" (Lerner and Tiedens 2006, 132), which likely leads them to take actions that have a low probability of succeeding but high payoffs (Leith and Baumeister 1996). When they act, angry people tend to be more punitive toward those they blame (Lerner, Goldberg, and Tetlock 1998).[15]

We can see these deliberative effects demonstrated in experimental work on altruistic punishment and ultimatum games. In altruistic punishment, the punisher receives no material benefit but imposes a cost on the party punished. Thus, people are willing to punish free-riders even when it is costly for them to do so and they cannot expect future benefits from punishing (Fehr and Gächter 2000). Jonathan Haidt has argued that a paradigm feature of human morality is this third-party enforcement of moral norms (2001). In one significant study, Ernst Fehr and Simon Gächter (2002) demonstrated that free-riding on a common good is less prevalent when altruistic punishment of free-riders is possible. When such punishment occurs, it is reported by punishers to express their anger and those who are punished perceive their punishers as angry.[16] A similar

[14] I am not arguing that such costs are always intended by the person feeling the blaming emotion, though they certainly sometimes are.

[15] For an excellent overview of recent empirical study of anger's effects on judgment and decision-making, see (Litvak et al. 2010).

[16] Notably, angry sanctions lead to positive behavior change even when free-riders interact with a new group of people that does not include their previous punishers.

finding concerns experimental work on ultimatum games. In such two-person games, one party (suppose it's me) controls some resources (say, $10) and makes an offer to another to split the resources in a particular fashion with another party ($8 for me, $2 for you). You have the opportunity to accept or reject the offer. Although game theory would predict that the splits offered should heavily favor the person controlling the resources and that all offers should be accepted, people tend to offer more than 40 percent of the resources and 15 to 20 percent of offers are rejected (Ochs and Roth 1989). The most perspicuous explanation of this behavior is that people expect the splits to be fair; if they are not, the split is angrily rejected even though the rejection leaves the rejector worse off than had she accepted (Pillutla and Murnighan 1996). Because 'offerers' anticipate the possibility of sanctioning reactions, they tend to offer more equitable splits.

I've been using *sanction* to specifically demarcate the cost-imposing function of the blaming emotions from the communicative aspect. As I argued above, one function of the blaming emotions is to communicate appraisals to others. While one way to bring about a change in another person's behavior is to successfully communicate an appraisal to them, the communication will ultimately only be successful if the other is willing to appraise herself as acting wrongfully and see her wrongness as a reason for change, apology, or restitution. However, even if the other person is unwilling or unable to see herself as acting wrongfully, placing a cost on particular ways she might behave can impact her chosen course of action. The threat of sanction can lead someone to refrain from a wrongful action not yet performed, not repeat a wrong he already performed, or not copy the successful wrongdoing of others. Importantly, I also assume that expressions of the blaming emotions, themselves, are experienced as sanctions by the emotion's target. Not only is it unpleasant in its own right to be the target of another person's blaming emotion, but Baumeister et al. suggest that one function of the blaming emotions may be to stimulate guilt in the person who is the target of the emotion (2007, 189). Thus, people tend to avoid actions that they know would lead to being the target of the blaming emotions of others in order to avoid psychological and physical sanctions.[17]

There are a number of considerations that bear on the appropriateness of the blaming emotions qua sanction, some of which we have touched on

[17] Again, note that while the sanctioning effects of the blaming emotions are at least sometimes directly intended by people who feel a blaming emotion (Fehr and Gächter 2000), they need not always be. However, even if people feeling blaming emotions do not intend that their emotions be experienced as harms by the targets of the blaming emotion, that doesn't mean they are not experienced as such by the targets.

already. Again, the lesson is that the question of whether a blaming emotion is appropriate, all things considered, is not solely determined by whether or not the blaming emotion fittingly appraises someone's conduct. For example, the propriety of sanctioning someone with a blaming emotion can be affected by the seriousness of the wrong done. Just as with communicative appropriateness, some extremely minor wrongs ought not be responded to, while the propriety of sanctioning a wrongdoer may increase with the seriousness of the wrong.[18] The person's own repentance, or lack thereof, is relevant because sanctioning an already repentant person may be unnecessary to affect his future conduct.

As each of us has only limited motivational and actional resources, the appropriateness of a blaming emotion qua sanction can also be affected by what other wrongs you might plausibly respond to. For example, you do better to get angry with the perpetrators of wrongs when you might successfully undo the wrong or positively influence the perpetrator. If, hypothetically, you became aware of two different wrongs committed by two different people that were approximately as severe, your blaming emotions would be better directed toward a wrongdoer who would be more swayed by the communicative and motivational significance of your blaming emotions. Thus, there is something to the phenomenon where people are more likely to feel a blaming emotion in response to a wrong that directly affects them or someone they know well, rather than a wrong that affects persons with whom they have little contact. Other things equal, your motivational resources are more likely to lead to beneficial outcomes if you address concerns with which you can profitably engage.[19]

A related, but distinct, issue concerns what I will term the fairness of sanctioning a wrongdoer with a blaming emotion. We can see this notion displayed first by returning to the phenomenon of hypocrisy, earlier raised in reference to the communicative appropriateness of a blaming emotion. If you habitually commit a certain type of wrong, your right to sanction others with the blaming emotions for similar wrongs will be called into question—especially if you protest the sanction of the blaming emotions when it is applied to you. Similarly, if two people jointly undertake to commit a wrong (say, robbing someone's home) but one of them plays

[18] I say "may" increase because I am doubtful that the relative deservingness of a wrongdoer, itself, is a sufficient reason to license a sanction for her. Explaining why would require that I develop a theory of desert and deservingness, which is a topic for elsewhere. For reasons why I am skeptical that desert can be a sufficient justification for sanctions, see (Dolinko 1991a; Dolinko 1991b).

[19] Technological advances that give you knowledge of wrongs done to little-known people far away don't contradict this point, though they do complicate it considerably.

more of a role in planning and executing the deed than the other, it is appropriate to sanction the “mastermind” to a greater degree with a blaming emotion. Thus, if you must choose where to direct your blaming emotions, fairness considerations speak to you blaming the mastermind more than the accomplice. I believe the notion of the unfairness of blaming emotions qua sanction also accounts for our appropriate reluctance to blame the victims of wrongs or injustices, even if the victims of such wrongs are complicit in, or partially responsible for, the wrongdoing. In such a situation, targeting the person who was wronged with a blaming emotion amounts to piling another bad thing on top of whatever misfortune the person has already suffered.

5. LINKING ACCOUNTS OF MORAL RESPONSIBILITY TO THE THREE FUNCTIONS

To this point, I have argued that the blaming emotions are significant for our moral responsibility practices as appraisals, communications, and sanctions. I have also argued that the appropriateness of a blaming emotion in one sense does not guarantee that the blaming emotion is appropriate in others. I have particularly focused on situations where a blaming emotion fittingly appraises another’s conduct but is at the same time communicatively inappropriate or is inappropriate as a sanction. I do not take myself to have exhausted all possible appropriateness considerations that bear on these three functions, though I do hope to have captured many of the most interesting and relevant appropriateness considerations for our practices of moral responsibility. I now want to discuss several prominent accounts of moral responsibility to suggest that differing accounts of moral responsibility are motivated by attention to different ways in which the blaming emotions are significant for moral responsibility.[20]

The notion that the blaming emotions involve *appraisals* of conduct as wrongful is implicit in a number of theories of moral responsibility. For example, John Martin Fischer and Mark Ravizza hold that someone is morally responsible for her conduct to the extent that she is an “appropriate candidate for at least some of the reactive attitudes on the basis of that behavior” (1998, 6). Fischer and Ravizza admit that “in some contexts it

[20] While this project seeks to locate the source of agreements and disputes about the concept of moral responsibility, my own view is that considerations of fit are the only appropriateness conditions of the blaming emotions that bear on a person’s moral responsibility because these considerations circumscribe the concept of *blameworthiness.* Unfortunately, I don’t have space to make that case here.

may not be justified or appropriate, all things considered, actually to have any reactive attitude to a particular agent" who is nonetheless morally responsible for her conduct (1998, 7).[21] Thus, their theory of moral responsibility needs a sense of appropriateness for the blaming emotions that can be apt, even if feeling a blaming emotion is not, all things considered, appropriate. I believe that the fittingness of a blaming emotion's appraisal is the notion they are searching for.

Fittingness also allows us to make sense of R. J. Wallace's idea that resentment, indignation, and anger share a distinct propositional object, namely, that "an expectation to which one holds a person has been breached" (1994, 12). The blaming emotions, Wallace believes, are fundamental to understanding the nature of moral responsibility. Coleen Macnamara has also recently noted the connection between the blaming emotions, appraisal and moral responsibility. When a person resents her brother for not helping her move as he had promised, Macnamara notes, her resentment is "a particularly deep form of moral appraisal" that responds to the meaning of the brother's insensitive action (2009, 89). This sense of appraisal is, on her view, one face of holding others morally responsible. Angela Smith has also recently urged that the fundamental question of responsibility is whether an action can be attributed to a person "in a way that makes moral appraisal, in principle, appropriate" (2007, 470) and that negative moral appraisal, in terms of a judgment of culpability, is entailed by feeling a blaming emotion like anger, resentment, or indignation toward someone on the basis of her conduct (2007, 467).

These views about the nature of moral responsibility are all unified in taking the fittingness of the blaming emotions to be connected to ascriptions of moral responsibility. Some, like Smith, hold that the question of whether a person is morally responsible for her conduct *just is* the question of whether an appraisal like that of the blaming emotions is fitting for someone on the basis of her conduct.[22] Others, like Macnamara, hold that this is one significant aspect of our practice of holding others responsible, but that it is not the only one. Macnamara argues that another important "face" of moral responsibility is found in communicative acts (2009, 90).

Several theorists beside Macnamara have developed accounts of moral responsibility that exploit the communicative function of the blaming emotions. Thus, Gary Watson urges "the negative reactive attitudes express

[21] On their terminology, the blaming emotions are clearly reactive attitudes.

[22] Thus, as I read Smith's account, the question of someone's moral responsibility is a question of the fittingness of certain appraisals, but is not a question of the fittingness of emotions.

a *moral* demand, a demand for reasonable regard . . . the reactive attitudes are incipiently forms of communication" (1993, 264). Taking up and extending Watson's idea, Michael McKenna claims that

> holding another morally responsible for doing morally wrong is a manner of communicating with her. In particular, it is a manner of responding to what she had done as on analogy with a conversation in which the blameworthy person's conduct has a significance . . . for her to *be* blameworthy is for her to be a fitting target of this manner of response, for her to be one with whom the relevant sort of communication is called for. (2004, 187–8)

Stephen Darwall concurs that the blaming emotions are communicative. He appeals to the idea that when I feel a blaming emotion in response to you stepping on—and then continuing to stand on—my foot, my feeling the blaming emotion is in part a demand that you remove your foot from mine (Darwall 2006, 17). Even theorists like Angela Smith, who do not regard the blaming emotions as fundamental to ascriptions of moral responsibility, hold that the appraisal of another's conduct as wrongful has a communicative dimension. She writes, "Moral criticism, by its very nature, seems to address a *demand* to its target. It calls upon the agent to explain or justify her rational activity in some area, and to acknowledge fault if such a justification cannot be provided" (2006, 381).

Finally, some other accounts of moral responsibility bring the sanctioning function to the fore. So, for example, Galen Strawson holds that in asking whether people are morally responsible, we are asking if they are

> responsible for their actions in such a way that they are, without any sort of qualification, morally deserving of praise or blame or punishment or reward for them. (2002, 441)

Note how Strawson thinks the question of moral responsibility *just is* the question of whether a person is deserving of sanctions like punishment, and that he must have the sanctioning function of blame (as an attitude) in mind if he is to think that whether someone deserves blame or punishment amounts to the same question.

Echoing Strawson's claim that the question of moral responsibility is a question of deservingness "without qualification," Derk Pereboom claims that

> For an agent to be morally responsible for an action is for it to belong to her in such a way that she would deserve blame if she understood that it was morally wrong, and she would deserve credit or perhaps praise if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent, to be morally responsible, would deserve the blame or credit just because she has performed the

action (given that she understands its moral status), and not by virtue of consequentialist considerations. (2007, 86).[23]

Also placing the sanctioning function at the fore of their analyses are Tamler Sommers and Neil Levy who, like Galen Strawson, are skeptics about desert-entailing moral responsibility.[24] Sommers claims that "we feel resentment when we feel that people have wronged us, and that they deserve blame (and perhaps punishment) for what they did" (2007, 327). Levy agrees that moral responsibility has a link to desert, but that the connection between moral responsibility and deserved sanction is less direct than Strawson and Sommers believe. On Levy's view, the thought that someone is morally responsible for wrongful acts she performs amounts to the claim that such a person no longer deserves the full protection of a right they would otherwise be entitled to: "a right against having their interests discounted in consequentialist calculations" (2011, 3). To have one's interests discounted in such calculations is to incur a cost on acting wrongfully: a sanction.

6. CONCLUSION

In this paper, I've argued that the blaming emotions relate to our practices of holding people morally responsible in three different ways: appraisal, communication, and sanction. I've shown that these ways in which the blaming emotions are significant for our moral responsibility practices are themselves associated with distinct considerations of appropriateness (fittingness, communicative appropriateness, and appropriateness qua sanction) and that these different considerations can come apart from one another. A blaming emotion can be fitting, for example, but inappropriate qua communicative considerations or fitting yet inappropriate as a sanction. Finally, I've suggested that the different functions of the blaming emotions and their characteristic conditions of appropriateness are quite

[23] Pereboom has been consistently advocating this account of the conditions of appropriateness for moral responsibility since the publication of *Living Without Free Will* (2001). I interpret Pereboom as being concerned with a sanction-based understanding of blame due to his emphasis on the potentially harmful effects of anger (Pereboom 2011).

[24] This "accountability" face of moral responsibility has also been emphasized by Gary Watson (1996). Watson also characterizes another, aretaic, face of responsibility that involves beliefs or judgments about where someone's conduct falls against some standard. His discussion bears some similarity to my account of the appraisal function of the blaming emotions. However, I am uncertain whether our analyses perfectly line up.

naturally seen as inspiring corresponding accounts of moral responsibility, itself.

I think this goes some distance toward accounting for the fact that there is so little agreement about the nature of moral responsibility, even after so much attention to it. In my view, all the accounts I've discussed get something right about the moral psychology of moral responsibility and its associated conditions of appropriateness, but they also ignore other important features of our moral responsibility practices. My analysis suggests moral responsibility may be best captured as a prototype concept: when we hold people morally responsible in normal interpersonal interactions, our responses typically conform to archetypal emotional reactions that involve all three aspects of the blaming emotions. If we aim to have a psychologically realistic picture of our moral responsibility practices, I believe we must have a tripartite theory. Thus, in response to the voluminous literature defending fittingness, communication, or sanction as *the* correct account of moral responsibility, my analysis suggests that such attempts require additional argument. There is little to be gained for supporting one account over another as the correct account of moral responsibility without some attention to our purpose in raising the question of someone's moral responsibility in a given context.

It's also no surprise on my analysis that theorists who emphasize different emotional functions in their characterization of moral responsibility disagree about the conditions under which people are morally responsible for what they do. In particular, we find a ready division between incompatibilists, who tend to emphasize the sanctioning function, and compatibilists, who tend to emphasize the appraisal and communicative functions. Moving forward with respect to points of dispute between the camps may be aided by further attention to the moral psychology of the blaming emotions and their conditions of appropriateness. In particular, I want to suggest that if we often implicitly attend to the different conditions of emotional appropriateness in our moral lives, our intuitive judgments about when people are morally responsible will be influenced by those considerations. Further, I expect that this influence will also be present when we think about people's moral responsibility in the context of philosophical thought experiments. This is a speculative claim, but it bears investigation, particularly in light of recent work on implicit and affective psychological process.[25]

[25] For an overview of some of the relevant empirical data, see (Bargh and Chartrand 1999).

There is no question that P. F. Strawson's "Freedom and Resentment" looms large over investigations into moral responsibility. One way in which Strawson influenced current debates is by motivating the idea that attributions of moral responsibility can be helpfully understood as "natural human reactions to the good or ill will or indifference of others towards us" (P. F. Strawson 1982, 53). I have here restricted my attention to just one set of natural human reactions, namely, the blaming emotions. If my account is plausible, these reactions to the quality of other's wills are a matter of significant complexity. Theorizing about moral responsibility must respect these intricacies if we are to progress.[26]

REFERENCES

Averill, James R. (1983). "Studies on Anger and Aggression: Implications for Theories of Emotion." *American Psychologist* 38 (11), 1145–60.

Bargh, John A, and Tanya L Chartrand (1999). "The Unbearable Automaticity of Being." *American Psychologist* 54 (7): 462–79. doi:10.1037/0003-066X.54.7.462.

Baumeister, Roy, K. D. Vohs, and C. Nathan DeWall (2007). "How Emotion Shapes Behavior: Feedback, Anticipation, and Reflection, Rather Than Direct Causation." *Personality and Social Psychology Review* 11 (2): 167.

Bennett, Jonathan (1980). "Accountability." In *Philosophical Subjects: Essays Presented to P.F. Strawson*, ed. Zak van Straaten. (Oxford: Clarendon).

Berkowitz, Leonard, and Eddie Harmon-Jones (2004a). "More Thoughts About Anger Determinants." *Emotion* 4 (2): 151–5.

—— (2004b). "Toward an Understanding of the Determinants of Anger." *Emotion* 4 (2): 107–30.

D'Arms, Justin and Daniel Jacobson (2000). "The Moralistic Fallacy: On the 'Appropriateness' of Emotions." *Philosophy and Phenomenological Research* 61 (1) (July): 65–90. doi:10.2307/2653403.

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

Dolinko, David (1991a). "Three Mistakes of Retributivism." *UCLA Law Review* 39: 1623.

—— (1991b). "Some Thoughts About Retributivism." *Ethics* 101 (3) (April): 537–59.

Ekman, Paul (1999). "Basic Emotions." In *Handbook of Cognition and Emotion*, eds. Tim Dalgleish and Mick Power, (New York: Wiley), 45–60.

[26] This paper has been much improved by helpful discussion with a friendly audience at the New Orleans Workshop on Agency and Responsibility. Antony Aumann, Andrea Scarpino, and David Shoemaker also provided particularly beneficial constructive comments and discussion. Finally, two anonymous referees from Oxford University Press supplied instructive comments.

Fehr, Ernst and Simon Gächter (2000). "Cooperation and Punishment in Public Goods Experiments." *The American Economic Review* 90 (4): 980–94.

——.(2002). "Altruistic Punishment in Humans." *Nature* 415, 137–40.

Fernández-Dols, José-Miguel, and María-Angeles Ruiz-Belda.(1995). "Are Smiles a Sign of Happiness? Gold Medal Winners at the Olympic Games." *Journal of Personality and Social Psychology* 69 (6): 1113–19. doi.10.1037/0022-3514.69.6.1113.

Fischer, John Martin and Mark Ravizza (1998). *Responsibility and Control: A Theory of Moral Responsibility*. (Cambridge, Cambridge University Press).

—— and Neal A. Tognazzini (2010). "The Physiognomy of Responsibility." *Philosophy and Phenomenological Research* (December). doi:10.1111/j.1933-1592.2010.00458.x.

Haidt, J. (2001). "The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgment." *Psychological Review* 108 (4): 814.

Hieronymi, Pamela (2004). "The Force and Fairness of Blame." *Philosophical Perspectives* 18 (1): 115–148. doi:10.1111/j.1520-8583.2004.00023.x.

Horstmann, Gernot (2003). "What Do Facial Expressions Convey: Feeling States, Behavioral Intentions, or Actions Requests?" *Emotion* 3 (2): 150–66. doi:10.1037/1528-3542.3.2.150.

Kraut, Robert E, and Robert E. Johnston (1979). "Social and Emotional Messages of Smiling: An Ethological Approach." *Journal of Personality and Social Psychology* 37 (9): 1539–53. doi:10.1037/0022-3514.37.9.1539.

Lazarus, R. S. (1982). "Thoughts on the Relations Between Emotion and Cognition." *American Psychologist* 37 (9): 1019–24.

—— (1984). "On the Primacy of Cognition." *American Psychologist* 39 (2): 124–9.

—— (1991). *Emotion and Adaptation*. (New York: Oxford University Press).

Leith, K. P, and Roy Baumeister (1996). "Why Do Bad Moods Increase Self-defeating Behavior? Emotion, Risk Tasking, and Self-regulation." *Journal of Personality and Social Psychology* 71 (6): 1250–67.

Lerner, Jennifer S., and D. Keltner (2000). "Beyond Valence: Toward a Model of Emotion-specific Influences on Judgement and Choice." *Cognition & Emotion* 14 (4): 473–93.

—— —— (2001). "Fear, Anger, and Risk." *Journal of Personality and Social Psychology* 81 (1): 146–59.

Lerner, Jennifer S., and Larissa Z Tiedens (2006). "Portrait of the Angry Decision Maker: How Appraisal Tendencies Shape Anger's Influence on Cognition." *Journal of Behavioral Decision Making* 19 (2) (April 1): 115–37. doi:10.1002/bdm.515.

Lerner, Jennifer S., Julie H. Goldberg, and Philip. E. Tetlock (1998). "Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility." *Personality and Social Psychology Bulletin* 24 (6) (June 1): 563–74. doi:10.1177/0146167298246001.

Levy, Neil (2011). *Hard Luck*. (New York: Oxford University Press).

Litvak, P. M, J. S Lerner, L. Z Tiedens, and K. Shonk. 2010. "Fuel in the Fire: How Anger Impacts Judgment and Decision-Making." *International Handbook of Anger*: 287–310.

McKenna, Michael (1998). "The Limits of Evil and the Role of Moral Address: A Defense of Strawsonian Compatibilism." *The Journal of Ethics* 2 (2) (June 1): 123–42. doi:10.1023/A:1009754626801.

—— (2004). "Responsibility and Globally Manipulated Agents." *Philosophical Topics* 32 (1/2): 169.

Macnamara, Coleen (forthcoming). "'Screw You!' & 'Thank You'." *Philosophical Studies*: 1–22. doi:10.1007/s11098-012-9995-3.

—— (2009). "Holding Others Responsible." *Philosophical Studies* 152 (October 23): 81–102. doi:10.1007/s11098-009-9464-9.

Ochs, Jack and Alvin E. Roth (1989). "An Experimental Study of Sequential Bargaining." *The American Economic Review* 79 (3) (June 1): 355–84.

Parkinson, Brian (1997). "Untangling the Appraisal-emotion Connection." *Personality and Social Psychology Review* 1 (1): 62–79.

——— (1999). "Relations and Dissociations Between Appraisal and Emotion Ratings of Reasonable and Unreasonable Anger and Guilt." *Cognition & Emotion* 13 (4): 347–85.

Pereboom, Derk (2001). *Living Without Free Will*. (Cambridge, MA: Cambridge University Press).

—— (2007). "Hard Incompatibilism." *Four Views on Free Will*: 85–125.

—— (2009). "Free Will, Love, And Anger." *Ideas y Valores: Revista Colombiana De Filosofía* (141): 169–89.

—— (2011). "Free Will Skepticism and Meaning in Life." In *The Oxford Handbook of Free Will*, ed. Robert Kane. 2nd edn. (New York: Oxford University Press).

Pillutla, Madan M., and J. Keith Murnighan (1996). "Unfairness, Anger, and Spite: Emotional Rejections of Ultimatum Offers." *Organizational Behavior and Human Decision Processes* 68 (3) (December): 208–24. doi:06/obhd.1996.0100.

Porter, Stephen and Leanne ten Brinke (2008). "Reading Between the Lies." *Psychological Science* 19 (5) (May 1): 508–14. doi:10.1111/j.1467-9280.2008.02116.x.

Prinz, Jesse (2004). *Gut Reactions: A Perceptual Theory of Emotion*. (New York: Oxford University Press).

—— and Shaun Nichols (2010). "Moral Emotions." In *The Moral Psychology Handbook*, ed. John Doris (Oxford: Oxford University Press), 111–48.

Roseman, Ira J., Cynthia Wiest, and Tamara S Swartz (1994). "Phenomenology, Behaviors, and Goals Differentiate Discrete Emotions." *Journal of Personality and Social Psychology* 67 (2): 206–21. doi:10.1037/0022-3514.67.2.206.

Scanlon, T. M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. (Cambridge, MA: Belknap, Harvard University Press).

Scherer, K. R. (1986). "Vocal Affect Expression: A Review and a Model for Future Research." *Psychological Bulletin* 99 (2): 143–65.

Shah, Nishi. (2003). "How Truth Governs Belief." *The Philosophical Review* 112 (4) : 447.

—— R. Banse, H. G Wallbott, and T. Goldbeck. (1991). "Vocal Cues in Emotion Encoding and Decoding." *Motivation and Emotion* 15 (2): 123–48.

Shaver, P., J. Schwartz, D. Kirson, and C. O'Connor (1987). "Emotion Knowledge: Further Exploration of a Prototype Approach." *Journal of Personality and Social Psychology* 52 (6): 1061–86.

Shoemaker, David (2011). "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility." *Ethics* 121 (3): 602–32.

Smith, Angela M. (2006). "Control, Responsibility, and Moral Assessment." *Philosophical Studies* 138 (December 15): 367–92. doi:10.1007/s11098-006-9048-x.

—— (2007). "On Being Responsible and Holding Responsible." *The Journal of Ethics* 11 (January 4): 465–84. doi:10.1007/s10892-005-7989-5.

Smith, C. A, and R. S Lazarus (1993). "Appraisal Components, Core Relational Themes, and the Emotions." *Cognition & Emotion* 7 (3): 233–69.

Sommers, Tamler (2007). "The Objective Attitude." *The Philosophical Quarterly* 57 (July): 321–41. doi:10.1111/j.1467-9213.2007.487.x.

Strawson, Galen (2002). "The Bounds of Freedom." In *The Oxford Handbook of Free Will*, ed. Robert Kane. (New York: Oxford University Press).

Strawson, P. F. (1982). "Freedom and Resentment." In *Free Will*, ed. Gary Watson. (Oxford, Oxford University Press).

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments*. (Cambridge, MA: Harvard University Press).

Wallbott, Harald (1998). "Bodily Expression of Emotion." *European Journal of Social Psychology* 28 (6) (November 1): 879–96. doi:10.1002/(SICI)1099-0992(1998110)28:6<879::AID-EJSP901>3.0.CO;2-W.

Watson, Gary (1993). "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In *Perspectives on Moral Responsibility*, eds. John Martin Fischer and Mark Ravizza. (Ithaca, NY: Cornell University Press) 119–50.

—— (1996). "Two Faces of Responsibility." *Philosophical Topics* 24 (2): 227–48.

Zajonc, R. B. (1980). "Feeling and Thinking: Preferences Need No Inferences." *American Psychologist* 35 (2): 151–75. doi:10.1037/0003-066X.35.2.151.

—— (1984). "On the Primacy of Affect." *American Psychologist* 39 (2): 117–23.

9

Unwitting Wrongdoers and the Role of Moral Disagreement in Blame[1]

Matthew Talbert

In "Culpable Ignorance," Holly Smith says that "[i]gnorance of the nature of one's act is the pre-eminent example of an excuse that forestalls blame" (543). So, to adapt an example of Smith's, a doctor who gives a patient the wrong treatment might avoid blame if she thought she was providing her patient with the best care she could. Whether the doctor avoids blame in this way will depend, in part, on whether she was blameworthy for being mistaken about what treatment was best. If the doctor prescribed the wrong treatment because of an earlier "benighting act" that impaired (or failed to improve) her cognitive position, then, *if we take the doctor to be culpable for the benighting act*, we may also regard her as blameworthy for her ignorance and for the consequences of her ignorance (547).

Smith's focus is on agents whose ignorance is circumstantial rather than moral: the doctor in the above example is unaware that the treatment she prescribes is not best for her patient, but she presumably is aware that she morally ought to give her patient the best treatment.[2] More recently, Gideon Rosen has applied an approach like Smith's to cases of both

[1] I would like to thank the participants at the 2011 New Orleans Workshop on Agency and Responsibility for their questions and comments on this paper, particularly Sarah Buss, T. M. Scanlon, Sarah Stroud, and Gary Watson. I am also very grateful to David Shoemaker and to two anonymous referees for Oxford University Press for their written comments.

[2] It is true that the doctor suffers from normative ignorance since she does not know that she has prescribed an inappropriate medication, but this is not what I mean by "moral ignorance." A doctor who suffers from moral ignorance (as I shall use the phrase) would not know that it is wrong to intentionally prescribe a patient the wrong medication. The doctor in the example suffers from circumstantial ignorance because she is not aware that, given her circumstances, her action will lead to bad consequences for her patient.

circumstantial and moral ignorance (Rosen 2004).[3] Rosen assumes that for an agent to be blameworthy for actions that issue from moral or circumstantial ignorance, the agent must be culpable for her ignorance. This assumption plays a crucial role for Rosen in a skeptical argument that calls into question many intuitive judgments of moral blameworthiness.

Though I reject Rosen's conclusion, I find his skeptical argument powerful. Indeed, and as I attempt to show, the argument is sufficiently strong that a relatively conservative approach to overturning it does not succeed.[4] Instead, if we are to avoid the skeptical conclusion, we must reject the plausible sounding assumption that unwitting wrongdoers are blameworthy only if they are culpable for their ignorance. To this end, I argue that while ignorance of the circumstances and consequences of one's actions often undermines blame, moral ignorance typically does not do so. For example, while I might not be blameworthy for injuring you if I was unaware that my action would have that result, I likely would be blameworthy if I were simply unaware that injuring you is impermissible. I will argue, moreover, that a morally ignorant wrongdoer can be blameworthy even if it is not her fault that she is ignorant of the moral status of her behavior, and even if it would be unreasonable to expect her to be aware of its status. In the context of making these points, I also try to shed light on the role that moral disagreement plays in our judgments of blameworthiness.

1. SKEPTICISM ABOUT MORAL RESPONSIBILITY

Gideon Rosen has advanced a skeptical perspective on moral responsibility based on the assumption that a wrongdoer is excused if she is nonculpably unaware that she does wrong. For Rosen, moral responsibility (for bad actions) amounts to being liable to the "sanctions" associated with moral blame; so, for Rosen's purposes, "we may simply identify moral responsibility with culpability or blameworthiness" (2004: 296). As Rosen puts it, "[s]kepticism about moral responsibility is thus the thesis that *confident positive judgments of blameworthiness are never justified*" (296).

Like many philosophers working on the subject, Rosen takes moral blame to involve certain emotional responses. For example, "you blame X for doing A, when you resent him or feel indignant towards him for

[3] Also see Rosen (2003). Michael Zimmerman (1997) offers a view similar to Rosen's; also see Zimmerman (2008).

[4] The conservative approach I have in mind is the one pursued by William FitzPatrick (2008). I discuss FitzPatrick's view in Section 2.

having done it" (296). Thus, if we judge that X is blameworthy, this means "that X is *liable* [in our judgment] to a negative emotional response of this sort for having done A, or equivalently, that some such response would be *appropriate* or *fitting*" (297). Rosen's skepticism amounts, then, to skepticism about the aptness of negative reactive attitudes like resentment. In what follows, I will understand moral responsibility in just the way Rosen suggests: the primary question will be about whether judgments of blameworthiness, and responses like resentment, are apt.

Rosen's skeptical argument proceeds as follows. Wrongdoing is either witting or unwitting. While a *knowing* wrongdoer may be directly blameworthy for her behavior, an unwitting wrongdoer will be blameworthy for her wrongdoing only derivatively or indirectly. This means that an ignorant wrongdoer will be morally responsible, in the sense of being open to blame, only if she is culpable for her ignorance.[5] However, according to Rosen, "[i]gnorance is culpable only if it derives from culpable recklessness or negligence in the management of one's opinion" and this prior recklessness will be culpable only if it was knowing or was itself the product of knowing mismanagement of one's opinions (302). So, as Rosen puts it, ignorance is culpable—and unwitting wrongdoing blameworthy—only if it results, at some point, from an akratic act in which the agent knowingly violated an epistemic or moral duty:

> One is responsible for the act done from ignorance only if one is independently responsible for something else . . . this entails that *the only possible locus of original responsibility is an akratic act*. . . . Our first sin must be a knowing sin—a sin done in full knowledge of every pertinent fact or principle. (307)[6]

Rosen applies this perspective to cases of circumstantial ignorance (like that of the doctor in the introduction), but also to cases of moral ignorance. For example, Rosen's view suggests that an "ancient slaveholder who . . . believes that it is morally permissible for him to buy and sell" slaves is blameworthy for his behavior "only if he is culpable for the moral ignorance from which he acts" (304).

The skeptical force of this position emerges when we consider the possibility that many unwitting wrongdoers have never committed the kind of akratic act that would make them culpable for their ignorance.

[5] Following the authors I discuss, I use "culpable" and "blameworthy" interchangeably.

[6] Zimmerman's view is similar: if "one is culpable for ignorant behavior, then one is culpable for the ignorance to which this behavior may be traced" (Zimmerman 1997: 418). However, "one is never in direct control of whether one is ignorant," so culpability for ignorant behavior must be indirect, which "presupposes direct culpability for something else. . . . Hence all culpability can be traced to culpability that involves lack of ignorance" (418).

Imagine an "ambitious capitalist who is mistaken about where to draw the line between permissibly aggressive business practices and reprehensibly ruthless business practices" (305). Rosen supposes (and I agree) that it is not hard to imagine a case in which the capitalist does something wrong without recognizing that his behavior is wrong and that he has been as reflective and careful as it is reasonable to expect him to be. Perhaps the capitalist's moral education was deficient, or perhaps the case he considers is "just a hard case and after thinking about it for a decent interval he has simply arrived at the wrong answer" (305). If we agree that "in reaching his [moral] conclusion our capitalist has not been reckless or negligent," Rosen thinks we should also agree "that his moral ignorance is not his fault" and that it would therefore "be a mistake to blame him for the wrong he does" (305).

Another example features Bill, who decides to tell a self-serving lie to his wife even though he "knows that it's just plain wrong to lie to your wife" (305). Again, it is possible that Bill had a defective moral education that led him to believe, "through no fault of his own, that while moral considerations have some weight, they are not in general decisive" (305).[7] Bill may thus be "blamelessly (though mistakenly) convinced that the balance of reasons comes down in favor of lying" (306). Rosen argues that in this case it would be unreasonable to blame Bill for telling the lie. Rosen asks: "Does it make sense to subject someone who blamelessly believes that he should do A, and then does it, to moral sanctions—to recrimination, resentment, righteous anger, contempt?" (306). Rosen thinks this doesn't make sense. Instead, Bill's wife should respond this way:

> Poor Bill. Through no fault of his own he found himself believing that all things considered, he should lie. Given that he found himself in that state, I can hardly fault him for lying. Holding the judgment fixed, the lie itself was a perfect manifestation of practical rationality. I can fault him for the lie only if I can fault him for believing that in the circumstances, his selfish interests were more important than my moral interests. Since by hypothesis, it is not his fault that he held this view, I have no option but to conclude that he is not properly culpable for his bad action. (306)[8]

Rosen's official skeptical conclusion is that "*it would be unreasonable to repose much confidence in any particular positive judgment of responsibility*" (308). This follows from the claim that an unwitting wrongdoer is blameworthy only if she has an akrasia-involving "*inculpating* history" (309) together with the claim that "it is almost always unreasonable to place

[7] Unlike the ambitious capitalist, Bill is mistaken about the force of moral considerations rather than about a moral principle.

[8] For discussion of a related example, see Rosen (2008: 605–9).

significant confidence in" the judgment that an agent has such a history (308). In a recent response to Rosen, William FitzPatrick has argued that in fact we often have good grounds for attributing akrasia to ourselves and others (FitzPatrick 2008: 593–9). However, as FitzPatrick notes, this leaves untouched Rosen's conclusion that blameworthy wrongdoing is always either knowing or the result of knowing wrongdoing (599). We are left, then, with a striking skeptical challenge, for very many ordinary wrongdoers may not have a relevant instance of akrasia in their past and so are not morally responsible for their wrongdoing.[9] Below, I consider Fitzpatrick's attempt to meet this skeptical challenge.

2. FITZPATRICK'S REPLY TO ROSEN

FitzPatrick accepts the "intuition at the core of [Rosen's] argument, that it is unfair to blame someone for an action done out of ignorance that he cannot fairly be blamed for having" (2008: 601). However, FitzPatrick rejects Rosen's claim that unwitting wrongdoers are culpable for their ignorance only if it resulted from knowing wrongdoing. According to FitzPatrick, an unwitting wrongdoer will also be culpable for her ignorance if she "could reasonably have been expected to take measures that would have corrected or avoided it" (609).

To make his case, FitzPatrick expands on Rosen's "ambitious capitalist" example and considers Mr Potter, the ruthless businessman in Frank Capra's 1964 film, *It's a Wonderful Life*. Potter is callous, vindictive, greedy, and dishonest, but let us suppose that he wrongly believes that his behavior is permissible. FitzPatrick argues that even if Potter did not akratically acquire his false moral beliefs, there are three central factors "that make most of us confident that Potter's moral ignorance is culpable" (605). First of all, Potter never engaged in the kind of moral reflection that might have helped him correct his moral ignorance, yet "[t]here were no relevant limitations in his social context or in his capabilities" that would have made such reflection unreasonably difficult (605). This suggests that "[t]he failure of adequate reflection" on Potter's part "was instead the result of voluntary exercises of vices such as overconfidence, arrogance, dismissiveness, laziness, dogmaticism, incuriosity, self-indulgence, contempt, and so on" (605). According to FitzPatrick, Potter "could thus reasonably have been expected to take steps that would have eliminated that [moral] ignorance, by refraining from exercising those vices and instead taking

[9] This is close to the skeptical conclusion advanced by Zimmerman (1997: 425).

advantage of" available opportunities for moral and epistemic improvement (605).

So on FitzPatrick's view, given Potter's general capacities, and the fact that "the opportunity for improved normative understanding was clearly present in his social context," it is reasonable to hold Potter to the expectation that he correct his moral ignorance (603–4). Thus, Potter is culpable for his ignorance and blameworthy for behavior that flows from it.

I agree with FitzPatrick that Potter is blameworthy for his bad behavior, but I am not convinced that it is reasonable to expect Potter to correct his moral ignorance. As FitzPatrick sees it, explaining Potter's moral ignorance in terms of his vices helps us see that this moral expectation is reasonable: Potter's failure to engage in reflection wasn't forced on him by circumstances or inability, it resulted from vicious choices that he was free to omit. I contend, however, that this emphasis on Potter's vices is unhelpful for the point FitzPatrick wants to make.

FitzPatrick characterizes Potter's failure to correct his moral ignorance as "the result of voluntary exercises of vices" (605). But what does Potter do *voluntarily*? He does not voluntarily behave in ways that he regards as impermissibly vicious—FitzPatrick's point hangs on Potter not being a *knowing* wrongdoer. Rather, Potter voluntarily exercises his vices in the sense that he acts voluntarily and his vices shape his practical judgments in characteristic ways. Now if Potter's vicious actions are voluntary in this sense, then he might have omitted them if he judged himself to have reason to do so, but one symptom of Potter having his vices is that he sees little in favor of such an omission. If Potter is afflicted with all the vices FitzPatrick mentions, he is not likely to see much in favor of exploring the opportunities for moral improvement available to him, so perhaps it is unreasonable to expect him to do so.

Neil Levy argues similarly against FitzPatrick that it is not reasonable to expect Potter to correct his vices because it is likely not subjectively rational for Potter to do so. When we think about what can reasonably be expected of a person, Levy says we should consider what that person can do by way of a rational reasoning procedure, and what an agent can do in this way is a function of her internal reasons, her "actual representations and proattitudes" (Levy 2009: 736). As Levy notes, even if Potter's failure to subject his values to scrutiny is a manifestation of epistemic and moral vices,

> by his lights, Potter governs his normative views adequately. He gives competing views the attention he takes them to deserve. . . . But if Potter does not see that he is managing his moral views badly, he has no (internal) reason to manage them any differently. Potter exhibits epistemic vices aplenty, but because he does not conceive of them as vices, he has no internal reason to refrain from so doing. (737)

Levy concludes that since Potter "could not rationally have taken advantage of the opportunities for moral improvement" with which he was presented, "we cannot reasonably expect him to do so" (735).[10]

There is also the question of how Potter acquired his vices. As FitzPatrick notes, it may seem "problematic that the vices Potter exhibits in his epistemically debilitating choices may trace back to his childhood and may be largely a result of moral (bad) luck" (607). FitzPatrick deflects this worry by noting that for most people "character traits are not merely given but are formed, reformed and continuously shaped by our choices from the point of moral maturity onward" (608). But this is not a helpful response to the problem of constitutive moral luck since it simply pushes the problem back to earlier stages of Potter's development. Apparently, Potter made poor choices as he shaped his character. But why did he make such poor choices? By hypothesis, he did not *knowingly* make poor character-forming choices, so perhaps Potter's tendency to make "epistemically debilitating" choices is explained by a tendency to see poor self-forming choices as choice-worthy. But in this case, Potter's self-forming choices would seem to be shaped by incipient versions of the vices that his self-formation is invoked to explain. And given the presence of these incipient vices, why should we expect Potter to make the right choice when it comes to choosing whether to act in a way that will strengthen his vices?

Finally, it is worth noting that FitzPatrick says that "cultural and historical contexts" may make it unreasonable to expect a wrongdoer to know better because "the relevant [moral] knowledge isn't reasonably available" (612). Aristotle, for example, may not be blameworthy for

[10] This discussion is taken up again in (Levy 2011: 124–8). A referee for Oxford University Press points out that we use the word "expect" in normative and descriptive senses (to use the referee's terminology). The descriptive sense presumably has to do with what we anticipate from an agent in view of his capacities and the context in which he acts, whereas the normative sense has to do with standards against which we measure agents. When I say, "I expect it to rain tomorrow," I use "expect" in the descriptive sense. I anticipate rain; it seems like a safe bet. Neither Levy nor FitzPatrick expect moral reform from Potter in this sense; neither regards it as a safe bet. (Though FitzPatrick is keen to draw attention to Potter's abilities and his social context and the way that these make reform at least possible; he regards such reform as a live possibility even if not as a likely outcome.) By contrast, both Levy and FitzPatrick agree that Potter falls short of our expectations in at least one normative sense: they both regard Potter's behavior as morally subpar. The question is whether this subpar behavior opens Potter up to moral blame. Here we might identify a slightly different sense in which "expect" can be normative, the sense in which we talk about *holding* someone to an expectation. Fitzpatrick thinks it is reasonable to hold Potter to our moral expectations because it is possible for him to live up to these expectations given his abilities and the context in which he acts. Levy thinks it is unfair to hold Potter to our moral expectations because he can fulfill these only by behaving irrationally.

thinking slavery permissible because slavery would have presented "a genuinely hard case for someone in Aristotle's circumstances" (FitzPatrick: 600 n. 24). Since social and cultural factors can turn a question that is easy *for us*—like "Is slavery wrong?"—into a difficult question for someone in ancient Greece, this undermines the culpability of the morally ignorant Greek. But Potter's vices seem to turn apparently easy moral questions into difficult ones, so perhaps these internal obstacles to moral knowledge should (at least on FitzPatrick's account) have the same excusing force as cultural impediments to moral knowledge.

3. REJECTING THE SKEPTICAL ARGUMENT

I agree with Levy that it is unreasonable to expect Potter to recognize and correct his moral ignorance. If this is right, and if it is also true that Potter never committed a relevant act of knowing wrongdoing, then he would not be culpable for his moral ignorance on either FitzPatricks or Rosen's account of culpable ignorance. What goes for Potter presumably goes for many other unwitting wrongdoers, so if we accept the assumption—endorsed by Rosen, FitzPatrick, and Levy—that unwitting wrongdoers are blameworthy only if they are culpable for their ignorance, then we should conclude that many (perhaps very many) ordinary wrongdoers are not blameworthy for their bad behavior. Fortunately, we need not accept this conclusion because there is good reason to reject the assumption that unwitting wrongdoers are blameworthy only if they are culpable for their ignorance.

The central instance of wrongdoing in *It's a Wonderful Life* occurs when Mr Potter keeps $8000 that George Bailey's uncle Billy misplaced. Potter keeps the money, believing (and hoping) that this will cause the foreclosure of Bailey Building and Loan, as well as the prosecution of George Bailey for bank fraud, and that this will leave the town of Bedford Falls ripe for economic exploitation. Recall that we are concerned here with whether someone like Potter is morally responsible for his behavior in the sense of being an appropriate target for blaming attitudes like resentment. Plausibly, if George Bailey were to resent Potter, this resentment would be provoked by the fact that Potter willingly acted so as to cause a bad outcome for George, and that he did so because of a desire to injure George and to extract an economic benefit from George's misfortune. These aspects of Potter's behavior make George's resentment natural and appropriate. If Potter's actions had lacked these features—if he had been coerced or had thought he was acting in George's best interests—then it would be inappropriate for George to resent Potter. But if Potter deliberately injured George, and did so for self-serving reasons, then George can quite reasonably point to these facts to explain and justify his resentment even if Potter happens to regard his behavior as permissible.[11]

The features of Potter's behavior just mentioned make George's blaming responses appropriate because of the way these features are tied up with Potter's expression of contemptuous judgments and attitudes towards George.[12] For example, if Potter deliberately injured George for self-serving reasons, then this behavior expresses the implicit judgment that George's welfare is unimportant in comparison with whether Potter achieves his aims. (In fact, it's not just that George's welfare is unimportant for Potter, he regards the possibility of injuring George as a reason for acting.) Given the contemptuous judgments that inform Potter's behavior, George can reasonably regard Potter's behavior as unjustifiable and morally offensive—rather than as merely harmful or unwelcome—and thus as proper grounds for resentment.

Similar points apply to Rosen's example of Bill. Recall that, through no fault of his own, Bill believes that his selfish reasons for lying to his wife trump the moral reasons against doing so. Rosen says that

> [a]nyone who bears our principles [about culpable ignorance] in mind and nonetheless judges that Bill is responsible for lying to his wife, is thereby committed to the view that somewhere in the story of that lie there exists a full-blown episode of altogether knowledgeable wrongdoing. (2004: 307–8)

"Responsibility" is an elastic term, so we can interpret the claim, "Bill is responsible for lying to his wife," in different ways. By hypothesis, Bill is not at fault for thinking that he has decisive reason to lie, so Bill is not responsible for lying to his wife in the sense of having played a particular sort of causal role in the process that led him to think that lying is the thing to do. But whether Bill is responsible for lying in this sense is not what is at issue. Again, the issue is whether Bill is morally responsible in the sense that attitudes like resentment would be a fitting response to him. The central question about Bill is whether it makes "sense to subject someone [like Bill] who blamelessly believes that he should do A, and then does it, to moral sanctions—to recrimination, resentment, righteous anger, contempt?" (306).

[11] In Section 5, I address the concern that Potter's nonculpable moral ignorance makes it unfair to blame him.

[12] For other applications of this kind of approach to blameworthiness, see T. M. Scanlon (1998) and (2008), Angela Smith (2005) and (2008), and my own (2008) and (2012a).

I take it that Rosen believes that Bill is responsible for lying, in the sense of being open to attitudes like resentment, only if he is responsible for lying in the sense of having played the right sort of role in bringing about his own tendency to favor lying. But Bill's case is a good example of why moral responsibility, in the sense of blameworthiness, does not require that agents have this sort "inculpating history." It is reasonable for Bill's wife to blame him because of the way his lying expresses Bill's morally faulty judgments and attitudes. Bill has these faults, and they contribute to his behavior, regardless of whether he is *at fault* for having them, and regardless of whether Bill genuinely believes that lying to his wife is the thing to do.

One of Bill's faults has to do with how his wife's interests rate when he is trying to figure out what to do: Bill is willing to overlook his wife's interests when they conflict with Bill avoiding trouble. It is reasonable, then, to attribute to Bill the judgment that his wife's interest in not being lied to can be overlooked, if that is how Bill can get what he wants. If Bill's wife were to find out how her interests rate with her husband, that he lied to her and the basis on which he did so, then she would have good grounds for blame. That is, it would be appropriate for her to be offended and hurt by what Bill's action expresses, for her to protest that her interests ought to rate more highly with Bill, to resent him for his callousness, to insist that he change his ways, and so on.

The general perspective developed above is as follows. Even if a wrongdoer is ignorant of the fact that her behavior is wrong, and even if this ignorance is not her fault, her actions may still express the contemptuous judgment that certain others do not merit consideration, that their interests do not matter, and that their objections can be overlooked. If one is injured by a wrongdoer who is moved by such judgments, then the attitudes and responses involved in moral blame are reasonable regardless of what the wrongdoer thinks about the moral status of her behavior. I will develop this perspective below, but there is enough here to see one way of rejecting the skeptical perspective on moral responsibility with which this paper began. Rosen's skepticism depends on assuming that unwitting wrongdoers are open to blame only if they are culpable for their ignorance. However, as I have argued, certain features of an unwitting wrongdoer's behavior can qualify her for blame regardless of whether she is culpable for her ignorance.

4. MORAL IGNORANCE AND MORAL DISAGREEMENT

Angela Smith has applied a view like the one I just outlined to show that agents are often morally responsible for things that are not under their direct control: their desires and emotions, their advertences and inadvertences, and so on. We are responsible for these things, on Smith's view, because of the evaluative judgments they express and the importance of these judgments for our interpersonal relations.

Smith gives special attention to the fact that "we often take what a person notices and neglects to have an enormous amount of expressive significance" (2005: 242). What a person notices attracts our attention because we assume a connection between what one notices and what one values. According to Smith, "if one judges some thing or person to be important or significant," this should "have an influence on one's tendency to notice factors which pertain to the existence, welfare, or flourishing of that thing or person" (244). And if one fails to notice such factors, this "is at least some indication that she does not accept this evaluative judgment [about the thing's importance]" (244). As Smith says, I may fail to "notice when my music is too loud," or that "my advice is unwelcome," or that "my assistance might be helpful to others," and *even if these failures are involuntary,* they may still indicate "that I do not judge your needs and interests to be important, or at least that I do not take them very seriously" (244).

Though she does not describe them this way, the examples Smith cites are instances of circumstantial ignorance. On Smith's view, then, judgments about a circumstantially ignorant wrongdoer's moral responsibility should track the plausibility of associating her behavior with interpersonally significant evaluative judgments—particularly judgments about the normative status of the needs and interests of those affected by the wrongdoer's actions and omissions.

As should be clear from the previous section, I am largely in agreement with Smith, but it is worth emphasizing that we cannot always infer that an agent does not care about something from the fact that she fails to notice how her actions (or omissions) will affect it. Smith is aware of this; she notes that in some cases of inadvertence, "the person in question may be extremely tired or under a lot of stress" and this may "block the normal inference from what a person notices to what she cares about" (244 n. 14). Indeed, we sometimes hesitate to make the inference from what a person notices to what she cares about even when we cannot point to stressors that intuitively explain an agent's inadvertence. Sometimes, and for no obvious reason, we just forget things, or fail to notice them, yet we may have as much concern for these things as we ought to have. This can be true even in cases in which an inadvertence has horrible consequences, such as when a parent mistakenly leaves a child in a hot car. In some of these cases, it is no doubt correct to infer that the parent has a condemnable lack of concern for the child's welfare. But in other cases, it is difficult to read the testimony of the parents involved and come away with the thought that their forgetting is best explained by a morally deficient degree of parental concern. Many of these parents seem to have been as concerned with their child's welfare as morality requires; yet, they left them behind all the same.[13]

We may often be unsure how to assess the sorts of cases just described, but my general point is uncontroversial: we should be cautious before we conclude that an unwitting wrongdoer's actions or omissions express evaluative judgments that might ground blame. More controversially, I would argue that such caution is particularly appropriate in cases in which the unwitting wrongdoer is ignorant of features of the context in which she acts, or of the likely consequences of her behavior. There is correspondingly less reason to be cautious about attributing blame-grounding judgments in cases of moral ignorance in which an agent is aware of the consequences of his behavior, but is unaware that it is wrong to bring about those consequences.

One consideration in favor of this last claim is obvious: if an agent is aware that act *A* will have consequence *C*, then, when she *A*'s, it is usually reasonable to attribute to her at least an implicit judgment about how the prospect of *C* bears on the question of whether to *A*. Of course, the fact that an agent knows that her action will cause *C* does not entail that she formed an *objectionable* judgment about the significance of *C*. The reason we should expect the actions of morally ignorant wrongdoers like Mr Potter to express objectionable judgments is that their moral ignorance is—unlike a lot of circumstantial ignorance—often a manifestation of a normative disagreement the agent has with those who object to his behavior. More specifically, the moral ignorance of such an agent is often tied up with a perspective that regards as unobjectionable the very thing that his victims take to make his behavior objectionable.

Mr Potter's moral ignorance, for example, is partly constituted by the fact that he thinks he assigns George's interests their proper weight in his practical deliberations. Of course, we disagree with him about the normative significance of George's interests, and since we take our position to be the right one, we regard Potter as lacking moral knowledge. In such a case—where a wrongdoer's moral ignorance is part and parcel of a profound normative disagreement between us and him—it is no surprise that his behavior expresses judgments and attitudes that are, by our lights, objectionable.

Of course, our normative judgments may conflict more or less profoundly with the judgments we attribute to morally ignorant wrongdoers,

[13] Cf. Levy (2011: 182). Gene Weingarten's (2008) Pulitzer Prize winning article, "Fatal Distraction," offers a detailed and affecting account of some of these cases.

and as the severity of this disagreement decreases, so too may the intensity of our blame. Take the case of Robert E. Lee. Those interested in burnishing Lee's reputation often note that he joined the secessionist cause in the US Civil War out of loyalty to the state of Virginia. Lee did wrong, I assume, in leading Confederate troops against the Union, but he thought he was doing what duty and loyalty required. Suppose, however, that Lee was motivated to support the Confederacy solely by racial hatred and a desire to see slavery preserved. In this case, Lee would still be a morally ignorant wrongdoer, but his defenders would face a much more difficult task in convincing us to excuse him. This is because Lee's actions would have been associated with judgments and attitudes that we find deeply objectionable. An excuse like, "Lee thought he was doing the right thing," would, I submit, have little influence on us if we found the judgments that informed his choices thoroughly repugnant. And if we do have some tendency to accept this excuse, I suspect it is because the judgments about the value and requirements of loyalty that purportedly guided Lee are not utterly foreign or repellant to us.

5. INTERNAL REASONS AND THE FAIRNESS OF BLAME

In this section, I respond to the claim that it is unfair to blame an unwitting wrongdoer who can't reasonably be expected to omit her bad behavior. Along the way, I develop the suggestion I made at the end of the last section that moral blame often rests on the recognition of a moral disagreement between ourselves and the one we blame. I will argue that what matters most for judgments of blameworthiness are the considerations that count as reasons for us (as issuers of blame), and not whether these considerations could have counted as reasons for those we blame.

Let us return to Neil Levy's discussion of Mr Potter's blameworthiness. According to Levy,

> [i]t is not reasonable to blame agents for actions they cannot (intentionally) omit by way of some reasoning procedure; we cannot hold them responsible for failing to do things they could do only by chance or through a glitch in their agency. (Levy 2009: 739)

For Levy, part of the problem with blaming Potter is that his actual proattitudes do not rationalize a decision to reassess his values and epistemic practices, and this means that he lacks a fair opportunity to avoid the actions for which we would blame him. As Levy sees it, "agents have a fair opportunity to avoid performing actions . . . only if they could have rationally chosen to omit the action, and what agents can do rationally is a function of their internalist reasons alone" (739).

I believe that the sense in which Potter has trouble avoiding wrong actions does not make it unfair to blame him. Let us agree that Potter has no internal reason to avoid a morally bad action, and that if he had avoided the action, this would have been because of a "glitch" in his agency, as Levy puts it. Importantly, this does not mean that Potter's *actual* bad behavior is the result of a glitch, mistake, compulsion, or anything similar. If someone is subject to a compulsion or to "glitchy" agency, then some of her actions may be unavoidable and she will have a claim to being excused for this behavior. But the basis of excuse here—the reason the compulsive agent is not appropriately targeted with resentment—is that her behavior is not under her control in such a way that it can express the objectionable judgments that would make blame appropriate. Usually, an agent's inability to avoid an action goes together with that action not being under the agent's control in this way, but cases of moral ignorance like Potter's are an exception. Even if Potter does not have rational access to doing the right thing, his bad behavior may be a manifestation of perfect reflective self-control and subjective practical rationality: he may be acting just as he likes and for reasons that really do speak in favor of so acting, given his proattitudes. But if Potter's bad actions are guided in this way by his judgment about reasons, then they can express the kinds of attitudes and evaluative judgments that I have argued make blaming responses appropriate.[14]

In addition, Potter presumably has the capacity to avoid wrongdoing in the sense that, for any wrong action, he would have refrained from performing that action if he had taken himself to have a decisive reason to do so. Of course, if Potter's actual constitution entails that he will see no such reason, then he will not rationally avoid his wrongdoing, and it is this lack of rational access to avoiding wrongdoing that Levy thinks makes blame unfair. But if Potter's bad actions are guided by his judgments about how to behave, and if he would have acted differently if he had recognized a reason to do so, then his wrongdoing is unavoidable mainly in the sense that an action that is wrong (by our lights) is bound to seem choice-worthy to Potter. Since this sort of unavoidability does nothing to make Potter's behavior less knowing, deliberate, and dismissive of the interests of those his actions affect, I do not see why it should make blame unfair.

So, while I agree with Levy that it is not reasonable to demand that Potter behave contrary to his internal reasons, what I think this tells us is

[14] There is an important relation here with the theme of Harry Frankfurt (1969). I develop this point in (2012a).

that having subjective rational access to avoiding an action is not a requirement on being properly blamed for that action. In other words, an agent may be blameworthy for performing an action that she had no internal reason to avoid. In my view, approaches like Levy's focus too much on the perspective of the blamed agent, on what can count as a reason for him, and what can be expected of him. In fact, I think that our judgments about blameworthiness are often formed—and rightly so—on the basis of the normative considerations that *we* recognize, and not on whether those we blame could have assigned the same normative weight to these considerations that we do.

Suppose we believe that Potter unjustifiably injured George by a certain action, and that while Potter willingly acted with the intent of injuring George, he could not have been expected to omit this action because, through no fault of his own, he had no internal reason to do so. If we regard Potter's treatment of George as unjustified, we at least think that Potter does not agree with us about the status of things that are *for us* important normative considerations. George, in particular, may think that his welfare is valuable, and that a person of good will who is appropriately sensitive to value will see a reason to avoid actions that are contrary to his welfare. From George's perspective, then, since Potter has failed to see his welfare as reason-giving, this suggests that he is not disposed toward George as a person of good will would be.

Because of George's judgment about the importance of his own welfare, Potter's action appears unjustifiable and offensive to George. Potter, of course, will see things differently: he thinks his treatment of George is entirely appropriate and that he has no reason to refrain from it. This may mean that Potter has no internal reason to omit his action, but regardless of whether this is the case, it is still true that Potter does not view the prospect of George's injury as a reason to refrain from his action. Thus, Potter's action expresses the offensive judgment that George's welfare is not particularly valuable.[15] And even if George agrees that Potter had no (internal) reason to refrain from his action, it is inappropriate to demand that George regard Potter's action as unobjectionable because this asks George to concede that his welfare is not normatively significant. Even if George

[15] What if a wild animal injured George? Would we say that the animal's behavior expresses a blame-grounding rejection of the significance of George's welfare? I think not. The behavior of animals is not morally significant in this way because a judgment like, "George's injuries don't matter," is not meaningfully attributed to them. I don't mean that nonhuman animals can never be described as being sensitive to reasons but rather that their behavior does not have the same significance for us as that of beings whose behavior is more richly and generally informed by evaluative judgments.

agrees that, in the internalist sense, Potter has no reason to care how he fares, it is still appropriate for George to insist that his welfare is valuable and to see Potter's rejection of this claim as a manifestation of ill will.

If we agree with George that his welfare is valuable, and that a person of good will would see his welfare as a source of reasons, we should conclude that Potter's judgment about the significance of George's welfare reasonably elicits blaming responses on George's part. By contrast, if we think that George's welfare has little normative significance, we are likely to find George's blame inappropriate. However, this conclusion would stem not from the thought that Potter can't reasonably be expected to recognize the significance of George's welfare, but rather from our disagreement with George about this significance.

It may be thought that Potter can reject the normative status of George's welfare in a morally relevant way only if he had rational access to an accurate judgment about this status.[16] I don't see why this should be so. What reason is there to think that the expressive significance, for George and for us, of Potter's behavior should hang on whether Potter might rationally have made different judgments about reasons? After all (and as I pointed out above), Potter's actual behavior, and his actual failure to agree with George about the significance of his welfare, need not have been caused by a glitch, but may be a thoroughly deliberate and controlled exercise of his agency. And suppose that Potter *did* have some internal reason to take George's welfare as a constraint on his behavior, but that he still unjustifiably and deliberately injured him. If we did not already think that Potter's action was offensive in a blame-grounding way, I do not see how adding this element to the story would make blame appropriate. What matters is that George believes (and we believe) that he has standing that makes Potter's treatment of him illegitimate. Regardless of whether Potter could have been rationally moved to accept our view about the treatment to which George is entitled, his knowing and willing behavior demonstrates that he rejects this view.

In one of his discussions of internal reasons, Bernard Williams characterizes blame as functioning like advice given after the fact—it tells wrongdoers what they ought to have done. As Williams says, "if 'ought to have' is appropriate afterwards in the modality of blame, then (roughly) 'ought to' was appropriate at the time in the modality of advice" (Williams 1995: 40). However,

[16] This would be similar to Gary Watson's claim that a wrongdoer flouts a moral demand in a way that justifies resentment only if he is capable of recognizing the validity of the demand (Watson 2004 : 234).

> 'ought to' in the modality of advice implies 'can,' because advice aims to offer something as a candidate for a deliberative conclusion. If ϕ-ing is not available to the agent, 'You ought to ϕ' cannot function as a piece of advice about what he should now do; when it is a matter of what I am to do, manifestly 'I cannot' acts as a stopper. (40)

Similarly, we might worry that if a wrongdoer has no internal reason to refrain from an action, this is a "stopper" on blame because there is no point to the advice that is supposedly implicit in blame.

I agree that when we blame a person, we are typically committed to the claim that she ought not to have done what she did. But the "ought" of blame is not always, or at least not solely, in the mode of advice. It can also be an "ought" that points to an ideal or to a moral fact: the fact, for example, that George deserves better treatment from Potter, and that a person of good will would not have disregarded George's interests and welfare. This sort of "ought" does not imply "can." Of course, as a form of advice, it may be futile for George to insist to Potter that his welfare matters. But as a form of *protest* this insistence is a natural way of expressing the moral offense and resentment involved in blame.[17] The reasons that matter most for blame are *our* reasons—the considerations recognized by the victim and by those who sympathize with her. Whether these considerations can also be reasons for the blamed party is, I think, of secondary importance.[18]

[17] I argue that moral blame is sometimes best construed as a form of moral protest in (2012a).

[18] A referee for Oxford University Press suggests that I might note the affinity between what I say in this section and, e.g., the sort of view once defended by Philippa Foot (1972). There certainly seem to me to be affinities here but I am not sure how to characterize them; I will simply note that I am very attracted to the sort of view Foot outlines. Another referee suggests that I can in fact say that it is reasonable to expect wrongdoers like Mr Potter to correct their moral ignorance. This suggestion relies on my characterization of these wrongdoers as expressing ill will, the fact that expressions of ill will (arguably) flout the demand to treat others with reasonable regard, and the fact that this demand is connected to our normative expectations of other agents. I don't doubt that this proposal is workable but (in this and previous work) I prefer to grant to opponents that certain moral demands and expectations are off the table in Potter's case (and other related cases). I think this helps to make the debate more clear than it would be otherwise. However, the referee's comment prompts me to note that my disagreement with FitzPatrick is in some ways less stark than I presented it in Section 2. FitzPatrick and I agree that Potter can reasonably be held to our moral standards in the sense that we regard him as a proper target of moral blame. (I suspect this is the sense in which the referee believes I can say that we properly expect better things of Potter.) However, for FitzPatrick, the legitimacy of blaming Potter depends on his having been able to avoid his moral ignorance. On my view, Potter's blameworthiness does not depend on this but only on the actual judgments that inform his behavior.

6. CIRCUMSTANTIAL KNOWLEDGE AND BLAME

On the account I have presented, when it comes to thinking about how knowledge relates to blameworthiness, we should consider what a wrongdoer needs to know in order for her actions to express the attitudes and judgments that make blame appropriate. What does an agent need to know before we attribute to her a judgment like, "your objections and interests can be overlooked when I am deciding how to act"? As I have argued, awareness that one's actions are wrong is not necessary for the presence or expression of such judgments. Much more relevant in this context is knowledge of the effects that one's actions will have on others. If a person knows that her action will injure someone who objects to this result, then it is reasonable to attribute to her the judgment that the other's injuries and objections are not a decisive reason to refrain from the action. Judgments like this help to ground blaming responses because of the way they call into question the standing of the other to raise objections to certain forms of treatment and to cite her interests as normative considerations.

Of course, we will not always think that an agent is open to blame when her actions express the judgment that another's interests and objections may be overlooked. As I suggested with my example of Robert E. Lee in Section 4, our views about an agent's blameworthiness vary with the degree to which we accept the judgments that move her. Most people believe that it is permissible to use violence in self-defense against an aggressor. If we take an instance of self-defense to be justified, then we likely think that whatever objections the aggressor might raise to the force used against him are rightly overlooked. Thus, we would be unlikely to blame the person who defended herself even if she acted on the judgment that, in the relevant instance, the aggressor's welfare can be overlooked. The injured aggressor may resent the one who injured him in self-defense, but by our lights, and by those of the one who acted in self-defense, this blame will appear inapt and unreasonable.

Similarly, if I uproot a plant, my action may express an implicit judgment that plants do not have standing to raise objections to the way I treat them. But I am not blameworthy here because this judgment is accurate; treating plants this way *is* unobjectionable. Whatever objections there may be to uprooting a plant are not grounded in the fact that plants are, in themselves, the sort of things that can be treated objectionably.

Suppose, however, that I am wrong about this, and that plants can suffer, and that they have a perspective from which to raise objections to my treatment of them. In the context of criticizing views like the one defended in this paper, Neil Levy proposes just this case:

> Suppose that there is a kind of harm that is objectively morally relevant, but of which we are ignorant. Suppose, for instance, that plants can be harmed, and that this harm is a moral reason against killing or treading on them. In that case, many of us are causally responsible for a great many moral harms. Are we morally responsible for them? Do we flout a moral requirement, and challenge plants' standing as objects to which some moral consideration is owed? No to all these questions: If we do not grasp the moral requirement, and this ignorance is not culpable, we do nothing blameworthy. (2005: 9)[19]

Now I agree that we are not blameworthy for stepping on plants in Levy's example, but the example is offered as an instance of exculpation by way of nonculpable *moral* ignorance and this is not what it shows. Levy's example fails to illustrate the exculpatory power of moral ignorance because what explains why we are not blameworthy in the example is that we lack crucial information about how walking on plants affects them. Since I do not know that plants can be harmed, my stepping on one does not express a denial of the significance of its being harmed. So, on the view I advocate, stepping on a plant does not express a judgment that could properly ground blame.

The problem is that Levy's example does not involve the relevant sort of moral ignorance. There is, of course, a kind of moral ignorance in the example: we lack knowledge of the moral status of stepping on plants (because we are ignorant of their capacity for suffering). But this moral ignorance derives from circumstantial ignorance and not from ignorance of a moral principle. In the example, I presumably know that I generally have moral reasons to not cause pain; I just don't know that stepping on plants causes them pain. However, the moral ignorance of someone like Mr Potter is just the reverse: he knows that his actions will cause George Bailey and others pain, but he does not know (because he does not believe) that this counts decisively against performing these actions.

For Levy's example to help us see that Potter's sort of moral ignorance counts as an excuse, the example would have to feature Potter's sort of ignorance. The example would need to be one in which I know, say, that stepping on a plant will cause it pain and I (nonculpably, but wrongly) believe that this pain does not matter. Now if we think that I am blameless in this revised version of Levy's example, we might draw a conclusion about the exculpatory significance of nonculpable ignorance of moral principles. Of course, I would not agree that I am blameless in this revised case. If

[19] I should note that Levy is considering psychopaths (rather than relatively normal wrongdoers like Mr Potter) when he offers this example, but I don't think this affects its usefulness here. David Shoemaker (2011) treats a similar example; I reply to Shoemaker in (2012b).

I knowingly (and unjustifiably) cause a plant severe pain, then I am open to resentment and indignation because my action dismisses the normative significance of the plant's pain, and this is something to which the plant (or at least one who is concerned for its welfare) could reasonably object.

CONCLUSION

I have argued against the assumption that morally ignorant wrongdoers are open to moral blame only if they are culpable for their ignorance. Thus, I reject skepticism about moral responsibility that depends on this assumption. On the view I have defended, the attitudes involved in moral blame are responses to the features of an action that make it objectionable from the perspective of the one who issues the blame. One important way that an action can appear objectionable to us is that it expresses a judgment with which we disagree about the significance of the needs and interests of those affected by the action. Whether a wrongdoer's action has this feature depends more on whether she is aware of the consequences of her behavior than on whether she regards her behavior as wrong.

REFERENCES

FitzPatrick, William J. (2008). "Moral Responsibility and Normative Ignorance: Answering a New Skeptical Challenge." *Ethics* 118: 589–613.

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

Frankfurt, Harry (1969). "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66: 829–39.

Levy, Neil (2005). "The Good, the Bad, and the Blameworthy." *Journal of Ethics and Social Philosophy* 1: 1–16.

—— (2009). "Culpable Ignorance and Moral Responsibility: A Reply to FitzPatrick." *Ethics* 119: 729–41.

—— (2011). *Hard Luck: How Luck Undermines Free Will and Moral Responsibility.* (New York: Oxford University Press).

Rosen, Gideon (2003). "Culpability and Ignorance." *Proceedings of the Aristotelian Society* 103: 61–84.

—— (2004). "Skepticism about Moral Responsibility." *Philosophical Perspectives* 18: 295–313.

—— (2008). "Kleinbart the Oblivious and Other Tales of Ignorance and Responsibility." *The Journal of Philosophy* 105: 591–610.

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

—— (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. (Cambridge, MA: Harvard University Press).

Shoemaker, David (2011). "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility." *Ethics* 121: 602–32.

Smith, Angela (2005). "Responsibility for Attitudes: Activity and Passivity in Mental Life." *Ethics* 115: 236–71.

—— (2008). "Control, Responsibility, and Moral Assessment." *Philosophical Studies* 138: 367–92.

Smith, Holly (1983). "Culpable Ignorance." *The Philosophical Review* 92: 543–71.

Talbert, Matthew (2008). "Blame and Responsiveness to Moral Reasons: Are Psychopaths Blameworthy?" *Pacific Philosophical Quarterly* 89: 516–35.

—— (2012a). "Moral Competence, Moral Blame, and Protest." *The Journal of Ethics* 16: 89–109.

——(2012b). "Aliens, Accountability, and Psychopaths: A Reply to Shoemaker." *Ethics* 122: 562–74.

Watson, Gary (2004). "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme." In *Agency and Answerability*. (New York: Oxford University Press), 219–59.

Weingarten, Gene (2008). "Fatal Distraction." *The Washington Post*, March 8, W8.

Williams, Bernard (1995). "Internal Reasons and the Obscurity of Blame." In *Making Sense of Humanity*, 35–45. (Cambridge: Cambridge University Press (1995)), 35–45.

Zimmerman, Michael (1997). "Moral Responsibility and Ignorance." *Ethics* 107: 410–26.

—— (2008). *Living with Uncertainty*. (Cambridge: Cambridge University Press).

10
Partial Desert

Tamler Sommers

1. INTRODUCTION

Moral luck occurs when agents are morally evaluated for things that are beyond their control (Williams, 1981; Nagel, 1979). A particularly clear kind of moral luck is "resultant luck," or luck in the way things turn out. Attempted murderers are judged less harshly than successful ones though both intended to kill their victims. Drunk drivers who have an accident resulting in the death of a pedestrian may be convicted for manslaughter or worse; drunk drivers who are caught without incident get nothing more than a suspended driver's license or compulsory participation in an online seminar. Some philosophers, most notably Aristotle, accepted moral luck as a fact of life. Contemporary philosophers, however, tend to regard moral luck as a serious problem or even a paradox because they employ a Kantian concept of moral desert according to which agents can be justly blamed or praised only for aspects of actions that are within their control.

The ongoing controversy over victim impact statements (VIS) shows that this conception of moral desert is pervasive within criminal justice systems as well, at least in the West. VIS are statements that express the grief and suffering of the victims of crimes as well as their views on what would be a suitable punishment for the offender. In the well known 1987 Supreme Court case Booth v. Maryland, the Justices overturned a lower court's use of a VIS in a capital crime. Writing for the majority decision, Justice Lewis Powell explains:

> The focus of a VIS . . . is not on the defendant, but on the character and reputation of the victim and the effect on his family. *These factors may be wholly unrelated to the blameworthiness of a particular defendant.* As our cases have shown, the defendant often will not know the victim, and therefore will have no knowledge about the existence or characteristics of the victim's family. Moreover, defendants rarely select their victims based on whether the murder will have an effect on anyone other than the person murdered. *Allowing the jury to rely on a VIS therefore could result in*

> *imposing the death sentence because of factors about which the defendant was unaware, and that were irrelevant to the decision to kill.* This evidence thus could divert the jury's attention away from the defendant's background and record, and the circumstances of the crime. ((Booth v. Maryland, 1987, p. 253; my italics)

The worry about VIS noted in the majority decision is precisely that it introduces an unacceptable degree of resultant luck in the sentencing process. According to Powell, punishment should be tied exclusively to the "personal responsibility and moral guilt" of the offender. The VIS contains information that the offender could not possibly have known or foreseen at the time of the crime. The amount of grief a particular family will feel, how vindictive or forgiving their natures happen to be—all of this is beyond the control of the offender. It has nothing to do with the offender's "decision to kill" and, according to Powell, is therefore unrelated to his moral guilt and personal responsibility.

Though the courts may try to minimize its effects in certain cases, it is clear that moral luck pervades our legal system and everyday lives. It seems right that a drunk driver who accidentally kills a pedestrian deserves more punishment than one who made it home safely, even if it is perhaps unfair for the difference to be so extreme. How can we account for our intuitions in such cases? One alternative would be to claim that although their intentions were the same, the drivers performed two different actions. The first driver is judged for the act of killing a pedestrian while driving drunk, the second for a normal DUI. Another would be to claim that the two drivers are equally deserving of blame, but that judgments about the proper punishment must take harm into account.[1] Finally, one may claim that the drivers deserve the same amount of punishment, but that for various consequentialist reasons we have to punish the second driver more harshly.

In this chapter, I defend a fourth alternative. I argue that we should not understand desert as impartial or "blind," connected only to the personal culpability of the agent. Rather, we should instead adopt a "partial" account according to which desert judgments are properly sensitive to the feelings, desires, and behavior of those most closely affected by the wrongdoing. Section 2 outlines in a little more detail the conception of moral desert that I wish to challenge. Sections 3 and 4 present several cases and variations that appear to undermine the impartial view and offers a new account of desert that can better account for our judgments in the cases. Section 5 introduces a

[1] This alternative may just push the problem back a step, however, since it does not explain why it is fair to punish wrongdoers for aspects of their behavior that are beyond their control.

relevant distinction in penal philosophy about the relationship between desert and proportionality. The final sections defend the partial account against common objections and offer reasons to prefer to it alternative accounts.

2. THE ACCEPTED FRAMEWORK FOR DESERT JUDGMENTS

Theories of moral desert—both compatibilist and incompatibilist—devote most of their attention to identifying general conditions or criteria that have to be met in order for agents to be blameworthy for their behavior.[2] The conditions differ depending on the theory, but they all focus exclusively on facts about the agent. This is true for both compatibilist conditions (e.g. reasons-responsiveness, attributability) and incompatibilist conditions (e.g. ultimate responsibility or the ability to do otherwise).[3] Only agents who meet these conditions are eligible to deserve blame and/or punishment for their actions.

Skeptics about desert can stop here since the conditions for moral responsibility in their accounts cannot be met. Nonskeptical theories, however, must also offer a way to determine how *much* blame or punishment an agent deserves. Although this aspect of the debate does not receive much attention, the formula seems to go as follows. Agent-centered facts determine if the agent is morally responsible and perhaps to what degree.[4] This judgment is then coupled with judgment about the gravity of the offense to determine the amount of blame or punishment the agent deserves. For the purposes of this paper, I will refer to judgments that combine (1) the severity of the wrongdoing and (2) the agent's moral responsibility for performing it as judgments about the agent's *personal culpability*. The accepted framework—or what I will sometimes refer to as the impartial conception of desert—regards personal culpability as determinative of how much blame or punishment the agent deserves.

For some, these remarks may seem so obvious as to be hardly worth mentioning. We are, after all, assessing what the agent deserves, so of course we look to facts about agents and their actions. And certainly, when

[2] This chapter focuses only on desert for morally wrong or bad actions.

[3] See e.g. Fischer and Ravizza (1998), Kane (1996), van Inwagen (1983). Exceptions to this rule, depending on one's interpretation, may include Strawson (1962), Wallace (1994), and Scanlon (2008).

[4] It is surprisingly difficult to find discussions of the *degrees* of moral responsibility in the philosophical literature. Perhaps this is because theories of moral responsibility tend to be framed in terms of necessary and sufficient conditions.

considered in abstract terms this approach to desert is intuitively compelling. It is also consistent with certain intuitions about fairness and our desire to be master of our own fate. When applied to particular cases, however, the framework can produce counterintuitive results and may therefore require revision. The modification I propose accepts that facts about personal culpability are *necessary* for making desert judgments, but denies that they are the whole story. Desert judgments, I argue, must in addition consider certain facts that are independent of the agent and the action. I refer to this as the "partial conception" of desert because it takes into account facts about particular individual victims—their behavior, desires, and attitudes—all of which can be beyond the offender's control.[5] In my revised framework, the agent's personal culpability sets a *spectrum* for how much blame and punishment can be deserved. But additional facts are required to make more precise determinations within that spectrum. On my account, then, agents who are equally culpable may deserve different amounts of blame and punishment depending on these facts.

3. PARTIAL DESERT

The following two cases offer some support for my proposal. They begin the same way:

> John is a 33-year-old graduate student at the University of Utah. He goes to a football game and gets very drunk. He plans to leave his car at the stadium and get a ride home from a friend, but there is miscommunication and his friend leaves without him. In general, John is morally opposed to drunk driving and almost never does. But it is almost impossible to get a cab, so John reluctantly drives home in his intoxicated state. Just before he reaches his house, he has an accident causing him to swerve into a driveway where a young girl was playing. The girl is killed instantly. Panicked, still drunk, not thinking clearly, he leaves the scene and goes home. As he sobers up, he is overcome with remorse. He considers turning himself in but is terrified of going to jail and decides against it.

Now the story splits into two directions.

First Scenario

> The police track down John, arrest him, and put him on trial. Perhaps because the death involved a child, as well as a hit and run, the DA

[5] As should be clear, I use "partial" here to contrast with "impartial" rather than with "wholly" or "fully."

> manages to convict John for homicide and the judge sentences him to the death penalty. Since this takes place in Utah, John dies at the hands of a firing squad.

Judgment: Most, I imagine, would call this an unjust verdict. The killing, after all, was completely unintentional. John showed no ill will whatsoever towards his victim or anyone else. True, he made the decision to drive drunk but there were many people in far worse condition than John who drove home from the game and were lucky enough to avoid this tragedy. It is truly a case of terrible moral luck that his accident resulted in the death of a child. To be sure, John deserves a harsh sentence for his crime, John himself would likely agree with that. But few would say that he deserves to *die* for it.

Second Scenario

> The police are unable to discover who caused the accident. The parents of the child are grief stricken, completely distraught. Their daughter meant everything to them. Since the police are overtaxed and the case has gone cold, they vow to find the culprit themselves. They cash out their retirement funds, sell their house, and hire the best private investigators. Eventually they discover that it was John who caused the death of their daughter. The father goes to John's house, taking his gun. When he sees John, he is overwhelmed with anger and grief. The image of his daughter playing in their driveway flashes through his head. He takes out his gun and shoots John in the heart, killing him.

Judgment: In this scenario, by contrast, it seems far more plausible that John gets what he deserves. At the very least, John seems significantly more deserving of his fate in the second scenario than the first. (If "John gets what he deserves" were on a Likert scale, I imagine people would be much closer to "agree" in the second scenario than the first.) Furthermore, *John himself* would likely feel the same way—I certainly would in his shoes. Facing the firing squad, John might be furious at the injustice and the unlawfulness of the verdict. But looking down the barrel of the father's gun, he may think: "I ended the life of this man's child. If he wants to shoot me, that's his right. I have this coming, it's what I deserve."

I should emphasize we are not evaluating the morality of the father's action, but rather whether John receives what he deserves. These are separate matters. To take a grisly example, imagine that a gang of rapists coincidentally choose as their victim a man who is a serial rapist himself. We might say that gang acted immorally but nevertheless that the serial rapist got precisely what he deserved. My claim, then, is not that the father acted rightly in shooting John. It is that John seems to deserve his fate (being shot in the heart) more when it is the father, rather than State, who carries it out. Yet the accepted framework cannot account for this judgment. In both cases, John's personal culpability and his punishment (being shot in the heart) are identical. Our judgments about what John deserves, then, do not seem to be based entirely on facts about the wrongdoing and John's responsibility for performing it.[6]

For readers who lack my intuitions about these cases, my argument will likely not have much force—at least not yet. Those who share my intuitions but still wish to preserve the impartial conception of desert must explain why we come to different judgments in the two cases. One might appeal to consequentialist considerations, but desert—as a backwards-looking concept—is essentially nonconsequentialist in nature; it would be surprising if the difference in intuitions were sensitive to such factors. More importantly, it is not clear that the consequences are better in the second case than in the first. Indeed, they may be worse, since the father will likely be imprisoned himself, causing even more suffering for himself and his wife.

One might object that our intuitions in the first case are responding to the *legal injustice* of the verdict. After all, involuntary vehicular manslaughter is not a capital crime. The DA would probably have to fudge the evidence or mislead the jury to get the conviction. Perhaps we are feeling more lenient towards John in the first case because of the legally unfounded conviction for homicide. I agree that there seems to be a legal injustice in the first case, but I do not think it can account for the difference in intuitions. After all, the law is not being respected in the second scenario either. Federal or State law does not allow for parents of victims to take the law into their own hands. And if it is unjust in principle to issue a capital sentence when there is no *mens rea* on the criminal's part, then it should be equally unjust for the father to carry out the killing himself.

A more promising strategy might appeal to our natural sympathy for the father. We may feel that the father's actions were understandable in a way that the State's was not. We may even believe that the father deserves his vengeance, which then affects our judgment about what *John* deserves. I agree that our judgments may be sensitive to our sympathy for the father, but it is not clear that this is a distorting influence rather than an appropriate one. My claim is that our desert judgments *should* be sensitive to our sympathy for the particular victims of wrongdoing. Again, imagine the case

[6] We may also imagine an analogous set of cases in which John receives what seems like too lenient a sentence. In the first case, the judge gives him probation and no jail time. In the second, the father finds John, sees that he feels tremendous guilt, that he is horrified by what happened, and decides not to turn him in to the police. Again, I would suggest that John seems more deserving of the lighter sentence in the second case than the first.

from John's perspective. If I were John, I would feel enormous sympathy for the child's father, and my sympathy might lead me to think it is in large part up to him as an individual to determine what I deserve. Unless we are already committed to the impartial conception, I see no reason why we should regard this sympathy as a distortion of John's judgment. One might try to turn this reply into another objection to the partial conception. Perhaps the difference in judgments can be traced to what happens when we take the agent's perspective. But again, there is no reason to think that this is a distorting influence—why shouldn't we take the agent's perspective into account? The reply: "Because the agents' subjective perspective is irrelevant to objective judgments about what they deserve" begs the question. It is true that agents are likely to be biased in their own favor, or at times feel excessive unwarranted guilt. But this just means we should be careful about how we interpret the agent's perspective, not that we should ignore it entirely.[7]

4. THE COMPLEXITY OF DESERT

Two more cases may help to illustrate the relevance of the victims' feelings to the offender's deservingness. Both are variations of the second case in which the father shoots John. The variations focus on the father's feelings after the shooting.[8]

Third Scenario

> The father recognizes that he acted in a moment of blind rage and despair, and regrets the shooting immediately. He calls for an ambulance but it arrives too late, John is dead. Although there is a small glimmer of satisfaction that John will not get away unpunished, on balance he feels

[7] As an anonymous referee notes, this case has an additional complication, namely that the victim is dead and her wishes regarding the punishment are unknown. (Or she may be too young to have well-considered feelings about John's punishment.) This raises the question of how desert might be affected if her parents or close relatives had different wishes regarding the punishment—for example, if the mother felt more retributive and the father more merciful. I agree that this is a difficult and important question, one that my "partial" view of desert must address. For the purposes of this more programmatic paper, however, it's enough to point out that such factors (agreement or disagreement among the relatives or those closely connected to the victim) actually matter, even if we cannot yet specify how much. On the impartial conception of desert, these factors would be irrelevant.

[8] These variations are inspired by Chandra Sripada and his comments on the Flickers of Freedom blog. I am grateful for his contributions as well as many others on that post. See: <http://agencyandresponsibility.typepad.com/flickers-of-freedom/2010/08/can-there-be-partial-as-opposed-to-impartial-desert.html>.

> worse than before. John did not mean to hurt his daughter, people drive drunk all the time. As the father looks down at John's lifeless body, the senselessness of his revenge seems tangible. The only thing he seems to have accomplished is the waste of another life. The father wishes desperately that he had simply turned John into the police.

Judgment. Before learning about the father's feelings, it seemed that John deserved his fate (or at least that he was more deserving than in the first scenario). But the father's regret seems to undermine this intuition. Now my reaction resembles when John was executed by the State—the punishment seems excessive and undeserved.

Scenario 4

> The father recognizes that he acted in a moment of blind rage. Still, upon reflection, he feels that justice was done. He recognizes that this act will not bring his daughter back, and that nothing will alleviate the suffering he feels in her absence. But at least he has paid his debt to her and did not allow the person who killed her to get away with it. He feels a strange sense of peace, although his grief is just as acute. The father calls 911 right away and confesses to the crime. He accepts responsibility for his action, and waits for the police to come and arrest him.

Judgment: Now my initial intuitions that John deserved his fate are, if anything, even stronger than in the second case. The father has performed the act, owns up to it, and even feels a small degree of satisfaction. He has risked and sacrificed a great deal to bring about the punishment. He accepts responsibility and will now go to prison. Perhaps if we were in the father's place, we would not feel or act this way. But there is a significant sense in which it is not up to us, because we did not suffer from the offense. Again, if I can inhabit John's perspective (now from beyond the grave), I would accept that I had received what was coming to me.

It may seem that I am edging (or hurtling) towards a *reductio* of my own position. Judgments of *John's* deservingness are supposed to be sensitive to the father's feelings about his act of revenge after performing it? To how much the father risked and sacrificed to make it happen? To his willingness to accept responsibility and punishment? In the remainder of this chapter I hope to minimize the incredulity that accompanies such questions and argue that the answer to all of them is a simple "yes."

5. CARDINAL AND ORDINAL PROPORTIONALITY

I mentioned earlier that when considered abstractly, there is a good deal of intuitive plausibility to the impartial conception of desert. This is due in large part to a long-standing (but insufficiently analyzed) idea that the punishment should fit the crime. The proportionality principle is a hallmark of retributive or "just-desert" theories of criminal justice, and indeed one of the main objections to rival utilitarian theories is that it would allow for disproportionate punishments. The criminologist Andrew von Hirsch calls the principle a "basic requirement of fairness" and describes it as follows:

> (1) [T]he principle of proportionality concerns how much punishment one deserves; (2) deserved punishment should be commensurate to the degree of blameworthiness of the conduct; and (3) blameworthiness depends both on the harmfulness of the conduct and on the degree of culpability of the actor blameworthiness depends both on the harmfulness of the conduct and on the degree of culpability of the actor." (Von Hirsch 1978: 622)

But the proportionality principle has some well-known difficulties as well. The primary problem concerns our inability to determine what kind of punishment is commensurate with a given crime. What is the deserved punishment for armed robbery? Ten years in prison? Fifteen years? Two years and probation? Flogging? As Von Hirsch recognizes, desert theorists have been notoriously unsuccessful at offering principled answers to these questions. Von Hirsch presents the problem in the form of a dilemma. If we assume that a particular crime warrants a specific quantum of punishment, then we must presuppose "a heroic kind of intuitionism: that if one only reflects enough, one will 'see' the deserved quanta of punishment for various crimes." (Von Hirsch 1992: 76) Unfortunately, no one seems to have such illuminating intuitions and it is implausible to think that more moral reflection will remedy this. The other option is to employ a "range-only" view of desert according to which a criminal's personal culpability determines only the upper and lower limits of deserved punishment—a view defended by Norval Morris (Morris 1982). Morris's view allows for a wide range of deserved punishments for a particular crime. It is only when punishments fall outside this range that our intuitions give us a clear sense that the punishment is *undeserved.* A heroic form of intutionism, then, is not required for the range-only view.

According to Von Hirsch, however, Morris's account is open to what he calls "a fundamental objection": it would allow two offenders who are equally culpable to receive different punishments. And this is just the sort of unfair outcome the proportionality principle is supposed to rule out. Von Hirsch's proposed solution to this dilemma employs a distinction between *cardinal* and *ordinal* proportionality. Cardinal proportionality is absolute: it "anchors" the severity of the punishment to the culpability of the criminal (which includes the harmfulness of the crime). Cardinal proportionality must remain "range only," issuing upper and lower limits where punishments would obviously be either too severe or too light. Ordinal proportionality is relative. It has two aspects. The first is parity: like crimes must be treated alike. If two criminals are equally culpable then they should receive the same punishment. Second, punishments must be proportionate relative to one another. If one crime is twice as serious as another, the punishment should be twice as serious as well. The leeway that cardinal proportionality allows in deciding the anchoring points of the scale explains why we cannot perceive a single right or fitting penalty for a particular criminal. Once the anchoring points of the scale have been fixed, however, the more restrictive requirements of ordinal proportionality begin to apply.

In practical terms, the idea would be roughly as follows. We do not know precisely what the punishment should be for, say, car theft. There are a range of punishments that might be proportionate for this crime and in absolute terms, proportionality just requires that we stay within this range. We do know, however, that car theft is a less serious crime than armed robbery. So the proportionality principle requires that (a) two equally culpable car thieves receive the same punishment, and (b) armed robbers receive a more severe punishment than car thieves. Von Hirsch offers university grading practices as an analogy. The standards for "A" papers and "B" papers and so forth are real but indeterminate (cardinal proportionality) and may depend on nonmerit based factors about the university. But once those standards are set, fairness requires that we give papers of equal merit the same grade (ordinal proportionality).

The initial intuitive resistance to the idea of partial desert is rooted in our commitment to ordinal proportionality. But the depth of this commitment is open to question. In fact, outside of the context of criminal justice, it's not clear that we are committed to ordinal proportionality at all. Imagine that a woman decides to leave her philandering husband and he replies: "I understand you're angry, but fairness requires that you don't leave me. Bill's wife stayed with him and he's had several more affairs than I have." Would the wife be moved by this consideration? Should she be? Everyday life is filled with cases like this—acts of infidelity, betrayals of trust, insulting or offensive remarks, and many others. We do not imagine that there is a correct response or punishment, one that is tied only to the agent's personal culpability. Nor do we cry foul when people who are equally culpable do not receive the same amount of blame or punishment. We leave it up to the relevant parties to determine the right response, within certain boundaries.

What does survive outside the context of criminal justice is our commitment to *cardinal* proportionality. We maintain that there is a range of appropriate blame or punishment responses and that responses outside of this range would be undeserved. Whether the betrayed spouse asks for a trial separation, files for divorce, gives the partner another chance is largely up to her. All of these are proportionate responses. But imprisoning the spouse or killing him or even cutting off all access to the children would be disproportionate. The husband (as well as third parties) might legitimately complain that the treatment is undeserved. It is significant that von Hirsch employs the practice of essay grading to illustrate the importance of ordinal proportionality. Certainly, it is a desert-based practice, but when a student writes a bad paper, there is no victim.[9] Offenses or crimes, by contrast, have identifiable victims who have suffered at the hands of the offender. The presence of victims is a morally relevant factor that affects our understanding of proportionality and desert. Exactly how is the topic of the next section.

6. THE DESERT SPECTRUM

On my account, whenever a desert-based practice involves victims as well as agents, personal culpability can only set a spectrum for how much blame and punishment the offender deserves. For more precise determinations, we must take facts about particular victims into account. The impartial conception regards the feelings and desires of the offended parties to be irrelevant to desert. By contrast, I see them as *essential* for determining an appropriate response within the spectrum of deserved responses. How narrow or broad is the desert spectrum set by personal culpability? This is a tough question. It seems to vary depending on the kind of case that we are judging. My hunch is that the spectrum is broadest in cases where the offense is accidental or the result of negligence. This is why there is such a wide range of deserved outcomes in the John cases. At one end of the spectrum is the outcome of the second case—John being shot in the heart. The other end might include outcomes that allow John to go free. Imagine that the parents track John down and confront him. They express their anger and grief and sense of loss. They see that John is consumed with remorse and has been since the accident. The meeting gives the parents a sense of peace and closure. They see that John would not do well in prison, and they decide, after some tortured reflection, not to turn him into the police. Many might call this a just outcome. Others might disagree and claim that the parents were admirable in showing mercy, but that John was clearly getting less than he deserved. There is room for reasonable disagreement on this question. But compare this outcome to one in which John,

[9] Aside from the instructors who have to comment on them.

through a plea bargaining agreement, is offered a suspended sentence—in spite of the protests of the child's parents. It seems undeniable that John is more deserving of his freedom when the parents bestow it on him. I am not claiming that the parents' wishes and attitudes determine John's deservingness entirely. John could not deserve a month long cruise to the Galapagos Islands no matter what the parents' wanted. Nor could he deserve to be tortured for twenty consecutive years. Again, we may reasonably disagree about the end points of the spectrum. But the accidental nature of the crime does seem to yield a strikingly broad range of deserved punishments. By contrast, if John had deliberately killed the young child in cold blood, the range might be significantly compressed.

7. PHILOSOPHICAL BUSYBODIES

The philosophical temperament may rebel against the looseness of this account. To some, it will appear arbitrary, irrational, unsystematic, and perhaps even antithetical to the project of providing a principled basis for desert assignments. Certainly, this has been the reaction of the legal academy to the rapidly growing victim's right movement that involves the victims in the sentencing process. Yet as I have noted, the demand for impartiality and rational consistency is completely at odds with our everyday practices, where victim involvement is expected and welcomed. Indeed, the partial view is probably most intuitive when punishment is not at issue, and the question concerns how much blame to assign to the offender. Should Sarah blame her sister Emma for not remembering her birthday because she was stressed about her job? That is up to Sarah—no theory should tell her the right or rational response. Certainly, impartial considerations should play some role in desert assignments, especially in more severe cases of wrongdoing. But the partial account allows for this by maintaining the commitment to cardinal or range-only proportionality. Why should we aspire to more precision than this? The common assumption that theories should dictate to people exactly how much they should blame the person who wronged them deserves scrutiny.

The increasing popularity of restorative or restitutionary movements in penal philosophy is relevant here. These movements have emerged out of the increasing dissatisfaction with the depersonalized, process-oriented, excessively rationalistic nature of our current criminal justice system.[10]

[10] See e.g. Barnett (1977), Strang and Braithwaite (2000), Van Ness (1993), and Zedner (1994).

According to Lucia Zedner, the system "has transformed the drama and emotion of social interaction and strife into technical categories which can be subjected to the ordering practices of the criminal process." (Zedner 1994: 231) Proponents of restorative justice argue that the blend of retributive (desert-based) and utilitarian principles of our current system is unjustly one-sided in its focus on the criminal. The victim is just a faceless vessel for wrongdoing—like a poor essay that justifies a low grade. This has the effect of alienating the victims and diminishing their self-respect even further.

The criminologist Nils Christie has famously accused the criminal justice system of "stealing conflicts" from their rightful owners. Lawyers, he claims, are particular good at this form of larceny:

> Lawyers are . . . trained into agreement on what is relevant in a case. But that means a trained incapacity in letting the parties decide what they think is relevant. If the offender is well educated, ought he then to suffer more, or maybe less, for his sins? Or if he is black, or if he is young, or if the other party is an insurance company, or if his wife has just left him, or if his factory will break down if he has to go to jail, or if his daughter will lose her fiancé, or if he was drunk, or if he was sad, or if he was mad? There is no end to it. And maybe there ought to be none. (Christie 1977:.8)

This extraordinary passage should trouble more than just lawyers. For if we replace "lawyers" with "philosophers" or "desert theorists" in Christie's remarks, we expose some of the dubious assumptions and aspirations in our current approach to moral desert. Our ever-more refined accounts of blame and punishment—accounts that are supposed to apply across the board, no matter what the relevant parties might think—may have the effect of stealing conflicts from particular individuals. This whole approach exhibits a "trained incapacity" to let individuals decide which factors are relevant and how much. Zedner's accusation seems apt as well. By focusing entirely on impartial conditions of criminal culpability, our theories transform the drama and emotion of social interaction and strife into technical (often metaphysical) categories which can be subjected to a systematic ordering process of desert attribution.[11]

Let me conclude this section by describing a real criminal case that occurred recently in Grand Junction, Colorado, one that resembles the hypothetical example I introduced earlier. A woman was driving in the early morning, drunk and high on methamphetamines. On the way, she had an accident and hit a sanitation worker. The worker's legs were

[11] One might interpret Strawson (1962) as making a similar point about the moral responsibility skeptic who believes that "blame is metaphysical." "The metaphysics," Strawson writes, "is in the eye of the metaphysician." (p. 24).

shattered, causing him to endure ten surgeries. During her trial, the worker testified and asked the judge to give the woman a lenient sentence. He explained that he could relate to her predicament, that he had been in a dark place once, and he hoped she could eventually get to a better place. The prosecutors and judge took his desires into account, dismissed the most serious charges, and the woman received the minimum sentence allowable—still over five years in jail.

Here we have a clear violation of ordinal proportionality. Many people who are equally culpable in Colorado have been given much higher sentences and many will again. But did the judge and prosecutor violate a "basic requirement of fairness"? Is this case clearly unjust? The victim of the offense is satisfied, more than he would be if his wishes had not been considered. The woman is still receiving a punishment within a reasonable cardinal range. Is it our business as philosophers to complain about the verdict, those of us who have not suffered in any way from this crime? Doing so, in my view, would make us "philosophical busybodies" sticking our collective nose where it doesn't belong.

8. CONCLUSION

For all the insights it provides, there is a crucial difference between the restorative justice critique and my own. Proponents of restorative justice regard it as an alternative to retributivist or desert-based approaches to criminal justice. In my view, we should *incorporate* facts about individual victims into the way we understand desert for wrongdoing, criminal and noncriminal. My account, then, is not an alternative but rather a revised version of desert theory or retributivism. In order to make a judgment about what John deserves in our original case, we have to know more about what John and the parents want and believe. For partial desert judgments, it matters whether the parents are vindictive or forgiving. And it matters what John himself feels when he looks into their eyes.

Defenders of our current criminal justice system like to think that as an enlightened society, we have transcended the revenge feelings and practices of our barbarous past and replaced them with "justice," which is rational and not subject to emotional bias. But the retributivist project in the West has struggled to develop a coherent notion of "just-deserts" that does not appeal in any way to our natural disposition for vengeance. In my view, the fears of allowing emotions into the equation are way overblown. No one is advocating for a return to the days of endless tribal warfare. There is a middle ground, one that allows individual victims to *influence* our desert judgments under certain defined parameters, but not to determine them.

These remarks lead to what may be the strongest objection to my argument. One might claim that I am conflating two distinct concepts: (1) what the offender deserves and (2) the just outcome of the crime. A critic might concede that to determine the just outcome, we must take other factors besides the deservingness of the offender into account—the costs of the punishment and perhaps even the victims' interests and facts about what *they* deserve.[12] Our judgments about the John cases, then, reflect our intuitions about the just outcome of the crime rather than intuitions about what John deserves.[13]

In response, let me first distinguish consequentialist factors from other justice-related factors independent of the just-deserts of the offender. I agree that we can distinguish desert judgments from "all-things-considered" judgments about blame and punishment that take consequentialist considerations into account. Recall, however, that the different desert judgments in the John cases could not be traced to such considerations. I agree as well that judgments about John's *moral responsibility* for the offense are distinct from judgments about what John deserves for having performed it. Moral responsibility on this account is constituted by agent-centered factors in a way that desert is not.[14] The objection, then, must be that we need to distinguish John's deservingness from other justice-related judgments about his case. Interpreted in this manner, however, the objection just begs the question by assuming a conception of desert that is tied only to the personal culpability of the offender. If we do not employ this conception from the outset, then there is no reason to think the justice-related judgments are distinct.

The partial conception has some significant advantages as well. Since it lacks the commitment to generality and objective precision, it is less vulnerable to the endless array of counterexamples and theoretical difficulties that have plagued desert theory to date. The partial conception may also offer a new way of addressing the "paradox" of moral luck that I described at the outset of this chapter. The reason the drunk driver who

[12] In penal philosophy, this type of account is known as the "hybrid view" developed by Paul Robinson (1987) among others.

[13] Once again, this objection is inspired by comments on my Flickers of Freedom post. A related objection, raised by an anonymous referee, is that I am conflating the notions of what John deserves and those of what "serves him right." In response, let me first say that I'm not sure there is a substantive difference between the notions of desert and those like "serves him right" and "had it coming to him"—although I recognize that many or most philosophers will disagree with me on this point. Second, even if there is an important difference between those notions, it's not clear that the distinctions explain our different intuitions regarding desert in the John cases.

[14] Thanks to Tim Scanlon and Sarah Buss for convincing me on this point.

has an accident involving a pedestrian deserves more blame and punishment than the drunk driver who makes it home without incident is that there are victims in the former case. Although the drivers had the same degree of control over the offense, desert judgments must take the harm and interests of victims into account as well. Since there are no victims for the second driver, he deserves significantly less blame and punishment. One can apply a similar strategy to other kinds of moral luck, including the most pervasive—constitutive luck. If our method for settling on a desert concept is reflective equilibrium, the ability of the partial account to address a problem as ancient and intractable as this one should count as a considerable virtue in its favor.[15]

REFERENCES

Barnett, Randy E. (1977). "Restitution: A New Paradigm of Criminal Justice." *Ethics* 87.4 : 279.

Christie, Nils (1977). "Conflict as Property." *British Journal of Criminology* 17.1: 1–15.

Fischer, John Martin, and Mark Ravizza (1998). *Responsibility and Control: A Theory of Moral Responsibility.* (Cambridge: Cambridge University Press).

Kane, Robert (1996). *The Significance of Free Will.* (New York: Oxford University Press).

Morris, Norval (1982). *Madness and the Criminal Law.* (Chicago: University of Chicago Press).

Nagel, Thomas (1979). *Mortal Questions.* (Cambridge: Cambridge University Press).

Robinson, Paul (1987). "Hybrid Principles for the Distribution of Criminal Sanctions." *Northwestern University Law Review* 82, 19–42.

Scanlon, Thomas (2008). *Moral Dimensions: Permissibility, Meaning, Blame.* (Cambridge, MA: Belknap, Harvard University Press).

Strang, Heather, and John Braithwaite (2000). *Restorative Justice: Philosophy to Practice.* (Aldershot,: Ashgate).

Strawson, P. F. (1962). "Freedom and Resentment." *Proceedings of the British Academy* 48 : 1–25.

Van Inwagen, Peter (1983). *An Essay on Free Will.* (Oxford: Clarendon Press).

Van Ness, Daniel W. (1993). "New Wine and Old Wineskins: Four Challenges of Restorative Justice." *Criminal Law Forum* 4.2: 251–76.

[15] Thanks to Sarah Buss, Dave Shoemaker, Chandra Sripada, two anonymous referees, and the commentators at the *Flickers of Freedom* blog for valuable comments. This project has benefited immensely from the discussion at the 2011 New Orleans Workshop on Agency and Responsibility organized by Dave Shoemaker.

Von Hirsch, Andrew (1978). "Proportionality and Desert: Reply to Bedau." *Journal and Philosophy* 75.11: 622–4.

—— (1992). "Proportionality in the Philosophy of Punishment." *Crime and Justice* 16: 55–98.

—— (1993). *Censure and Sanctions*. (Oxford: Clarendon Press).

Wallace, R. Jay (1994). *Responsibility and the Moral Sentiments*. (Cambridge, MA: Harvard University Press).

Williams, Bernard Arthur Owen (1981). *Moral Luck: Philosophical Papers, 1973–1980*. (Cambridge Cambridge University Press).

Zedner, Lucia (1994). "Reparation and Retribution." *Modern Law Review* 57. 2: 228–50.

11

Values, Sanity, and Responsibility*

Heidi L. Maibom

1. SANITY AND MORAL INTRANSIGENCE

According to one influential view of responsibility, due to Harry Frankfurt (2003), what is central to responsibility is a person's ability to evaluate her own will and identify with some volitions and distance herself from others. If she is free to have the will she wants, she is responsible. Whether this lucky situation is the result of someone else's will is beside the point. Someone who is free to do what she wants and free to want what she wants to want has "all the freedom it is possible to desire or to conceive" (2003: 333). The view has been criticized for not accommodating our intuitions that an insane person is neither legally nor morally responsible. For Susan Wolf, there is one more freedom required for responsibility. For us to be responsible, we must be able to:

(a) evaluate ourselves sensibly and accurately, and
(b) transform ourselves insofar as our evaluation tells us to do so. (385)

This is the freedom of sanity.[1]

Frankfurt was interested in cases like addictions and compulsions, where having the right kind of second-order volitions might absolve one from moral blame, even if one is not free to have the will one wants. We are inclined not to hold an addict responsible for taking drugs, because we think he cannot help himself. Nevertheless, though he may have little choice in the matter of wanting to take the drug of his addiction, he can at least resist. He can want to not want to take the drug. We can distinguish

* Thanks to audiences at the first New Orleans Workshop on Responsibility, and to Vanessa Davis, Annie Larivée, Melanie Mortensen, Darien Shanske, David Shoemaker, and two anonymous referees for helpful comments and suggestions.

[1] I focus on the critique of Frankfurt, though Wolf's critique affects all so-called Deep Self Views, including those of Gary Watson and Charles Taylor (Watson 1975, Taylor 1976).

between addicts on the basis of their so-called deep selves; one might be a willing addict, the other unwilling. Wolf thinks Frankfurt's view cannot accommodate our intuitions about the insane because they have the will they want to have, yet we hesitate to hold them (fully) morally responsible. The trouble is that the insane are not able to know "the difference between right and wrong" (382) and they cannot but possess "values that are unavoidably mistaken" (383).

Wolf uses the example of the son of an evil dictator to illustrate her point. JoJo's father habitually sent people to prison, torture chamber, or death on a whim. Having been under his father's educational regimen, JoJo grows up performing just the kinds of actions that his father did. He does so willingly; just like a willing addict satisfies his addiction. He has the first-order volitions that he wants to have and the freedom to act on them. Contrary to what Frankfurt claimed, this is *not* all the freedom anyone could want, Wolf argues. For it seems unavoidable that JoJo should have the values—second-order volitions—that he does, and therefore it is not right to blame him:

> These are people, we imagine, who falsely believe that the ways in which they are acting are morally acceptable, and so, we may assume, their behavior is expressive of or at least in accordance with these agents' deep selves. But their false beliefs in the moral permissibility of their actions and the false values from which these beliefs derived may have been inevitable, given the social circumstances in which they developed. If we think that the agents could not help but be mistaken about their values we do not blame them for the actions those values inspired. (382)

Wolf's point is not that *we* must be the authors of our second-order volitions. The inevitability of which she speaks does not refer to the *origin* of such values, but to their continued existence as values that characterize an agent's deep self. What is central to sanity is the ability to *transform* one's values, and thereby one's deeper self. Seriously wrongheaded values trap a subject, in a manner of speaking, by making her incapable of evaluating them realistically. Since the insane are stuck with their second-order volitions, but we are not, they are not responsible whereas we are. That JoJo espouses the values that he does shows that he is insane:

> Sanity, remember, involves the ability to know the difference between right and wrong, and a person who, even on reflection, cannot see that having someone tortured because he failed to salute you is wrong plainly lacks the requisite ability. (382)

This suggests that in order to possess the abilities required for sanity and responsibility mentioned above (a and b), we must first have:

(c) the ability cognitively and normatively to recognize and appreciate the world for what it is. (383)

At least three conditions violate (c) according to Wolf: mental illness, seriously deviant upbringings, and membership of groups that espouse significantly different moral values than ours. In essence, any person who has values that contrast sufficiently with ours, and who holds on to such values doggedly, suffers from essentially the same deficit as the insane, barring laziness, distraction, and greed. The consequence is that people with significantly different moral values are not responsible, or at least not fully responsible. This includes male chauvinists of our father's generation, Ancient Greek slave-owners, and Nazis. In effect, we should get used to the idea that those who do great evil, by our standards, are not sane and therefore not responsible (386–7).

The concerns raised here are typically associated with what is sometimes known as the epistemic condition on responsibility, i.e. the condition that an agent must be aware of what she is doing. And there certainly seems to be many ways in which an agent may not be entirely aware of what she is doing. What Wolf seems to do is to lump many such cases together under the heading "insanity." There are many reasons to think this is infelicitous. Terminology aside, there are deeper issues with characterizing people who are insane and agents who are not, but who hold morally intransigent and divergent views, as suffering from the same epistemic deficit. Insanity and what I shall call moral intransigence are psychologically, legally, and morally quite distinct. On the one hand, people who are insane typically suffer from hallucinations and delusions, but do not possess values that differ significantly from our own. On the other, in many, if not most, cases of divergent and intransigent moral values, there is an inferential route from a person's values and beliefs, or from values and beliefs accessible to her, to the recognition that what they are doing is wrong. Such cases are therefore not generally ones in which the agent is *unable* to see that what they are doing is wrong. This is presumably why we can be held responsible although we no doubt also have a number of wrongheaded values. But if we can be held morally responsible, so can people from different cultures (though not, perhaps, from *all* other cultures). Conversely, if they are not responsible, so neither are we. The way to avoid this consequence is to maintain that only those who do wrong while fully knowing it are blameworthy. This, however, deconstructs our notion of blameworthiness in an unacceptable way.

In what follows, I shall use criminal responsibility as a guide to moral responsibility. This is not because I think the former is prior to, or more important than, the latter. Rather, I think the conditions under which we hold people criminally responsible reveal some of our deep intuitions about when someone is morally responsible. This is particularly true when it comes to the question of sanity.

2. INSANITY IN LAW

The widely used *McNaughtan Rule* stipulates that in order to be insane the defendant must suffer from a mental illness which is causally implicated in the criminal action that the defendant is on trial for: (Moran 1981: 169)[2]

> To establish a defence on the ground of insanity it must be clearly proved that, at the time of committing the act, the party accused was labouring under such a defect of reason from disease of the mind, as not to know the nature and quality of the act he was doing, or if he did know it, that he did not know that what he was doing was wrong.
>
> If the accused was conscious that the act was one which he ought not to do, and if that act was at the same time contrary to the law of the land, he is punishable.

Contrary to what is sometimes thought, mental illness is *not* synonymous with insanity. A mental illness does not simply deprive a person of responsibility for her actions, but it can sometimes overwhelm her, and when it does, she cannot be held responsible for the actions performed under its influence. The law acknowledges that mental illness can affect everyone alike, and though it may affect a person profoundly, it rarely blots out her preexisting character or moral sensibilities. As such, a crime committed by a person suffering from a mental illness may not, in any interesting sense, be the result of that person's illness, but of their deficient moral character. A person can be both bad and mad.

Most of those judged to be not guilty by reason of insanity[3] suffer from paranoia.[4] The majority of successful insanity pleas in Canada (Rice and

[2] If not directly used, the Rule often serves as the basis of acts concerning the responsibility of persons judged to be insane. In Canada, the *Criminal Code* section 16 specifies that "(1) No person is criminally responsible for an act committed or an omission made while suffering from a mental disorder that rendered the person incapable of appreciating the nature and quality of the act or omission or of knowing that it was wrong." In Britain, the corresponding act concerns diminished responsibility. According to section 52 of the Coroners and Justice Act 2009 (the Act), "a person who kills or is a party to the killing of another is not to be convicted of murder if they were suffering from an abnormality of mental functioning which: (a) arose from a recognised medical condition, (b) substantially impaired their ability to do one or more of the following: understand the nature of their contact, form a rational judgement, or exercise self control, and (c) provides an explanation for [their] acts and omissions in doing or being a party to the killing. An abnormality of mental functioning provides an explanation for the conduct if it causes, or is a significant contributory factor in causing, the defendant to carry out that conduct."

[3] In Canada, "not criminally responsible on account of mental disorder" has replaced "not guilty by reason of insanity" (Criminal Code section 672.34).

[4] The DSM-IV lists three different types: paranoid personality disorder, paranoid subtype of schizophrenia, and the persecutory type of delusional disorder.

Harris 1990), New York (Steadman et al. 1983), Colorado (Jeffrey et al. 1988), and Oregon (Rogers et al. 1984) involve defendants who have been diagnosed with psychosis, mostly commonly schizophrenia (often 80 percent or higher). Psychosis is characterized by delusions, hallucinations, and disorganized thought and speech. Consequently, people deemed insane are usually disturbed individuals who have profound problems with reality generally, not just moral reality if such a thing exists. For instance, Daniel McNaughtan, who gave the name to the famous insanity rule, killed Edward Drummond because he mistook him for the British Prime Minister Robert Peel, whose secretary he was. He believed himself to be persecuted by Peel and his political party to such an extent that his life was becoming unbearable. He also thought they were planning his demise. Andrea Yates was suffering from severe postpartum psychosis and believed that, were her children to live, they would eventually face eternal damnation. To kill them was therefore an act of mercy in her mind. So she drowned her five children in a bathtub, one after the other. James Hadfield thought he was something of a new Messiah whose lawful execution or death would bring about the second coming of Christ, thereby saving humanity from much unnecessary suffering. He shot at King George III—missing by a wide margin—because he knew that attempted regicide was punishable by death.[5] Henry Maudsley (1898) tells the story of a father who believed that he was fighting a fierce snake only to find that he had killed his infant son.

In typical successful insanity pleas it is the subjects' delusions that cause the problem, not their moral compass. Yates and Hadfield clearly thought they were acting for the greater good, and at the expense of their own well-being. Had Yates and Hadfield been right about their mistaken beliefs, their actions would not have been culpable. The so-called *As-if* rule stipulates that someone who acts on beliefs about the world that are due to delusions or hallucinations has an excuse on the condition that had their beliefs been true, their action would have been morally acceptable (Reznik 1997). The complimentary *If-only* rule specifies that in the absence of mental disorder, the subject would not have performed the action in question. And, in fact, Yates or Hadfield would probably not have done what they did had they not been in the grip of strong delusional beliefs. The *As-if* and the *Only-if* rules typically characterize successful insanity pleas. In other words, it is not usually the case that a person judged to be not guilty by reason of insanity is someone whose values are foreign to us or are such that we cannot identify with them, in some sense.

[5] He could not kill himself as he believed suicide was not permitted by God.

Some, like Lawrie Reznik (1997), think that the notion of a good character is central to insanity pleas. For the defendant to have an excuse, in cases of insanity, he must be shown to have acted *out of character*, in some sense. His actions may be the result of his mental illness, for instance. Character, for Reznik, is similar to the deeper self that Frankfurt refers to.[6] We are responsible for actions performed "in character" even if we are not responsible for our character or, to be more precise, our *moral* character.[7] If, however, "a person changes from a good character to an evil one, commits an offense, then changes back again, the good character has an excuse" (227). Mental illness and brain injury can result in such character change. If the disorder is irreversible—a progressive brain disease, for instance—the individual cannot be held responsible, and if the disorder *is* reversible, we cannot blame the good character should we succeed in restoring it. Lack of prior convictions, due regard for the well-being of others, etc., are all signs of a good character.

The Yates case helps illustrate the importance of character. Yates was diagnosed with postpartum psychosis. As already mentioned, psychotic subjects appear to inhabit, at least part of the time, a different reality from the rest of us, and usually experience great difficulties functioning normally. Parents suffering from postpartum depression or psychosis often fantasize about killing their infants. Prior to the murders, Yates had been catatonic and unable to breastfeed her youngest child properly. She had a history of postpartum depression and attempted suicides, but no criminal record. Yates's husband testified to her declining condition and insisted that she had undergone a personality change. She had expressed doubts about her ability to be a good mother, but ultimately claimed to have drowned her children to save them from an eternity of suffering in Hell. Her history of relatively blameless behavior suggests that an otherwise morally upstanding citizen—to put it somewhat primly—had been overcome by mental illness.

People judged not guilty by reason of insanity are not always so benevolent, of course. Where Yates and Hadfield thought of themselves as minimizing harm by their actions, subjects of command hallucinations believe they are simply obeying orders. Following orders, however, is usually not

[6] Reznik explicitly rejects Frankfurt's notion of personhood. This, however, seems based on a misunderstanding of Frankfurt's philosophical position.

[7] A person's moral character consists in "the set of dispositions that explain his ethical beliefs, his moral sentiments, and ethical conduct." (Reznik, 223) Talking about *moral* character is required at least for legal purposes. If someone has a temporary change in character, but it does not affect their moral outlook or their moral dispositions, it cannot excuse any crime that they may commit in their altered state.

an excusing condition in civilian life when the action is illegal. Yet, people who kill because they hallucinate God commanding them to do so are sometimes deemed insane even though their actions would not have been legally or morally acceptable had their beliefs been true (at least in a secular society). Kim John, together with Francis Philip, bludgeoned a Catholic nun to death and set a priest on fire at the Cathedral of the Immaculate Conception in St Lucia. He was subsequently diagnosed with delusional disorder, paranoid type.

Kim John claims to have been under a divine injunction to target the Catholic Church, which he also believed was persecuting him. He blamed the Church for his own troubles and for "child abuse, molesting, stealing the tithes and offerings, buying off the land, poisoning the water and food, robbing and taking funds" (Larbey 2007). Unlike McNaughten, Kim did not think that his life was in danger. He did not, therefore, have the excuse that had his beliefs been true, his act would have been reasonable. Nevertheless, his delusions were judged to be of such a nature as to make him unfit for punishment (*Francis Philip & Kim John v. The Queen*). He claims to have had visions from an early age of "[s]elling of human beings, children like cargo, women raped and murdered, tortured, castrations. The falling of the wicked in flames, seeing them burn up, skeletons, ashes, smoke, fire, weeping and wailing" (Larbey 2007). He also thought he could do magic. The pervasiveness of Kim's delusions made the Appellate Court think it useless to try to assess his moral character, which, presumably, they thought obliterated by, or buried under, his mental illness.

It is notable that not all command hallucinations excuse criminal behavior. Ronald Lafferty claimed to have had a divine revelation that he was to "remove" certain individuals that stood in God's way. Conveniently, they were all people he blamed for encouraging his wife to leave him with their children. He talked his brother Dan into killing their other brother's wife Brenda and their 15-month-old daughter. Ron received the death sentence in 1985. In the much publicized 1996 retrial, his lawyer entered an insanity plea, arguing that Ron's religious ideas and "revelations" were signs of a delusional mind. The prosecution countered that irrational and strange ideas—including the idea that one receives revelations from God—are a stable of religion. But we cannot simply assume that all religious people are insane. That would render much of the population and the political establishment in large parts of the world insane.[8] Instead, witness for the

[8] The DSM-IV specifies that "culturally sanctioned" responses (p. xxxi) or political, religious, or sexual deviant behavior are not signs of mental illness unless they are symptoms of a dysfunction of the individual associated with present distress, disability,

prosecution Dr Stephen Golding maintained that Ron was a zealot, and zealots "are [not] mentally ill, per se." "A zealot is simply someone who has an extreme, fervently held belief" and is willing to go "to great lengths to impose those beliefs, act on those beliefs." (Krakauer 2003: 305).

Ron's revelations took place within a community of fellow Mormons who all thought they were receiving revelations from God and were in the habit of discussing them. The communal nature of the revelations and other religious beliefs suggest that Ron was not mentally ill. Delusions caused by mental illness are typically delusions of a single individual, not ideas encouraged and nurtured by a group of people. The religious nature of Ron's ideas clearly did not explain his actions either because the group as a whole refused to act on Ron's "removal" revelation. Several individuals were quite disturbed by the violent nature of his revelations. This suggests that the violence of the revelations were a reflection of Ron's fantasies and urges. In short, they reflected his bad character. Did he even believe that they were revelations, one might ask, given how convenient they were given his aims and goals? Ron certainly had odd beliefs, but were they delusional? He stopped working and paying taxes because he thought he should not have to do either; he mistreated his wife and children for not obeying him unquestioningly, and so on. This is more suggestive of an immature and entitled attitude than of a subject suffering from delusions.

Was Ron mentally ill? Dr Golding suggested that perhaps he suffered from Narcissistic Personality Disorder. However, as we have seen, mental illness is only a necessary, not a sufficient, condition for insanity. What is crucial is that it be plausible the criminal act was the result of the illness. Dr Golding argued that there was no reason to think that Ron's narcissism caused him to believe that he was receiving revelations from God or to orchestrate the killings of others. Few narcissists engage in the sort of harmful activities that Ron did, even if they are grandiose, self-centered, and lacking in empathy for others. By contrast to Ron, Kim John was diagnosed with delusional disorder, persecutory type, which is associated with anger and violence (American Psychiatric Association 2000). A person who suffers from persecutory delusions believes "he or she is being tormented, followed, tricked, spied on, or ridiculed." (2000: 299).[9] This

or an increased risk of suffering "death, pain, disability, or an important loss of freedom" (p. xxxi).

[9] Command hallucinations are not always violent. When they are, people are less likely to comply with them (My Lee et al. 2004). Those that do are more than twice as likely as those who do not to have a recent history of violence. Furthermore, a subject's beliefs about the effects of the action, how socially acceptable it is, and so on, affect whether or not she will comply with the command (Beck-Sander, Birchwood, and

makes it rather likely that the disorder played an important role in Kim's beliefs about the Church's responsibility in his misfortunes. Compare such delusions with Ron Lafferty again. Ron blamed a group of people for his wife leaving him and he was, in part, right. Clearly he failed to take into account his own role in the split, but these individuals *did* encourage her to leave and take the kids. In short, Ron had the quite reasonable belief that a number of individuals encouraged his wife to abandon him. His ensuing belief that they should be killed was therefore not the result of delusions that these individuals were (partly) responsible for his wife's departure. Quite likely, it was the result of his desire for revenge.

3. LESSONS OF INSANITY

We are now in a better position to consider whether those who have significantly different values suffer from the same epistemic deficit as those who are insane. Wolf appears to have used JoJo as an example of how moral intransigence collapses into insanity. So let us begin by thinking a bit more about that example. JoJo has perpetrated a great number of crimes over many years, including murder and torture. He has done so freely, with a callous disregard for the well-being of others, and without remorse. Our main evidence for him having a mental disorder is his twisted values. His is not a generally decent moral character overwhelmed by disease. There is little question of his actions being compelled. Could he have evaluated his twisted values objectively and, finding them lacking, would he have been able to change them? Well, imagine JoJo becoming a subject and someone else the dictator ruling as he and his father before him had ruled. Would he come to regard torturing or killing someone on a whim as wrong? My hunch is that, yes, he would be able to see that such actions were wrong, and he would cease to embrace such values once he was at the wrong end of them, as it were. So it certainly *seems* as if JoJo has the ability to evaluate and change his values.[10]

Would JoJo's actions have been acceptable had his beliefs been true (the *As-if* rule) and his actions the direct result of his illness (the *Only-if* rule)? Would torturing and killing people on a whim have been morally permissible if torturing and killing people on a whim were morally permissible?

Chadwick 1997). Having command hallucinations to harm someone are most likely not sufficient to cause people to comply in the absence of other deluded beliefs (McNiel, Eisner, and Binder 2000). Kim John, however, seemed to have no shortage of delusions.

[10] According to David Faraci and David Shoemaker (2010), most ordinary subjects also judge JoJo to be blameworthy for his actions.

Obviously. But this cannot be the right reading of the *As-if* rule since *any* conviction could pass the rule. Rather, it is its acceptability within *our* moral and legal system that is at issue, and within those torturing or killing people on a whim is *not* permissible. This means that the possession of divergent moral or legal values is not typically an excusing condition *even if* such values are the result of mental disorder. As mentioned above, command hallucinations are an exception. Here the defendant's actions may *not* have been justified had their beliefs been true. In these cases, it is usually the pervasiveness of delusions that excuses, as in the case of Kim John. As in other cases of insanity, it is not the subject's *moral* values that are the primary issue. Indeed, if the Church were responsible for the actions John thought it was, it would be culpable. This fact shows just how much overlap there is between John's moral values and ours. The main problem with JoJo, however, is his values.

This brings us to the *If-only* rule, which JoJo also does not satisfy. He does not suffer from any currently recognized mental illness. Someone might argue that had JoJo been raised otherwise, he would not have performed the actions that he did, wherefore we cannot hold him responsible. It is instructive to note that this is *not* Wolf's argument. *How* people come to have the values they do is irrelevant. What matters are the values with which they end up. If we are to believe Wolf, most values that diverge from ours in significant respects render the agents who possess them unable to evaluate themselves "sensibly and accurately" and change themselves accordingly.

JoJo is, of course, just one example. The fact that JoJo is not unable, rather than merely unwilling, to change his values does not show that there are not people who are incapable of comprehending that certain types of actions are morally or legally impermissible, such as torture, murder, rape, or persecution of others. What would it take, one wonders, to be *incapable* of such a thing? Two interpretations present themselves. Either the agent's ability to evaluate and change values is directly affected, or it is intact, but her environment does not yield opportunities for such evaluation and for changing her values to be more aligned with values that we now regard to be right. In the latter case, in particular, we need to distinguish those who have sufficient access to values and information for us to say that they *hold on* to their values from those whose environment and values are such that we cannot expect them to be able to reach values similar to those we now adopt. But first we must explore what knowing that what one is doing is wrong amounts to.

The law typically does not care whether you are ignorant of the law or whether you *agree* with it. Nevertheless, some readings of the *McNaughtan Rule* demand that the defendant know that what he was doing was *morally* wrong. In other words, the person must understand that the action committed was not merely illegal (*malum prohibitum*), but also bad in itself (*malum in se*). It is not just the ability to appreciate that, say, killing *in general* is *malum in se* that can be required for legal responsibility, for most people judged not guilty by reason of insanity understand that. What must be meant is that, at the time of the crime, the defendant was able to understand that committing *this* action was morally impermissible.

Being able to understand that an action is morally wrong cannot translate to truly believing, at the time, that what you are doing is wrong. For, according to Roy Baumeister, "[m]ost people who perpetrate evil do not see what they are doing as being evil." (1997: 1).[11] Our assessment of a criminal act is usually different from that of the perpetrator. Perpetrators tend to trivialize their transgressions and think that, given the circumstances, they could not have helped doing what they did. By contrast, most victims magnify the wrongness of the event and the wickedness of perpetrator's intentions (Baumeister 1997). In the laboratory, too, people minimize their own wrong doings while they magnify wrongs perpetrated against them (Baumeister, Stillwell, and Wotman 1990). Not surprisingly, therefore, there is a good correlation between a person's access to justification for violence and the degree to which they engage in it (Berkowitz and Powers 1979; Calvete 2008, Schwartz, O'Leary, and Kendziora 1997, Zelli et al. 1999).

Crimes like violent assault or homicide are often the result of an escalated process of reciprocal provocation (Berkowitz 1978, Luckenbill 1977, Berkowitz 1978). It is not uncommon for the victim to throw the first punch in such confrontations (Wolfgang 1958). Consequently, the person who ends up killing another often sees himself as a victim. He was, after all provoked, threatened, or attacked, and his action was justified under the circumstances (Katz 1988). In such escalations, one person sees his act of retaliation as equitable whereas the other usually regards it as an unjustified escalation. As a result, both end up feeling victimized (Stillwell, Baumeister, and Del Priore 2008). Clearly, therefore, many convicted criminals did not see their actions as impermissible at the time.

The above makes it clear that we cannot simply demand that the subject understands that there is some general moral injunction against the action at hand. The average killer, like the average insane person, does not disagree with the injunction against killing. Typically, he thinks that *this* instance of

[11] Philosophers will be familiar with the idea from *The Meno* where Socrates argues that nobody who "recognizes evils for what they are" desires them (77C).

killing is not wrong or at least that it is not *that* wrong. The circumstances were sufficiently mitigating—he insulted me, she was cheating on me, he attacked me first, she is a blot on my honor—that I should not be held responsible for murder (but perhaps for some lesser crime). Though it is quite possible that the criminal is right in some of these cases, in most cases the court judges that he should have realized that the situation did not call for that degree of violence. The standard here is usually that of a "reasonable man." So perhaps this is a better way of thinking about the requirement that one understands that one's action is *malum in se*: *if* there is an inferential link from beliefs and values that the person has at the time of the crime to the recognition that the action is morally impermissible, *then* the person is responsible and culpable should they nevertheless perform the action (cf. Williams 1981).

The condition is too weak, however. A person might defend herself by pointing out that she did not embrace the relevant values at the time. If she does not believe killing, or *this* form of killing, to be wrong there is no inferential route to her appreciating that *this* killing was wrong. But clearly we would still hold her responsible. Being on the fence about, or disagreeing with, a moral prohibition does not seem to excuse. Consider honor killing in the West.

People who commit honor killings believe that killing to preserve or reinstate honor is morally permissible by contrast to other forms of killing, which they agree are impermissible. This is causing something of a stir in North America. In Canada, Mohammad Shafia, his second wife, and their son were recently found guilty of killing Shafia's first wife and the couple's three daughters. Apparently, the three daughters objected to wearing the hijab, wanted to spend time with their friends after school, have boyfriends, and so on. The story is very similar to another recent honor killing in Toronto where Aqsa Parvez was murdered by her father, Muhammad, and her brother, Waqas, for bringing "insult" to the family for similar reasons (*CBC News*, June 15, 2010). The two Waqas's received life sentences.

Both Mohammad Shafia and Muhammad Pavez and their families are recent immigrants from, respectively, Afghanistan and Pakistan, where values differ significantly from those of the West. In both countries honor killings are disturbingly common, and though prohibited by law, are rarely punished. Pakistan has a long tradition of *karo-kari*, the practice of killing women for their perceived immoral behavior, e.g. adultery, having been raped, refusing an arranged marriage, or wanting a divorce. Often, girls and women are killed on mere suspicion of some sort of affiliation with a male outside the family. Honor killers doubtlessly regard their actions as justified. The Shafias and Pavezs evidently acted in accordance with their native morality.[12] They do not *accept* that killing to save honor is wrong.

In the case of honor killings, it is fairly clear that failure to accept or adopt a moral or legal norm does not deprive someone of responsibility. Honor killings are cases of moral intransigence, not of moral inability. They are no different from typical homicides. Neither killer believes they are truly culpable. However, this lack of recognition is not usually thought to be mitigating. By contrast, it is often thought to make the offender *more* culpable. For their action is not the result of a mistake or an accident, it is the result of adopting values that we find despicable. It is their beliefs about the moral value of Jews, in particular, that make Nazis so despised. It is the belief that women are chattel and have no value other than to serve men and bear male children that makes many of the practices in the Middle East so horrific. The inferential link we require for responsibility, therefore, must be to the current values of the person in question *or* to values that we judge that she *ought* to have had under the circumstances.

If we assume that culture does not inhibit people's ability to evaluate or change their values—I shall argue for this shortly—then what is at issue with honor killers is whether they *hold on to* their values or whether they do not have access to relevant values and information that would enable them to change them in the right way. Wolf suggests that if one is brought up in a particular moral milieu, which allows, perhaps encourages, certain wrongs, one is not really in a position to accurately evaluate or change those values. It remains obscure, however, why *we* have the capacity to change our wrongheaded values. My hunch is that Wolf thinks that there is a pretty straightforward inferential route from our current moral values to future, improved, ones. For instance, our indifference to the death and suffering of nonhuman animals, particularly the ones that we eat, is quite plausibly culpable. We have the capacity to arrive at valuing the life and well-being of nonhuman animals because of values and beliefs that we either possess or that are readily available in our environment. We believe that if an action or policy creates unnecessary or avoidable suffering, then we have a prima facie reason not to perform or institute it. We also know—or if we do not actually know, we could easily come to know—that factory farming creates a great amount of suffering, and that we do not need to consume as much meat as we do for proper nutrition. From those beliefs, there is a relatively straightforward inferential route to the belief that factory farming is wrong and that we ought to oppose policies that permit it. As we shall see, this line

[12] This is not to say that many people living in those countries do not regard honor killings with horror, and would never engage in, nor condone such acts.

of reasoning shows not only that *we* are responsible, but also that there usually is little question of someone's culture inhibiting her responsibility. If *we* are responsible for our wrongdoings, so are the slaveholders of Ancient Greece and the male chauvinists of our father's generation. If they are not, neither are we.

Michele Moody-Adams (1994) has argued that people in cultures that permit, or encourage, practices that we condemn, such as slavery, are typically exercising so-called affected ignorance. That is, they chose not to question, seek information or otherwise know about these practices of wrongdoing.[13] I suspect that there is another form of affected ignorance that derives from the degree of difficulty involved in endorsing values that significantly diverge from the culture at large, not to mention advocating a societal change of standards. It requires some imagination to envisage a different moral order. It is, perhaps, an affected lack of imagination. Wolf maintains that the ancient Greeks cannot be held responsible for their attitudes towards slavery. The question, however, is whether there is a not too onerous inferential route from beliefs and values that the ancients possessed that leads to the recognition that slavery is morally wrong.

At the time of its practice, slavery was widely regarded as a terrible fate. When Andromache bewails the death of Hector in *The Illiad*, she decries the fate of the citizens of Troy: "all who will soon be carried off in the hollow ships and I with them—And you, my child, will follow me to labor, somewhere, at harsh, degrading, work, slaving under some heartless master's eye" (Book 24, line 860 ff.).[14] In Xenophon's *Symposium*, Antisthenes suggests that enslaving others is a crime: "Want prompts a thousand crimes, you must admit. Why do men steal? why break burglariously into houses? why hale men and women captive and make slaves of them? Is it not from want?" (Xenophon 2008a: §27) In *Hellenica*, Xenophon not only talks about the great lengths that people will go to, to avoid being "reduced to" slavery, but also recounts of the Spartan general Callicratidas who refused to enslave the Methymnaeans because they were fellow Hellenes (Xenophon 2008b). In the *Politics*, Aristotle refers to people who "affirm that the rule of a master over slaves is contrary to nature, and that the distinction between slave and freeman exists by law

[13] Another feature of affected ignorance is the rephrasing of wrongs in relatively inoffensive sounding language. For instance, the message on the Rwandan radio encouraging the Hutus to kill the Tutsis was "cut down the tall trees" (Dallaire 2004).

[14] The sentiment appears to have been typical in all times where slavery was a predictable result of conquest. Thus, in *Beowulf* "[a] Geat woman too sang out in grief; with hair bound up, she unburdened herself of her worst fears, a wild litany of nightmare and lament: her nation invaded, enemies on the rampage, bodies in piles, slavery and abasement." (line 3150 ff.)

only, and not by nature; and being an interference with nature is therefore unjust." (*Politics*, Book I, Part III) In *Rhetorica ad Alexandrum*,[15] the Sophist Alcidamas says "The deity gave liberty to all men and nature created no one a slave" in reference to the Thebans freeing the Messenians, who had been taken slaves by the Spartans (Garlan 1988: 125).

The passages above suggest that it was not unthinkable to the Ancient Greeks that slavery was wrong. Clearly some people thought slavery was unjust. Furthermore, the elements for a realization of the moral wrongness of slavery were there. Every free person wanted to remain free and regarded slavery as degrading and awful (1988). They recognized that slaves were fellow human beings, that they were capable of suffering, that suffering was morally relevant, that they themselves might be in danger of enslavement by others, and so on. All the talk of the baseness and stupidity of slaves that we also find in the extant literature seems designed to protect an affected ignorance of the wrongness of the institution. After all, as long as *you* are not a slave but *others* are your slaves, it is to your advantage to embrace a norm that permits slavery. If we were to make a comparison to current culture, we might point out that we, too, are in the possession of everything we need to know to recognize that factory farming, for instance, is an immoral practice. Far from being *incapable* of recognizing that this is so, we are choosing to ignore it because it is difficult to imagine not eating meat or eating meat much more rarely, to imagine a societal change, to eat differently from everybody else, and so on, not to mention the economic difficulties that would be involved. Despite these difficulties, however, we are hardly *unable* to recognize the wrongness of the practice.

There are other features of cultural value systems that speak against them having the power to deprive agents of responsibility. It is no secret that to talk of *a* value system of a particular culture is something of an idealization. Though there may be agreement about the most serious forms of transgressions, a society is not characterized by complete agreement about moral norms. Furthermore, such values are subject to change. And, as a matter of historical fact, values *do* change. Such change may be a relatively fluid affair, or be more cataclysmic. Now, *people* instantiate or embody norms (Moody-Adams 1994). For change in values to be possible, people must be able to change their norms: evaluate them, adopt them, defend them, relinquish them, overthrow them, and so on. The point is obvious on reflection. *If* it were true that people brought up in societies where slavery was

[15] Written around the same time as *Rhetoric,* it was traditionally attributed to Aristotle, but might have been written by Anaximander.

sanctioned by law and common morality were thereby incapable of thinking it was wrong, *then* we should expect no moral change. But such change *did* happen; not in Ancient Greece, but in Europe and the Americas. Therefore, adoption of seriously wrongheaded norms does not, by itself, deprive a subject of responsibility (or reduce said responsibility).

The fact that slavery *was* abolished suggests two things. First, there is an inferential route from values and beliefs the subject did have or values and beliefs that she was capable of gaining access to (without unreasonable hardship) to holding the belief/adopting the value that slavery was wrong. Second, a person who is capable of embracing one set of norms is capable of evaluating and changing said norms. The capacity for self-evaluation and change is, after all, a ubiquitous feature of our abilities. When I was a child, I believed in God, now I do not. More pertinently, perhaps, I once thought abortion was wrong, now I do not. This says something about my belief and value formation abilities. Possessing skills, tastes, habits, and values do not prevent change. To the contrary, it reveals the capacity to acquire, evaluate, update, and change such skills, tastes, habits, and values. Unless the subject has been exposed to some physiological or psychological insult, if she possesses values, she has the capacity to think about them, consider their worth, and change them if required. Consequently, people from other cultures typically have the capacity to self-evaluate and change their values accordingly. Furthermore, in many, if not most, instances their environment is sufficiently rich to contain values and information sufficient for them to change their values to what we now take to be the right ones. Consequently, they can be held responsible for their wrongdoing because their adoption of wrongheaded values was based on affected ignorance.

None of this shows that *all* cases of cultural differences in value are due to affected ignorance, nor that there are no agents who are *genuinely* incapable of comprehending that some of their values are questionable. But we have seen that affected ignorance characterizes many such cases. And where it does not, it is not the capacity to evaluate or change one's values that is at issue, but whether the person's value and environment would have allowed her to reach values close enough to what we think are the right ones. The slavery example suggests that affected ignorance is the main culprit behind divergent moral values. But to make this argument requires a more thoroughgoing exploration. Suffice it to say that affected ignorance is an important factor behind the holding of intransigent and divergent values.

The usual examples of cultural differences in values are, at any rate, quite different from the prototypical cases of insanity. The insane rarely possess values that differ from ours, nor do they suffer from specifically moral deficits.[16] They typically suffer from delusions and hallucinations that affect many different areas of their lives, not just their values.[17] These delusions or hallucinations create the disturbance that gives rise to the criminal action. For instance, an insane person might think that he is facing a creature very different from the one he is actually struggling with, like the infanticidal father thinking he's fighting a snake; he might think that the person has designs on his life and take himself to be acting in self-defense, as apparently McNaughtan did; or he could believe he is committing a harm only to avoid a greater future harm, as in the case of Yates. Instead of espousing substantially different values, the insane are typically wrong about nonmoral matters of fact. In other cases, their madness creates lacunas in their moral outlook. A person suffering from command hallucinations might think that she ought to carry out an otherwise prohibited action, for instance kill someone. But even in these cases there is rarely a radical change in their moral compass.

When it comes to insanity, then, determinations of someone's responsibility depend in no small measure on whether their belief formation is subject to undue influences as a result of mental disorder (or physiological insult), to what extent, and how it affects their values. Depression is not typically an excusing condition, for though the subject's thoughts are affected by the depression—her life seems lackluster and meaningless—it is unlikely to affect her in such a way that she becomes unable to tell right from wrong or become incapable of acting in accordance with her values.[18] At the other extreme, someone in the grip of psychosis whose vision of the world has been distorted may kill her child in the mistaken belief that she is preventing future greater harm. The prototype of insanity looks little like the prototype of intransigent values that differ from ours.

[16] This is why I think psychopaths are not insane (Maibom 2008).

[17] One might argue that people from different cultures possess a whole range of false beliefs, which play a distorting role similar to that of delusions or hallucinations, and that therefore the analogy between insanity and culturally induced values holds. Not so. First, people who have looked for such differences in beliefs that would be relevant to the morally divergent views have had difficulties finding them (Brink 1989; Doris and Plakias 2008). Second, by contrast to ordinary beliefs, subjects tend to be deeply convinced by the truth of their delusions although they are often bizarre. A delusion is not justified by the available evidence—it is isolated from other relevant beliefs—and is often held despite overwhelming reasons not to believe it. And hallucinations involve as-if perceptions, which is not characteristic of false beliefs generally. Third, insofar as we are all subject to culturally induced beliefs, that are similar to delusions, either we are as little responsible for holding values derived from them as are people from different times or cultures, or we are all responsible.

[18] Or rather, we may excuse her for small failures that seem to flow from her anhedonic state, but should she kill someone we will be skeptical that "the depression made her do it."

4. CONCLUSION

If there is an inferential route from someone's beliefs, values, etc. to knowledge that what she is doing is wrong, and it was not too onerous to follow, then we can hold her responsible. Moreover, if she does not know what she is doing is wrong and she does not possess beliefs or values from which there is an inferential route to such knowledge, but she could acquire such knowledge, beliefs, or values without too much hardship given her environment and her mental condition, she can also be held responsible. This is, I have argued, the best interpretation of the epistemic condition on responsibility, at least when it comes to moral intransigence. Choosing to ignore that what one does is wrong *or* simply disagreeing with one's community about what is morally right or wrong cannot be held up as an excuse.

Wolf is right to point out that *if* a person is unable to objectively evaluate or appropriately change her values, *then* she should be excused *ceteris paribus*. She cannot be held responsible for performing actions that she was unable to know were wrong. However, culture does not, in general, inhibit moral change, nor do divergent upbringings. The very fact that we possess values suggests that we are capable of changing them, barring madness or brain damage. Whether a person can change her values so that they are more in accord with what we now believe are the right ones depends, of course, on her values and beliefs and the information that her environment affords. I argued that ancient Greek slaveholders can be held responsible, and I see little reason to think male chauvinists of our father's generation cannot also be blamed. It may turn out that most cases of culturally divergent values are due not to *inability*, but to *inexpediency*. Unless we simply assume that one can only be held responsible if one thinks of one's action as wrong, we should not suppose that people we usually judge to be the most evil—e.g. perpetrators of genocides, slaveholders, and pedophile rapists—are the least responsible. Such a view reduces all culpability to moral incontinence. But as we have seen, most who do wrong do not take themselves to do so. These wrongs are not typically perpetrated by individuals who are *unable* to see the error of their ways, but by individuals who are *unwilling* or are *neglecting* to do so; they engage in affected ignorance. In the case of cultural differences, there is often an inferential route from values and beliefs that are held by the person to what we take to be the right values. For instance, there is little question that the Hutu killers were able to recognize the humanity of their Tutsi neighbors, that they recognized the prohibition on killing, etc., etc. It was, however, expedient to ignore this, so those who were not coerced into engaging in the genocide chose to ignore the wrongness of their actions. Consequently, they did not think of their actions as wrong. Yet, we can surely hold them responsible.

REFERENCES

American Psychiatric Association (2000). *Diagnostic and Statistical Manual of Mental Disorders,* 4th edn. Text Revision *(DSM-IV)*. (Washington DC: American Psychiatric Association).

Aristotle (1984). "The Politics." Trans. Jowett. In: *The Complete Works of Aristotle.* ed. J. Barnes. (Princeton, NJ: Princeton University Press).

Baumeister, R. (1997). *Evil: Inside Human Violence and Cruelty.* (New York: Henry Holt).

——Stillwell, A., and Wotman, S. (1990). "Victims and perpetrator accounts of interpersonal conflict: Autobiographical narratives about anger." *Journal of Personality and Social Psychology* 59, 994–1005.

Beck-Sander, A., Birchwood, M., and Chadwick, P. (1997). "Acting on command hallucinations: A cognitive approach." *British Journal of Clinical Psychology* 36, 139–48.

Beowulf. (2000). Trans. Seamus Heaney. (London: W. W. Norton & Co).

Berkowitz, L. (1978). "Is criminal violence normative behavior? Hostile and instrumental aggression in violent incidents." *Journal of Research in Crime and Delinquency* 15, 148–61.

——and Powers, P. (1979). "Effects of timing and justification of witnessed aggression on the observers' punitiveness." *Journal of Research in Personality* 13, 71–80.

Brandt, R. B. (1959). *Ethical Theory: The Problems of Normative and Critical Ethics.* (Englewood Cliffs, NJ: Prentice-Hall).

Brink, D. (1989). *Moral Realism and the Foundation For Ethics.* (Cambridge: Cambridge Uuniversity Press).

Calvete, E. (2008). "Justification of violence and grandiosity schemas as predictors of antisocial behavior in adolescents." *Journal of Abnormal Child Psychology* 36, 1083–95.

CBC News, June 15, 2010. Father, son plead guilty to Aqsa Parvez murder. <http://www.cbc.ca/news/canada/toronto/story/2010/06/15/parvez-guilty-plea.html>.

Dallaire, R. (2004). *Shake Hands with the Devil: The Failure of Humanity in Rwanda.* (Toronto: Vintage Canada).

Doris, J. and Plakias, A. (2008). "How to argue about disagreement: Evaluative diversity and moral realism." In W. Sinnott-Armstrong (ed.), *Moral Psychology, vol. 2. The Cognitive Science of Morality: Intuition and Diversity.* (Cambridge, MA: MIT Press), 303–31.

Faraci, D. and Shoemaker, D. (2010). "Insanity, Deep Selves, and Moral Responsibility: The Case of JoJo." *Review of Philosophy & Psychology* 1, 319–32.

Frankfurt, H. (2003). "Freedom of the will and the concept of a person." In G. Watson (ed.), *Free Will,* 2nd edn. (New York: Oxford University Press), 322–36.

Garlan, Y. (1988). *Slavery in Ancient Greece.* (Cornell: Cornell University Press).

The Iliad. (1990). Trans..Robert Fagles. (Bath: The Bath Press).

Jeffrey, R., Pasewark, R., and Bieber, S. (1988). "Insanity pleas: predicting Not Guilty by Reason of Insanity adjudications." *Bulletin of the American Academy of Psychiatry and Law* 16, 35–9.

Katz, J. (1988). *Seductions of Crime: Moral and Sensual Attractions in Doing Evil.* (New York: Basic Books).

Krakauer, J. (2003). *Under the Banner of Heaven: A Story of Violent Faith.* (New York: Anchor Books).

Larbey, C. (2007). "The secret lives of Kim John and Francis Philip. What do we really know about the Cathedral killers?" *St. Lucia Star*, May 18, 2007.

Luckenbill, D. (1977). "Criminal homicide as a situated transaction." *Social Problems* 25, 176–86.

McNiel, D., Eisner, J., and Binder, R. (2000). "The relationship between command hallucinations and violence." *Psychiatric Services* 51, 1288–92.

Maibom, H. (2008). "The mad, the bad, and the psychopath." *Neuroethics* 1, 167–84.

Maudsley, H. (1898). *Responsibility in Mental Disease.* (New York: D. Appleton and Company).

Moody-Adams, M. (1994). "Culture, responsibility, and affected ignorance." *Ethics* 104, 291–309.

Moran, R. (1981). *Knowing Right From Wrong.* (New York: The Free Press).

MY Lee, T., Chong, S., Chan, Y., Sathyadevan, G. (2004). "Command hallucinations among Asian patients with schizophrenia." *Canadian Journal of Psychiatry* 49, 838–42.

Reznik, L. (1997). *Evil or Ill: Defending the Insanity Defence.* (New York: Routledge).

Rice, M. and Harris, G. (1990). "The predictors of insanity acquittal." *International Journal of Law and Psychiatry* 13, 217–24.

Rogers, J., Bloom, J., and Manson, S. (1984). "Insanity defences: contested or conceded?" *American Journal of Psychiatry* 141, 885–8.

Schwartz, M., O'Leary, S., and Kendziora, K. (1997). "Dating aggression among high school students." *Violence and Victims*, 295–305.

Steadman, H., Keitner, L., Braff, J., and Arranites, T. (1983). "Factors associated with a successful insanity plea." *American Journal of Psychiatry* 140, 401–405.

Stillwell, A., Baumeister, R., and Del Priore, R. (2008). "We're all victims here: Towards a psychology of revenge." *Basic and Applied Social Psychology* 30, 253–63.

Taylor, C. (1976). "Responsibility for self." In A. E. Rorty (ed.) *The Identities of Persons.* (Berkeley, CA: University of California Press), 281–99.

Watson, G. (1975). Free agency. *Journal of Philosophy*, LXII, 205–20.

Williams, B. (1981). "Internal and external reasons." In his *Moral Luck.* (Cambridge: Cambridge University Press), 101–13.

Wolf, S. (2003). "Sanity and the metaphysics of responsibility." In G. Watson (ed.), *Free Will,* 2nd edn. (New York: Oxford University Press), 372–87.

Wolfgang, M. (1958). *Patterns in Criminal Homicide.* (Philadelphia, PA: University of Pennsylvania Press).

Xenophon. (2008a). *Symposium.* Trans. H. G. Dakyns. Project Guttenberg (accessed January 14, 2012).

Xenophon. (2008b). *Hellenica*. Trans. H. G. Dakyns. Project Guttenberg (accessed January 14, 2012).

Zelli, A., Dodge, K., Laird, R., Lochman, J., and Conduct Problems Prevention Research Group. (1999). "The distinction between beliefs legitimizing aggression and deviant processing of social cues: Testing measurement validity and the hypothesis that biased processing mediates the effects of beliefs on aggression." *Journal of Personality and Social Psychology* 77, 150–66.

12

Fairness and the Architecture of Responsibility[1]

David O. Brink and Dana K. Nelkin

In this essay, we explore a conception of the nature and structure of responsibility that draws on ideas about moral and criminal responsibility. Though the two sorts of responsibility are not the same, the criminal law reflects central assumptions about moral responsibility, and the two concepts of responsibility have very similar structure. Our conception of responsibility draws on work of philosophers in the compatibilist tradition who focus on the choices of agents who are reasons-responsive and work in criminal jurisprudence that understands responsibility in terms of the choices of agents who have capacities for practical reason and whose situation affords them the fair opportunity to avoid wrongdoing.[2] We treat these two perspectives as potentially complementary and argue that each can learn things from the other. Specifically, we think that criminal jurisprudence needs a more systematic conception of the capacities for normative competence and that ideas from the reasons-responsive literature

[1] This essay is fully collaborative. The authors are listed in alphabetical order. The ideas for this essay grew out of a graduate seminar that we taught together on the topic of partial responsibility in 2008 and were refined in a seminar on responsibility that DB taught in 2011. Versions of this material were presented at the University of Illinois, the University of Calgary, the Murphy Institute at Tulane University, Cornell University, the University of Western Ontario, and the New Orleans Workshop on Agency and Responsibility. We would like to thank audiences on those occasions for helpful feedback. We owe special thanks to input from Craig Agule, Sarah Aikin, Amy Berg, Mitch Berman, Michael McKenna, Per Milam, Richard Miller, Michael Moore, Stephen Morse, Derk Pereboom, Erick Ramirez, Sam Rickless, Tim Scanlon, David Shoemaker, Jada Twedt Strabbing, Sarah Stroud, Matt Talbert, Michael Tiboris, and Gary Watson.

[2] The philosophical work in the reasons-responsive wing of the compatibilist tradition on which we draw includes Fischer and Ravizza 1998, Wallace 1994, Wolf 1990, and Nelkin 2011. The criminal jurisprudence work on which we draw includes Hart 1957 and 1961, Moore 1997, and Morse 1994, 2002, and 2003.

on moral responsibility, some familiar and some novel, can fill this need. However, we think that moral philosophers tend to focus on the capacities involved in responsibility and so tend to ignore the situational element in responsibility recognized in the criminal law literature. Our conception of responsibility brings together the dimensions of normative competence and situational control, and we factor normative competence into cognitive and volitional capacities, which we treat as equally important to normative competence and, ultimately, responsibility. Moreover, we argue that normative competence and situational control can and should be understood as expressing a common concern that blame and punishment presuppose that the agent had a fair opportunity to avoid wrongdoing. Thus, we treat the value that criminal law theorists associate with the situational element of responsibility as the umbrella concept for our conception of responsibility, one that explains the distinctive architecture of responsibility.

This essay aims to motivate and articulate this sort of fair opportunity conception of the architecture of responsibility. It is part of a larger project that develops this conception and applies it to issues of partial responsibility, involving insanity and psychopathy, immaturity, addiction, provocation, and duress. The details and applications of the fair opportunity conception of responsibility are interesting and important, and we hope to address them more fully elsewhere. But the framework itself is important and requires articulation.

1. RESPONSIBILITY, BLAME, AND THE REACTIVE ATTITUDES

P. F. Strawson famously highlighted the link between ascriptions of responsibility and the reactive attitudes (Strawson 1962). The reactive attitudes involve emotional responses directed at oneself or another in response to that person's conduct. Reactive attitudes include hate, love, pride, gratitude, anger, regret, resentment, indignation, and forgiveness. So understood, the reactive attitudes form a large and heterogeneous class. Some of these reactive attitudes have little direct connection with moral praise and blame and responsibility. Consider the difference between anger and resentment. Anger need not have moral content. I might be momentarily angry or upset with a very young child who has carelessly damaged a treasured keepsake of mine. But resentment would seem to be out of order. Resentment seems to involve a kind of anger or upset that presupposes that one has been mistreated or wronged by another in some way. This kind of moral judgment does not seem to apply to a very young child. The idea that assumptions about responsibility are embedded in the reactive attitudes makes most sense if we focus on this narrower class of reactive attitudes that are *moralized* (cf. Wallace 1994: ch. 2).

In particular, we want to focus on the attitudes and practices of praise and blame, especially as they reflect assumptions about responsibility. Some negative reactive attitudes, such as regret, don't seem to implicate responsibility at all. Bernard Williams describes the case of a truck driver who, through no fault of his own, hits and kills a child who has darted into the street (1976: 28). As Williams claims, it is appropriate for the driver to feel a kind of *agent-regret* at being the instrument of the child's death, which is distinct both from the regret or horror that bystanders might feel and from guilt for having been responsible for wrongdoing.

Normally, blame only makes sense if the agent is responsible for some kind of wrong. Some philosophers distinguish between two kinds of blame and responsibility. For instance, Gary Watson distinguishes between responsibility as *attributability* and as *accountability* (1996). An agent is responsible in the attributive sense, roughly speaking, when her actions reflect *the quality of her will* in the right way. Some kinds of blame can be a fitting response to the quality of the agent's will. For instance, A might have hard feelings toward B if B injures A through malice, recklessness, or negligence. Here, our reactive attitudes track the insufficient regard that B shows A's interests and rights. But attributability does not guarantee accountability. Consider a situation in which we find out that though B's actions exhibit malice, in no relevant sense did B have an opportunity to do otherwise, perhaps because he suffers from a serious mental illness and is not a competent decision-maker as a result. In these cases, we are likely to think that B was not at *fault* or *culpable* and so not *accountable* for the harm he did. Although hard feelings may remain and be perfectly appropriate in such cases, reactive attitudes involving resentment and indignation cease to be appropriate and tend to dissipate. Attributability is necessary but not sufficient for accountability. In this essay, we are especially interested in responsibility as accountability and its connection with reactive practices and attitudes involving blame, and we rely on an intuitive understanding of the reactive attitudes that seem to be especially responsive to accountability.[3] It is this sense of blame and responsibility that we take to be most relevant to the sort of responsibility required for punishment and to capture what is common to both moral and legal responsibility.

[3] Here, we are in agreement with Watson (1996: 276, 2011). By contrast, T. M. Scanlon develops a conception of blame that seems to presuppose only attributability, not accountability (2008: ch. 4, esp. 202).

Thus, our focus in what follows will be restricted to attitudes of praise and blame involving accountability, with special attention to attitudes and practices of blame, rather than praise. We do so because our aim is to combine insights from criminal jurisprudence, which focuses on criminal acts that involve wrongdoing, with those from moral theory. But we believe that there are natural ways of extending what we say here about attitudes of blame and blameworthy actions to praise and praiseworthy actions.

Strawson links responsibility and reactive attitudes, such as those of resentment and indignation, in a biconditional fashion.

> Reactive attitudes involving blame and praise are appropriate just in case the targets of these attitudes are responsible.

Call this biconditional claim *Strawson's thesis*. Strawson's thesis can be interpreted in two very different ways, depending on which half of the biconditional has explanatory priority.

According to the first interpretation, there is no external, or response-independent justification of our attributions of responsibility. This reading fits with Strawson's view that our reactive attitudes and ascriptions of responsibility, as a whole, do not admit of external justification. Particular expressions of a reactive attitude might be corrigible as inconsistent with a pattern of response, but the patterns of response are not themselves corrigible in light of any other standard. Similarly, particular ascriptions of responsibility might be corrigible in light of patterns in our ascriptions of responsibility, but the patterns themselves are not corrigible in light of any other standards. Responsibility judgments simply reflect those dispositions to respond to others that are constitutive of various kinds of interpersonal relationships. This is a *response-dependent* interpretation of Strawson's thesis.

This response-dependent interpretation of Strawson's thesis is probably the right interpretation of Strawson.[4] But as a systematic, rather than an interpretive, matter, we favor an alternative interpretation of Strawson's thesis that is *realist*, rather than response-dependent. This interpretation stresses the way that the reactive attitudes make sense in light of and so *presuppose* responsibility. As such, the reactive attitudes are *evidence* about when to hold people responsible, but not something that constitutes them being responsible. It's true that the reactive attitudes are appropriate if and only if the targets are responsible, but it's the responsibility of the targets

[4] Watson defends this response-dependent interpretation of Strawson's thesis, at least on interpretive grounds (1987: esp. 222). Wallace defends this interpretation of Strawson's thesis in its own right (2004: 19), though we think other elements in Wallace's account fit better with an alternative realist reading.

that makes the reactive attitudes toward them fitting or appropriate. In the biconditional relationship between responsibility and the reactive attitudes, it is responsibility that is explanatorily prior, according to this realist interpretation. Strawson points out that the limits of our reactive attitudes are indicated by our practices of exemption and excuse. Because the realist believes that the reactive attitudes presuppose responsibility, she can appeal to our practices of exemption and excuse to help understand the conditions under which we are responsible. This will be a response-independent conception of responsibility.

A response-independent conception of responsibility is hostage to traditional worries about freedom of the will. The problem of free will is the problem of reconciling responsibility with determinism, because responsibility may seem to presuppose freedom of the will, and freedom of the will may seem incompatible with determinism. Our realist approach to responsibility and the reactive attitudes is best articulated as a version of compatibilism that denies that responsibility requires a form of freedom that would be undermined by the truth of determinism. In particular, because our practices of exemption and excuse track forms of normative competence and situational control, rather than the truth of determinism, they promise to ground a compatibilist conception of responsibility. Though we will articulate this compatibilist interpretation of our project, we cannot defend it here (though we return to these issues briefly in Section 7 below).[5]

Although Strawson focuses primarily on the relation between the reactive attitudes and responsibility, his thesis fits well with a particular approach to punishment and criminal responsibility. In particular, the realist interpretation of Strawson's thesis fits with a broadly retributive approach to blame and punishment, precisely because the retributivist thinks that the reactive attitudes and our practices of blame and punishment can be appropriate responses to culpable wrongdoing, where culpable wrongdoing is wrongdoing for which the agent is responsible. To see this, it will be useful for us to say more about both blame and punishment.

[5] Because we think that reactive attitudes involving praise and blame presuppose that the targets of these attitudes are responsible, we accept the need to provide a response-independent conception of responsibility and to answer skeptical doubts about responsibility. Consequently, we see response-dependent conceptions of responsibility as offering skeptical solutions to skeptical worries (cf. Kripke 1982: 66–7). We view skeptical solutions to skeptical problems as, at best, a kind of fallback solution to be entertained only after straight solutions have clearly failed. For a fuller exploration of the compatibilist aspects of this conception of responsibility, see Nelkin 2011. We believe that at least some of what we say here can be accepted by incompatibilists who accept *further* conditions on responsibility, beyond that of indeterminism.

When agents are responsible (accountable) for doing wrong blame is appropriate. Blame typically involves both *censure* and *sanction*. When we blame someone, we not only censure her conduct but also censure the agent herself for engaging in that conduct. Parents are often warned to disapprove the bad conduct of their children but not to blame them. This is because blame involves finding fault in the agent and that seems to assume that the agent is responsible (accountable) and could have avoided the conduct. Being blameworthy licenses various kinds of sanction, often informal and sometimes formal. Blame itself can involve overt reproach, which is a kind of sanction, whether directed at another or at oneself. Sometimes reproach is the only appropriate sanction. But sometimes blameworthiness licenses other informal sanctions, such as public rebuke or social distancing. And in other cases, blameworthiness might license various kinds of punishment, whether personal, social, or legal. To be blameworthy is to be a fitting object of blame, censure, and sanction. It is to be deserving of these attitudes and responses. No doubt, where sanctions are appropriate, they have to be proportionate, and there may be cases in which one is blameworthy and yet it is not on balance appropriate to blame or sanction. But, presumably, even in these cases there is a pro tanto case for blame and at least some informal sanction, if only self-reproach, as a fitting response to culpable wrongdoing.

On this view, punishment is a species of blame for culpable wrongdoing. On a broadly retributive view of the criminal law, this is true of legal punishment as well. We understand criminal punishment as the authorized deprivation of an agent's normal rights and privileges, because he or she has been found guilty of a criminal act.[6] Punishment is a form of blame, and like other kinds of blame, presupposes culpable wrongdoing. Legal retributivism, as we understand it, is the claim that legal punishment is justified on the basis of culpable legal wrongdoing. This claim can take *positive* or *negative* forms. According to positive retributivism, culpable wrongdoing is both necessary and sufficient for justifying punishment. The sufficiency claim admits of both *strong* and *weak* interpretations. According to strong sufficiency, culpable wrongdoing is a sufficient condition of justified proportional blame and punishment, whereas, according to weak sufficiency, culpable wrongdoing is sufficient for a pro tanto case for proportional blame and punishment. Weak sufficiency allows for the pro tanto case for retributive blame and punishment to be overridden in particular cases by nonculpability moral considerations, such as forgiveness or mercy. By contrast, according to

[6] Cf. Bedau and Kelly 2010. We aim for a normatively neutral and ecumenical definition of punishment and one that identifies punishment as involving deprivations of certain sorts, but not essentially involving the imposition of pain or suffering.

negative retributivism, culpable wrongdoing is necessary, but not sufficient, for justified punishment.[7]

Legal retributivism (in either version) has the virtue of explaining well the two principal forms of affirmative defense in the criminal law. According to the retributivist, justified punishment aspires to track culpable wrongdoing.[8] Wrongdoing and culpability are independent variables. Affirmative defenses, whose success justifies acquittal, deny either wrongdoing or culpability. *Justifications*, such as the necessity defense, deny wrongdoing, insisting that behavior that would otherwise be wrong is not in fact wrong in these circumstances. *Excuses*, such as the insanity defense, deny culpability or responsibility, claiming that the agent acted wrongly but was not responsible for her wrongdoing.

Here, the criminal law reflects the moral landscape well. Moral retributivism could be understood as the claim that moral blame (that presupposes accountability) and informal sanction are appropriate only as a response to culpable moral wrongdoing. It too has the virtue of explaining the two principal ways of avoiding blame—justifying and excusing conduct. Justification denies wrongdoing, and excuse denies responsibility for wrongdoing. Insofar as moral retributivism says that moral blame ought to track desert, where desert is the product of the two independent variables of wrongdoing and responsibility, it fits our moral defenses like a glove.

In this way, the realist interpretation of Strawson's biconditional can appeal to our understanding of excuses to provide a window on to the nature of responsibility.[9] An analysis of criminal law doctrines of excuse can be a part of this investigation. In this context, it is worth addressing the relationship between excuses and exemptions. The prototypical case of an exemption is a case in which an actor is not responsible for what he did because of quite general impairments of his agency. So, for instance, insanity and immaturity are sometimes described as exemptions. By contrast, excuses are sometimes claimed to be prototypically case-specific in which the agent is otherwise normal and responsible but acted

[7] Cf. Duff 2008. The view that we have called "negative retributivism" is sometimes called a "mixed theory" of punishment, because it requires more than one type of justificatory reason, typically, both retributivist and consequentialist. It is also worth noting that our definition of retributivism does not commit retributivists to endorsing the thesis that punishment is intrinsically good, as some retributivists claim.

[8] Precisely for this reason, a skeptic about moral responsibility will deny that any retributivist view of punishment can be correct. For a skeptical view and its relation to punishment, see Pereboom 2012.

[9] Moore describes excuse as the "royal road" to responsibility (1997: 548). Whereas the realist regards our practices of excuse as potential evidence of a response-independent conception of responsibility, a response-dependent conception will understand our practices of excuse as constitutive of responsibility.

inadvertently or was subject to coercion in a specific situation. Despite the existence of these two different kinds of prototypical cases, we think that it is a mistake to treat exemptions and excuses as disjoint classes. First, while Strawson and others include insanity among the exemptions, the criminal law treats insanity as an excuse. In fact, the criminal law includes all claims to less than full culpability in the single category of excuse. So there is some reason not to assume that exemptions cannot be excuses. Second, the prototypical cases are not exhaustive of the possibilities, as Strawson himself recognized (1962: 79). Strawson's partition is into cases in which the reactive attitudes are generally disabled in regard to a particular agent and cases in which they are selectively disabled due to inadvertence or compulsion. But there are at least three different dimensions on which these paradigm cases can be distinguished: *scope*, *duration*, and the *location* of the obstacle to culpability. Immaturity, for example, or even more temporary conditions, such as depression or even dementia due to dehydration, might undermine responsibility for all sorts of actions during the episode in question, and so have wide scope. In contrast, a particular perceptual deficit, or a compulsive disorder narrowly confined to one area, like kleptomania, might have a relatively narrow scope. Paradigm cases can also be distinguished on the basis of duration. A phobia, for example, might affect one's choices in a narrow area, but be lifelong, in contrast to a short spell of dementia caused by dehydration. The third dimension is the location of the obstacle as either within or outside of the agent. Immaturity is an example of the former, and low lighting conditions that prevent one from seeing someone else in need is an example of the latter. All three of these are separable in principle, but in the original paradigm cases, go together. For example, childhood is long-lasting, has a wide scope (though narrowing as one ages), and seems to be explained by the agent's own capacities. Not realizing one is stepping on another's toes, in contrast, is typically short-lived, narrow in scope, and explainable by something about the particular situation rather than one's capacities. Recognizing that considerations that mitigate culpability can fall in a variety of places along all three dimensions suggests to us that it would be most useful to consider all of the cases as ones involving potential excuses with varying degrees of scope and duration and with varying locations between the agent and the situation. On the proposal that we favor, exemptions are best understood as comparatively global or standing excuses.

The challenge for the realist interpretation of Strawson's thesis is to use the reactive attitudes as evidence to uncover an independent conception of responsibility that can support the reactive attitudes. If we study responsibility by studying excuses, we find that excuses factor into two main kinds on the location dimension. Some excuses reflect compromised psychological capacities of agents. We will conceptualize these as failures of *normative competence*. Insanity is the most familiar excuse of this type. But some excuses reflect no failure of normative competence. Instead, they reflect a lack of normal *situational control*. In such situations, though the agent is normatively competent, factors external to her deprive her of the fair opportunity to avoid wrongdoing. Coercion and duress provide excuses of this type. We think that attention to these two kinds of excuse provides the key to understanding the architecture of responsibility.

2. NORMATIVE COMPETENCE

If someone is to be culpable or responsible for her wrongdoing, then she must be a responsible agent. So we need to distinguish between responsible and nonresponsible agents. Our paradigms of responsible agents are normal mature adults with certain sorts of capacities. We do not treat brutes or small children as responsible agents. Brutes and small children both act intentionally, but they act on their strongest desires or, if they exercise deliberation and impulse control, it is primarily instrumental reasoning in the service of fixed aims. By contrast, we suppose, responsible agents must be *normatively competent*. They must not simply act on their strongest desires, but be capable of stepping back from their desires, evaluating them, and acting for good reasons. This requires responsible agents to be able to recognize and respond to reasons for action. If so, normative competence involves *reasons-responsiveness*, which itself involves both *cognitive* capacities to distinguish right from wrong and *volitional* capacities to conform one's conduct to that normative knowledge.[10]

It is important to frame this approach to responsibility in terms of normative competence and the possession of these capacities for reasons-responsiveness. In particular, responsibility must be predicated on the possession, rather than the use, of such capacities. We do excuse for lack of competence. We do not excuse for failures to exercise these capacities properly. Provided they had the relevant cognitive and volitional capacities, we do not excuse the weak-willed or the willful wrongdoer for failing to

[10] In framing our approach to the internal dimension of responsibility this way, we draw on previous work in the compatibilist tradition that emphasizes normative competence (Wolf 1990, Wallace 1994) and reasons-responsiveness (Wolf 1990, Wallace 1994, Fischer and Ravizza 1998, and Nelkin 2011) and distinguishes cognitive and volitional dimensions of reasons-responsiveness (Wallace 1994, Fischer and Ravizza 1998).

recognize or respond appropriately to reasons. If responsibility were predicated on the proper use of these capacities, we could not hold weak-willed and willful wrongdoers responsible for their wrongdoing. It is a condition of our holding them responsible that they possessed the relevant capacities.[11]

Normative competence, on this conception, involves two forms of reasons-responsiveness: an ability to recognize wrongdoing and an ability to conform one's will to this normative understanding. Both dimensions of normative competence involve norm-responsiveness. As a first approximation, we can distinguish moral and criminal responsibility at least in part based on the kinds of norms to which agents must be responsive. Moral responsibility requires capacities to recognize and conform to moral norms, including norms of moral wrongdoing, whereas criminal responsibility requires capacities to recognize and conform to norms of the criminal law, including norms of criminal wrongdoing.

Reasons-responsiveness is clearly a modal notion and admits of degrees; one might be more or less responsive. This raises the question how responsive someone needs to be to be responsible. This is an important and difficult issue, deserving more careful discussion than we can give it here. We make some preliminary remarks here, which we will refine in later sections. We might begin by distinguishing different *grades* of responsiveness. Here, we adapt some ideas from John Fischer and Mark Ravizza in their book *Responsibility and Control* about the responsiveness of the mechanisms on which agents act to our issue about how reasons-responsive the agents themselves are.[12] We propose to specify the degree to which an agent is responsive to reasons in terms of counterfactuals about how she would believe or react in situations in which there was sufficient reason for her to do otherwise.[13] An agent is more or less responsive to reason

[11] Sidgwick famously objects to Kant's conception of autonomy as conformity to principles of practical reason that this would prevent us from holding criminals responsible and would allow us to recognize only morally upright behavior as responsible (Sidgwick 1907: 511–16). The solution to this problem is for Kant to define autonomy in terms of *capacities* for conformity to principles of practical reason.

[12] The conception of reasons-responsiveness that Fischer and Ravizza defend is mechanism-based, rather than agent-based (1998: 38). By contrast, we favor a version of reasons-responsiveness that is agent-based, rather than mechanism-based, precisely because we think that responsibility and excuse track the agent's capacities, rather than the capacities of her mechanisms. For defense of the agent-based approach, see Nelkin 2011: 64–79 and McKenna 2012.

[13] For present purposes, in specifying an agent's capacities in terms of such counterfactuals, we can remain agnostic about whether capacities or counterfactuals have explanatory priority, in particular, whether capacities ground the counterfactuals or whether the capacities just consist in the truth of such counterfactuals.

depending on how well her judgments about what she ought to do and her choices would track her reasons for action.

We could begin this process by distinguishing two extreme degrees of responsiveness.

- *Strong Responsiveness:* Whenever there is sufficient reason for the agent to act, she recognizes the reason and conforms her behavior to it.
- *Weak Responsiveness:* There is at least one situation in which there is a sufficient reason to act, and the agent recognizes that reason and conforms her behavior to it.

However, it does not seem plausible to model normative competence in terms of either strong or weak responsiveness. Strong responsiveness is too strong for the same reason we gave for focusing on competence, rather than performance. We do not require that people actually act for sufficient reasons to do otherwise; it is the capacities with which they act that matter. The weak-willed are, at least typically, responsible for their poor choices. Moreover, weak responsiveness seems too weak. It treats someone as responsive in the actual situation even if she did not respond in the actual situation and there is only one extreme circumstance in which she would recognize and respond to reasons for action. The Goldilocks standard of responsiveness evidently lies somewhere between these extremes. Of course, there is considerable space between the extremes—the gap between always and once.

We might stake out an intermediate form of responsiveness in something like the following terms.

- *Moderate Responsiveness:* Where there is sufficient reason for the agent to act, she regularly recognizes the reason and conforms her behavior to it.

Moderate responsiveness is deliberately vague; it specifies a range or space of counterfactuals that must be true for the agent to be responsive. Ideally, we would be able to specify a preferred form of moderate responsiveness more precisely. But what is important for present purposes is that reasons-responsiveness is a matter of degree and that the right threshold for responsibility is probably some form of moderate responsiveness.

So far, this conception of responsiveness is coarse-grained in ways that might prove problematic. For one thing, it lumps together cognitive and volitional dimensions of responsiveness. But if they are independent aspects of normative competence, then we may need to assess responsiveness along these two dimensions separately. Moreover, it is at least conceivable that we might require different degrees of responsiveness in cognitive and volitional dimensions of competence. For instance, Fischer and Ravizza distinguish the cognitive and volitional dimensions of reasons-responsiveness in terms of "reasons-receptivity" and "reasons-reactivity" (respectively). Their conception of reasons-responsiveness is *mixed*, because it treats receptivity and reactivity *asymmetrically*. They combine moderate receptivity and weak reactivity (1998: 81–2). We ultimately reject this asymmetry, but it represents a conception of responsiveness worth considering.

Furthermore, this initial formulation of responsiveness assumes that we consider all situations in which there is sufficient reason to act together. But we may find it more informative to partition possibilities into groups, depending on the kinds of reasons at stake and other aspects of the situations in which agents finds themselves. For instance, in deciding whether an agent had sufficient volitional capacity to overcome fears that stood in the way of her performing her duty, we may think it best to restrict our attention to those counterfactuals in which she faced threats or fears of comparable kind or magnitude.

For these reasons, we may need to make our assessments of the degree of an agent's responsiveness more fine-grained in several ways. We address some of these complications below.

3. THE COGNITIVE DIMENSION OF NORMATIVE COMPETENCE

Normative competence requires the cognitive capacity to make suitable normative discriminations, in particular, to recognize wrongdoing. If responsibility requires normative competence, and normative competence requires this cognitive capacity, then we can readily understand one aspect of the criminal law insanity defense. A full account of the elements of insanity is controversial, as we will see. But most plausible versions of the insanity defense include a cognitive dimension, first articulated in the *M'Naghten* rule that excuses if the agent lacked the capacity to discriminate right from wrong at the time of action.[14]

Here is one place it might be important to distinguish between the demands of moral and criminal responsibility. Presumably, moral responsibility requires the ability to recognize moral norms, including norms that specify moral wrongdoing, whereas criminal responsibility requires the ability to recognize criminal norms, including norms that specify criminal wrongdoing.[15] The cognitive abilities to recognize these two different kinds

[14] *M'Naghten's Case*, 10 Cl. & F. 200, 8 Eng. Rep. 718 (1843).

[15] There is a debate about whether the cognitive dimension of the insanity test, expressed in *M'Naghten's* rule, should be formulated in terms of capacities for recognizing criminal or moral wrongdoing. British criminal law has focused on criminal wrongdoing, and American jurisdictions remain divided.

of norms might be different, with the result that it might be possible to be criminally responsible without being morally responsible and vice versa.

It is common to contrast reason and emotion. This common contrast might lead one to suppose that the cognitive dimension of normative competence is purely cognitive and does not involve emotion or affect. But this conclusion would be misguided. Emotional or affective deficits may block normative competence by compromising cognitive capacity. For instance, lack of empathy may make it impossible or very difficult to recognize actions as injurious and, hence, legally or morally wrong. There is also evidence that congenital damage to the amygdala, which is thought to be the part of the brain responsible for emotional learning and memory, may prevent the formation of normative or, at least, moral concepts. There is emerging research that shows that psychopathy involves both abnormalities in the amygdala and empathy deficits (Blair et al. 2005) Moreover, psychopaths have been thought to have trouble with a psychological test used to discriminate between moral norms and conventional norms.[16] These findings raise questions about whether psychopaths have moral concepts and so whether they have the cognitive capacity to distinguish moral right from wrong. Even if they lack cognitive moral competence, it doesn't follow that they lack the capacity to recognize legal wrongdoing. It is at least possible that some psychopaths might be criminally responsible without being morally responsible.[17] These are complicated issues that deserve fuller examination, but they illustrate ways in which emotion and affect can have a bearing on the cognitive dimension of normative competence. Here, emotional capacities may be *upstream* from normative cognition.

4. THE VOLITIONAL DIMENSION OF NORMATIVE COMPETENCE

But there is more to normative competence than this cognitive capacity. We assume that intentional action is the product of informational states, such as beliefs, and motivational states, such as desires and intentions. Though our beliefs about what is best can influence our desires, producing optimizing desires, our desires are not always optimizing. Sometimes they are good-dependent but not optimizing, when they are directed at lesser

[16] See Blair et al. 2005: 57–9. For some skepticism, see Aharoni et al. 2012.

[17] We suspect that severe psychopathy impairs, but does not eliminate, reasons-responsiveness and that it may make for a better moral excuse than a criminal excuse. Contrast Fine and Kennett 2004.

goods, and sometimes they are completely good-independent. This is reflected in cases of weakness of will in which we have beliefs about what is best (and perhaps optimizing desires) but in which we act instead on the basis of independent nonoptimizing passions and desires. This psychological picture suggests that being a responsible agent is not merely having the capacity to tell right from wrong but also requires the capacity to regulate one's actions in accordance with this normative knowledge. This kind of volitional capacity requires emotional and appetitive capacities to enable one to form intentions based on one's optimizing judgments and execute these intentions over time, despite distraction and temptation.

Here, emotion and appetite are *downstream* from cognition and play a separate, volitional or executive role. If one's emotions and appetites are sufficiently disordered and outside one's control, this might compromise volitional capacities necessary for normative competence. Consider the following obstacles to volitional competence.

- Irresistible desires or paralyzing fears that are neither conquerable nor circumventable, as perhaps in some cases of genuine agoraphobia or addiction.[18]
- Clinical depression that produces systematic weakness of will in the form of listlessness or apathy.
- Acquired or late onset damage to the prefrontal cortex in which agents have considerable difficulty conforming to their own judgments about what they ought to do, as in the famous case of Phineas Gage.[19]

Each of these cases involves significant volitional impairment in which agents experience considerable difficulty implementing or conforming to the normative judgments they form.

Notice that recognition of a volitional dimension of normative competence argues against purely cognitive conceptions of insanity, such as the

[18] Mele understands a desire as conquerable when one can resist it and as circumventable when one can perform an action that makes acting on the desire difficult or impossible (1990). The alcoholic who simply resists cravings conquers his impulses, whereas the alcoholic who throws out his liquor and stops associating with former drinking partners or won't meet them at places that serve alcohol circumvents his impulses. Conquerability is mostly a matter of will power, whereas circumventability is mostly a matter of foresight and strategy.

[19] Phineas Gage was a nineteenth-century railway worker who was laying tracks in Vermont and accidentally used his tamping iron to tamp down a live explosive charge, which detonated and shot the iron bar up and through his skull. Though he did not lose consciousness, over time his character was altered. Whereas he had been described as someone possessing an "iron will" before the accident, afterward he had considerable difficulty conforming his behavior to his own judgments about what he ought to do. The story of Phineas Gage is related, and its larger significance explored, in Damasio 1994.

M'Naghten test, which recognizes only cognitive deficits as the basis for insanity, and in favor of the more inclusive *Model Penal Code* conception.

> Mental Disease or Defect Excluding Responsibility: (1) A person is not responsible for criminal conduct if at the time of such conduct as the result of a mental disease or defect he lacks substantial capacity either to appreciate the criminality [wrongfulness] of his conduct or *to conform his conduct to the requirements of law.* (2) [T]he terms "mental disease or defect" do not include an abnormality manifested only by repeated criminal or otherwise anti-social conduct.[20]

The *Model Penal Code* conception of insanity is an important advance on the *M'Naghten* conception, precisely because it recognizes an independent volitional dimension to sanity and so recognizes a wider conception of insanity as involving significant impairment of *either* cognitive or volitional competence.

Recognizing the volitional dimension of normative competence may require revising the rationality or practical reason conceptions of responsibility employed by criminal law theorists such as Michael Moore and Stephen Morse. Strictly speaking, rationality conceptions of normative competence need not reject the volitional dimension of normative competence. There might be more to rationality than correct belief or knowledge. For instance, one might not count as practically rational unless one's appetites and passions are sufficiently under control to enable one to conform one's will to one's normative judgment.

As far as we can tell, Moore is noncommittal on this issue and could agree with these claims about the importance of the volitional dimension of normative competence, folding them into claims about rational capacities. However, Morse is skeptical about the volitional dimension of normative competence. In part because his skepticism finds echoes in Fischer and Ravizza's treatment of reasons-reactivity, it is worth considering his complaints about the volitional dimension in some detail.

In his essay "Uncontrollable Urges and Irrational People," Morse critically discusses proposals to treat wrongdoers with irresistible impulses as

[20] American Law Institute, *Model Penal Code* §4.01, emphasis added. The *Model Penal Code* is a model statutory text of fundamental provisions of the criminal law, first developed by the American Law Institute in 1962 and subsequently updated in 1981. The MPC was intended to serve as a model for local jurisdictions drafting and revising their criminal codes. Notice three differences between MPC and *M'Naghten*: (a) unlike *M'Naghten*, MPC includes *volitional, as well as cognitive,* capacities in its conception of insanity; (b) whereas *M'Naghten* makes complete incapacity a condition of insanity, MPC makes *substantial incapacity* a condition of insanity; and (c) whereas *M'Naghten* requires only capacity for normative recognition for sanity, MPC requires capacity for normative *appreciation*. Here, we focus only on (a), but all three points of contrast between MPC and *M'Naghten* are potentially significant.

excused for lack of control. He claims, not implausibly, that many with emotional or appetitive disorders are nonetheless responsible, because they retain sufficient capacity for rationality (2002: 1040). In discussing excuses that appeal to uncontrollable urges, he makes clear that his conception of rationality excludes volitional components.

> This . . . Essay claims that our ambivalence about control problems is caused by a confused understanding of the nature of those problems and argues that control or volitional problems should be abandoned as legal criteria [for excuse] (2002: 1054).

But why should we abandon a volitional dimension to normative competence and control? Morse focuses on the alleged threat posed by irresistible urges and makes several (incompatible) claims about them: (1) we cannot make sense of irresistible urges, (2) we cannot distinguish between genuinely irresistible urges and urges not resisted, (3) there are no irresistible urges, because under sufficient threat of sanction we can resist any strong urge.

Morse focuses on irresistible urges. This is already problematic, because it ignores the varieties of volitional impairment, which include not just irresistible urges but also paralyzing fears, depression, and systematic weakness caused by damage to the prefrontal cortex.

But consider what Morse does say about irresistible urges. He argues against the claim made by the majority in *Kansas v. Crane* that civil detention be limited to those who are dangerous to themselves or others on account of control problems that are the result of mental abnormality.[21] Morse plausibly claims that mental disease or abnormality, as such, is irrelevant to excuse (2002: 1034, 1040). All that mental abnormality signals is something about the cause of urges; by itself, it does not signify anything about the agent's capacities, and so cannot serve as an excuse (2002: 1040). That is surely right, but the Court in *Crane* did not say that mental abnormality was sufficient for excuse, but at most that it was necessary.[22] What was critical, the Court claimed, was whether the urges were sufficiently irresistible to present a control problem. A control problem can be understood as a lack of relevant volitional capacities. So the Court is just not making the fallacious argument that Morse rightly criticizes. Demonstrating that abnormality does not imply incapacity does not show that responsibility does not require volitional capacity. So Morse's criticism of the abnormal cause requirement does not support a rationality conception of agency that eschews volitional capacities.

[21] *Kansas v. Crane*, 534 US 407 (2002).

[22] Insofar as the Court is requiring a mental abnormality, perhaps defined on the disease model, we disagree. It is neither a necessary nor sufficient condition for excuse.

Morse goes on to claim that the idea of irresistible urges is not coherent and that, even if it was, we could not distinguish between irresistible urges and urges not resisted (1994: 1601, 2002: 1062). This is the problem of distinguishing between can't and won't. Finally, he asserts that even if we could distinguish between irresistible urges and urges not resisted, we would find that in actual cases the urges in question would be resistible. In discussing whether an addict's cravings are irresistible, Morse argues that they are not because if you hold a gun to the addict's head and tell him that you'll shoot him if he gives in, he can resist (2002: 1057–8, 1070). This is reminiscent of the sort of weak reactivity that Fischer and Ravizza defend and that Kant requires in *The Critique of Practical Reason.*

> Suppose that someone says his lust is irresistible when the desired object and opportunity are present. Ask him whether he would not control his passion if, in front of the house where he had this opportunity, a gallows were erected on which he would be hanged immediately after gratifying his lust. We do not have to guess very long what his answer would be. (Kant 1788: 30)

But these different complaints about irresistible urges are all resistible.

First, there seems to be no conceptual problem with irresistible urges. We can conceive of paralyzing emotions or irresistible desires, as Mele does, as emotional states or appetites that stand in the way of implementing the verdicts of practical reason that are virtually unconquerable and uncircumventable (1990). Resistibility is a modal notion. There is a question about how unconquerable or uncircumventable impulses must be to be excusing, and there may be evidential or pragmatic problems about identifying desires that are genuinely irresistible. But the concept of irresistible desires does not seem especially problematic.

Second, consider the worry that we cannot reliably distinguish between an inability to overcome and a failure to overcome such obstacles. First of all, this is an evidentiary problem, not a claim about the ingredients of normative competence. Moreover, this evidentiary problem seems no worse than the one for the cognitive dimension of normative competence, which requires us to distinguish between a genuine inability to recognize something as wrong and a failure to form correct normative beliefs or attend to normative information at hand. Making the distinction between can't and won't is a challenge, but not an insurmountable one, in either the cognitive or volitional case. For instance, there are neurophysiological tests for various forms of affective, as well as cognitive, sensitivity, such as electrodermal tests of empathetic responsiveness (Blair et al. 2005: 49–50).

Finally, consider Morse's claim that volitional capacity is easily demonstrated insofar as agents can always resist desires and temptations under sufficient threat. Morse's position here bears comparison with that of

Fischer and Ravizza. As we saw earlier, while they defend moderate reasons-receptivity, they require only weak reasons-reactivity.[23] In defense of weak reactivity, Fischer and Ravizza claim that reactivity is "all of a piece"—if you can conform in some cases, even one case, that shows that you can conform in any case. (1998: 73). Kant and Morse seem to agree. There are two problems here. First, they want to recognize an asymmetry between cognitive and volitional capacities. Yet, if reactivity were "all of a piece," then why not say the same thing about receptivity? If one can recognize some moral reasons, one can recognize any. Or if one can recognize them under some circumstances, then one can recognize them under any. This would be to accept weak reasons-receptivity, which both Morse and Fischer and Ravizza reject. Second, they are committed to claiming, at least about reasons-reactivity, that one can't have weak responsiveness without having moderate responsiveness. Anyone who can resist an urge in one extreme situation can resist it in others. But we see no reason to accept this psychological stipulation. An agoraphobe might have such a paralyzing fear of public spaces that she would be induced to leave her home only under imminent threat of death. There's no reason to assume that we cannot have weak reactivity without moderate reactivity (cf. McKenna 2001, Watson 2001, Pereboom 2006, and Todd and Tognazzini 2008).

Our own view is that weak reactivity is simply implausible as a general reactivity condition on responsibility. Cases in which a person would only react differently under a threat of imminent death, because of a paralyzing fear or compulsion, for example, seem to be cases in which we should excuse.[24] If a desire is really only resistible in this one counterfactual case, then we think that the agent is not responsible, or at least not fully responsible, in the actual case. That doesn't mean that we can't detain him if he is dangerous to himself or others, but it would mean that it would be inappropriate to blame and punish him.

On closer inspection, it seems Morse is really ambivalent between two different kinds of skepticism about the volitional dimension of normative competence and its significance. In some moments, he denies that there is any separate volitional dimension to normative competence, beyond the cognitive dimension. At other times, he recognizes the need for a separate

[23] It is worth noting that Fischer's and Ravizza's view is *doubly* asymmetric insofar as they require receptivity to at least some *moral* reasons, but require reactivity only to reasons in general, not necessarily moral ones (1998: 79). We reject this sort of asymmetry, as well.

[24] Mele's example of the agoraphobe, who will not leave his house, even for his daughter's wedding, but would leave it if it were on fire, seems coherently described as one in which someone is weakly reactive, but nevertheless, not responsible.

volitional dimension but claims that it is easily satisfied because volitional conformity to what one judges right and wrong is “all of a piece.” We hope to have shown that neither form of skepticism is especially promising.

5. SITUATIONAL CONTROL

An important part of an agent’s being responsible for wrongdoing that she chose and intended consists in her being a responsible agent. This we have conceptualized in terms of normative competence and analyzed into cognitive and volitional capacities. Evidence for this view is that one seems to have an excuse, whether complete or partial, if one’s normative competence is compromised in significant ways. The most familiar kinds of excuse—insanity, immaturity, and uncontrollable urges—all involve compromised normative competence.

But there is more to an agent being culpable or responsible for her wrongdoing than her being responsible and having intentionally engaged in wrongdoing. Moreover, excuse is not exhausted by denials of normative competence. Among the factors that may interfere with our reactive attitudes, including blame and punishment, are *external* or *situational* factors. In particular, *coercion* and *duress* may lead the agent into wrongdoing in a way that nonetheless provides an excuse, whether full or partial. The paradigm situational excuse is coercion by another agent, as when one is threatened with physical harm to oneself or a loved one if one doesn’t assist in some kind of wrongdoing, for instance, driving the getaway car in a robbery. Though criminal law doctrine focuses on threats that come from another’s agency, hard choice posed by natural forces seems similarly exculpatory, as in Aristotle’s famous example of the captain of the ship who must jettison valuable cargo in dangerous seas caused by an unexpected storm (*NE* 1110a9–12). Situational duress does not compromise the wrongdoer’s status as a responsible agent and does not challenge her normative competence, but it does challenge whether she is responsible for her wrongdoing.

The details of duress are tricky. Some situational pressures, such as the need to choose the lesser of two evils, may actually *justify* the agent’s conduct, as is recognized in *necessity* defenses. If the balance of evils is such that the evil threatened to the agent is worse than the evil involved in her wrongdoing, then compliance with the threat is justified. But in an important range of cases, coercion and duress seem not to justify conduct (remove the wrongdoing) but rather to *excuse* wrongdoing, in whole or in part. In such cases, where the evil threatened is substantial but less than that contained in the wrongdoing, the agent’s wrongdoing should be excused because the threat or pressure was more than a person could or should be expected to resist.[25] The *Model Penal Code* adopts a reasonable person version of the conditions under which a threat excuses, namely, when a person of reasonable firmness would have been unable to resist, provided the actor was not himself responsible for being subject to duress (section 2.09).

Whereas the situational aspect of responsibility was recognized by classical writers, such as Aristotle, Hobbes, and Locke, it has been less prominent in more recent philosophical discussions of responsibility. Perhaps because of case law and doctrine involving duress, criminal theorists, such as Moore, and Morse, have clearly recognized the importance of the situational component of responsibility (Moore 1997: 554, 560–1, Morse 1994: 1605, 1617, and Morse 2002: 1058). They explain the rationale for this situational component and the associated excuse of duress in terms of the *fair opportunity to avoid wrongdoing*. The idea is that normatively competent agents, through no fault of their own, due to external threat or hard choice may lack the fair opportunity to avoid wrongdoing. In normal cases, this opportunity may just blend into the background, taken for granted. But in cases of duress it is absent. This not only explains why duress should be excusing but also alerts us to the importance of this opportunity in the normal case, where duress is absent.

6. TWO MODELS OF NORMATIVE COMPETENCE AND SITUATIONAL CONTROL

We think that this emerging picture of the architecture of responsibility in which normative competence and situational control are the two main elements of responsibility is quite attractive. Others have thought so too.

[25] Exactly when duress justifies and when it excuses is an interesting and difficult question. Much will depend on how the necessity and balance of evils doctrines are understood. Suppose A threatens to rape B's loved one if B doesn't kill C, who is innocent. On one interpretation, this case fails the balance of evils test (murder is worse than rape), so it tends to excuse, rather than justify. But this may be less clear if the balance of evils test is performed using a moral balance employing agent-centered prerogatives. We can't get a clear handle on the difference between duress justifications and duress excuses until we fix the moral conception employed in the balance of evils test. But we think that it is safe to assume that however exactly the lines are drawn on these issues about interpreting the balance of evils test there will be some duress excuses. That is especially plausible, we think, when we recognize that duress and excuse can be scalar. That is sufficient to justify our architectural assumption that there should be a separate wing for situational control, whether or not that wing is densely populated.

For example, in "Negligence, Mens Rea, and Criminal Responsibility" H. L. A Hart seems to accept such a view.

> What is crucial is that those whom we punish should have had, when they acted, the normal capacities, physical and mental, for abstaining from what it [the law] forbids, and a fair opportunity to exercise these capacities. (Hart 1961: 152)

And Moore endorses this idea.

> Hart thus subdivides the ability presupposed by his sense of 'could' into two components. One relates to the equipment of the actor: does he have sufficient choosing capacity to be responsible? The other relates to the situation in which the actor finds himself: does that situation present him with a fair chance to use his capacities for choice so as to give effect to his decision? (Moore 1997: 554)

We see two different conceptions of how these two factors relate to responsibility.

On one conception, normative competence and situational control are individually necessary and jointly sufficient but independent factors in responsibility. On this conception, there is an appropriate degree of competence and an appropriate degree of situational control that can be fixed independently of each other and which are both necessary for responsibility, such that falling short in either dimension is excusing. On this picture, we assess an agent in each area separately. We figure out whether she had the relevant capacities (e.g. were they "normal" or "sufficient"), and then we figure out whether she had the fair opportunity to exercise them.

This has been the conception of the architecture of responsibility that we have articulated so far. However, an alternative conception of normative competence and situational control is possible that treats them as individually necessary and jointly sufficient but at least sometimes interacting. On this picture, how much and what sort of capacities one needs can vary according to situational features. So, for example, there might be situations in which the wrongdoing in question was especially clear, such as a murder or an assault, and in which there was no significant provocation, duress, or other hard choice. We might think that culpable wrongdoing, in such cases, requires less in the form of cognitive or volitional capacities than in cases in which the normative issues are less clear or in which there is substantial provocation or duress. Or hold constant the wrongdoing in question and compare the interaction of situational factors and competence in different individuals. It's plausible to suppose that normative competence requires an ability to make one's own normative judgments and hold to them despite temptation, distraction, and peer pressure. It's also plausible to suppose that adolescents have less independence of judgment and ability to resist peer pressure than their adult counterparts. But then we might be more excusing of adolescent wrongdoing committed in groups than of comparable adult behavior committed in groups. If so, differential competence may explain why the same level of situational pressure may be excusing for some and not for others (Brink 2004). But then we might prefer a conception of normative competence and situational control, in which they are potentially interacting, rather than independent, dimensions of responsibility. Such a conception would also imply that the requisite levels of normative competence and situational control are not invariant, but rather context-dependent.

7. RESPONSIBILITY AS THE FAIR OPPORTUNITY TO AVOID WRONGDOING

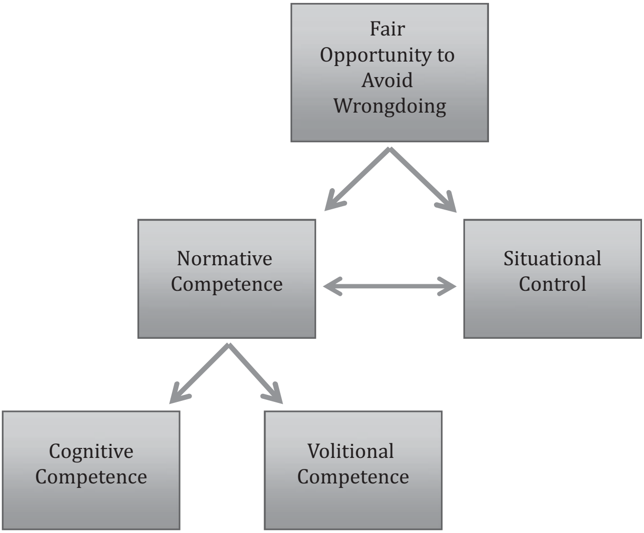
So far, the conception of responsibility emerging from our discussion is a two-factor model twice over. We factor responsibility into normative competence and situational control, and we factor normative competence into partially independent cognitive and volitional capacities, which we treat as equally important. This kind of two-factor model seems plausible, in significant part because it promises to fit our practices of excuse in both moral assessment and the criminal law pretty well. Perhaps this is adequate justification. But it would be nice if there were some unifying element to its structure.

One possible umbrella concept is *control* (the concept Fischer and Ravizza associate with reasons-responsiveness). Freedom from coercion and duress, cognitive competence, and volitional competence all seem to be aspects of an agent's ability to control her actions. But control seems important, at least in part, because it seems *unfair* to blame agents for outcomes that are outside their control.

But this suggests that the umbrella concept should perhaps be *fairness*, in particular, the fair opportunity to avoid wrongdoing (the concept that Moore and Morse associate with the situational component), because failure of either normative competence or situational control violates the norm that blame and punishment be reserved for those who had a fair opportunity to avoid wrongdoing.

A related proposal would be to treat the umbrella concept as the fair opportunity to avoid *blame*. For, we might also say that fairness requires that we not blame those who lacked normative competence or situational control. While the opportunity to avoid blame is no doubt a by-product of the fair opportunity to avoid wrongdoing, we think that it is the latter concept that should be primary in our understanding of responsibility. For one thing, if we understand responsibility (as accountability) as the condition that justifies blame, then appeal to the fair opportunity to avoid blame threatens to introduce circularity into our account. As we would then be explaining what makes actions potentially blameworthy in terms of the fair opportunity to avoid blame. That seems like too small an explanatory circle. Moreover, making the primary focus on blame seems potentially narcissistic—too focused on the agent's moral ledger. It may be that one can avoid blame if one has the fair opportunity to avoid wrongdoing, but being responsible does not seem to be primarily about the agent avoiding blame. Rather, the ability to avoid blame seems consequential on responsibility. We can explain this, without focusing unduly on the responses of others, if we treat the fair opportunity to avoid wrongdoing as the primary concept.

If we treat the fair opportunity to avoid wrongdoing as the key to responsibility, we get the following picture of the architecture of responsibility.



One way to see the importance of the fair opportunity to avoid wrongdoing is to think about why strict liability is problematic. Strict liability offenses are those for which wrongdoing is sufficient and culpability is not required. While tort law recognizes some strict liability offenses, it is significant that the penalties for such torts are civil monetary damages. Strict liability is extremely problematic within the criminal law, where guilt results in blame and imprisonment. Indeed, the main strict liability crime is statutory rape (consensual intercourse by an adult male with an underage female), where a reasonable belief that the female was not a minor is not exculpatory. But this conception of statutory rape is anomalous. Statutory rape is not treated as a strict liability crime in all jurisdictions, and where it is so treated, it is widely viewed as morally problematic. Indeed, the *Model Penal Code* condemns criminal offenses for which conviction does not require culpability (section 2.02).[26] Why exactly are strict liability crimes so problematic?

In "Legal Responsibility and Excuses" H. L. A. Hart suggests that the criminal law conditions liability on culpability out of respect for "the efficacy of the individual's informed and considered choice in determining the future" (1957: 46). A corollary of this concern with individual autonomy is the demand for the fair opportunity to avoid wrongdoing. This principle is at work in support of the fundamental legal principle of *legality*. Legality is the doctrine that there should be no punishment in the absence of public notice of a legal requirement. The principle of legality is usually defended as part of fair notice. Ex post facto or retroactive criminal law would be unfair, because it would punish those for failing to conform to behavioral expectations of which they had not been apprised in advance. Ex post facto law thus threatens individual autonomy and its demand of fair opportunity to avoid wrongdoing. But a similar rationale is at work against strict liability crimes. Just as it would be unfair to convict actors for failing to conform to standards that had not been promulgated in advance, so too it would be unfair to convict actors for failing to conform to standards (promulgated in advance) that they did everything within their power to obey. Conviction without culpability denies the fair opportunity to avoid wrongdoing.

Some have suggested that fairness requires normative competence as a condition of responsibility.[27] Others have suggested that the fair

[26] For a brief discussion of the anomalous position of strict liability crimes within the criminal law, see Dressler 2009: ch. 11.

[27] See Wallace 1994: 109, 116, 161–6. The substance of our conception of responsibility agrees with Wallace on many points, including the idea that responsibility requires normative competence, that normative competence has both cognitive and volitional dimensions, and that normative requirement is a condition of the fairness of blame and punishment. But there are important differences. Wallace's account of excusing and exempting conditions (chs. 5–6) fails at many points to distinguish properly between justification and excuse. Also, like Fischer and Ravizza and other philosophers, Wallace tends to focus on the kind of normative competence that fairness requires for responsibility, thus ignoring or at least underestimating the importance of the kind of situational control that fairness also requires for responsibility. Finally, and perhaps most importantly, Wallace insists on adopting a response-dependent interpretation of this conception of responsibility insofar as he claims that

opportunity to avoid wrongdoing requires situational control. We put these ideas together and conclude that the guiding principle underlying our conception of responsibility is the fair opportunity to avoid wrongdoing.

Because the fair opportunity to avoid wrongdoing is the central value in our conception of responsibility, it is perhaps worth addressing skepticism about the fair opportunity to avoid wrongdoing. There is a familiar kind of skepticism about responsibility that appeals to incompatibilism and determinism. One form of skepticism alleges that responsibility presupposes the Principle of Alternative Possibilities (PAP), requiring that one could have done other than one did, which determinism shows to be false. Our conception of responsibility is not immune to these sorts of skeptical threats, because we understand responsibility in terms of the fair opportunity to avoid wrongdoing and we understand the fair opportunity to avoid wrongdoing to require alternative possibilities. To have the fair opportunity to avoid wrongdoing, it must be true that when one commits wrong, one could have done otherwise. There are two kinds of challenge to PAP. First, PAP may seem to require indeterminism. Second, alternative possibilities may seem unnecessary for responsibility because of Frankfurt-style counterexamples in which an agent chooses to act in a normal and responsible way and in which someone else (the counterfactual intervener) stands ready to intervene to cause her to act that way in case she should waver (Frankfurt 1969).

These are familiar philosophical issues about responsibility that are quite complicated, and this is not the place to defend compatibilism in any detail. But we can and should indicate the lines along which we would like to defend our commitments. Our response is to understand the "could have done otherwise" clause in PAP as consistent with both the existence of typical Frankfurt scenarios and with determinism. This requires a compatibilist understanding of the ability to do otherwise. A variety of such compatibilist accounts have recently been defended.[28] The account we favor is one according to which one has the general ability to do otherwise, manifested in various cognitive and volitional capacities and opportunities,

we cannot give an account of the conditions under which an agent is responsible independently of our reactive attitudes and practices of excuse (1994: 19). But while the reactive attitudes and our practices of excuse may be *evidence* about the conditions of responsibility, on our conception, they point to a conception of responsibility grounded in the fair opportunity to avoid wrongdoing, which itself requires normative competence and situational control. On this conception, it is possible to give a response-independent account of the conditions under which the reactive attitudes are justified.

[28] For example, there have been several attempts to reconcile PAP with determinism by understanding the ability to do otherwise in terms of the having of a disposition. See, for example, Smith 2003, Vihvelin 2004, Fara 2008. For criticism, see Clarke 2009.

and, at the time of the action, nothing impairs these capacities and opportunities in a relevant way. Such a condition can be satisfied in Frankfurt-style cases, and, we claim, in deterministic settings.[29] There is, to be sure, another sense of the ability to do otherwise, in which the agents in Frankfurt-style cases lack such an ability, but this is not the sense we believe is required by the fair opportunity to avoid wrongdoing. The legitimate threats to the opportunity to avoid wrongdoing and, hence, responsibility come not from the fact that our actions are caused but rather from specific impairments of cognitive and volitional capacities and specific kinds of external interference with our ability to conform to the relevant norms.

8. PARTIAL RESPONSIBILITY AND EXCUSE

An important feature of our conception of responsibility is that the components of normative competence and situational control are *scalar* concepts that admit of degree. This implies that responsibility is itself scalar and that *partial* responsibility is a real phenomenon.[30] Insanity and immaturity are clearly scalar concepts. But so are the ideas of paralyzing emotions, irresistible urges, and a disabling depression. So too is the hardness of a choice, posed by coercion or natural forces, and the question of whether it is more than a reasonable person could resist. Partial responsibility is reasonably well recognized within our assessments of moral responsibility. But the recognition of partial responsibility suggests a basis for criminal jurisprudence reforms.

Criminal law trials are *bifurcated* into two phases—the *guilt phase* and, if guilty, the *sentencing phase*. Verdicts at the guilt phase are generally *bivalent*—guilty or not guilty. There are various possible reasons for a finding of not guilty. The prosecution may fail to prove some element of *actus reus* or *mens rea* beyond a reasonable doubt, or the defense may provide adequate evidence of justification, by demonstrating necessity, or excuse, by demonstrating duress or insanity. But the factors determining both justification and excuse would appear to be scalar. One might conclude that in an ideal world guilt

[29] The skeptic might reply that a determined past prevents one from exercising one's capacities. But, as Nelkin has argued (2011: 64–79), the fact that the past is deterministic does not make it interference or prevention in the relevant sense any more than is the counterfactual intervener in a Frankfurt scenario. See Wolf (1990: 94–116) for a different kind of defense of the conclusion.

[30] One recent discussion of the scalar character of reasons-responsiveness is Coates and Swenson 2012. For a defense of the scalar character of immaturity and the resulting partial responsibility appropriate for juvenile justice, see Brink 2004.

verdicts should be *multivalent* reflecting the degree of justification or excuse the defendant possessed. Insofar as our bivalent system recognizes only perfect necessity as a justification and complete or substantial lack of normative competence as an excuse, its conception of guilt and punishment is overinclusive and punishment is not proportional to just deserts. Perhaps there are legitimate pragmatic reasons for opposing an infinitely multivalent system. But even a *trivalent* system that recognized *partial guilt*, as well as guilty and not-guilty verdicts, would provide more retributive justice.

Justifications do not bear on responsibility, as excuses do. So we focus on partial excuse, rather than partial justification. While some European criminal justice systems recognize partial excuse, American criminal law does not. Instead, we recognize only full excuses, relegating considerations that are relevant to excuse but fall short of full excuse to the sentencing phase of a trial.[31] Some commentators have called for the adoption of a trivalent system for the guilt phase in which the accused can be found guilty, not guilty, or guilty but partially responsible.[32]

A trivalent system might be superior to our current bivalent one. Supplementing a bivalent system with sentence mitigation is an imperfect proxy for recognizing partial excuses. Though one can mitigate sentences to respond to various nonculpability factors (e.g. mercy or forgiveness), mitigation is a poor response to reduced culpability. First, mitigation is not always possible where there are mandatory sentencing rules. Second, mitigation, where possible, is largely discretionary. But it shouldn't be discretionary whether to hear and evaluate considerations that would tend to excuse. If complete incompetence is relevant to the guilt phase and a full excuse, then partial competence is also relevant to the guilt phase and a partial excuse, because it appeals to the very same factors that a full excuse appeals to, only to a reduced degree. In effect, the claim is that excuses, whether full or partial, belong to the guilt phase of a trial and that sentence mitigation should be confined to nonculpability considerations (or, perhaps, in a trivalent system, to those culpability considerations that fall below the threshold necessary for partial excuse).

Our conception of the architecture of responsibility recognizes the scalar nature of the components of responsibility and so justifies findings of partial responsibility. If, as the positive retributivist believes, punishment and the reactive attitudes should be proportional to culpable

[31] The one exception to the generalization that American criminal law does not recognize partial excuse is the provocation defense, under which intentional homicides committed with adequate provocation reduce to an offense of voluntary manslaughter.

[32] See Morse 2003. Morse's position is especially interesting, because this defense of partial excuse is a reversal of his previous skepticism about partial excuse (Morse 1984).

wrongdoing, then just deserts require that we recognize partial excuses. Otherwise, we punish more than just deserts. As we have seen, not all retributivists are of this sort. Yet all retributivists, positive and negative, believe that punishment and the reactive attitudes should not be *out of* proportion to culpable wrongdoing. If this is correct, then justice requires that we recognize partial excuses. Presumably, our moral assessments and practices of moral blame can and do already reflect recognition of partial excuses, albeit imperfectly. The conception of responsibility we have sketched suggests that we should recognize a doctrine of generic partial excuse in the criminal law as well and abandon that doctrine only if and when it proves administratively unfeasible or to have worse retributive outcomes overall.

REFERENCES

Aharoni, E., Sinnott-Armstrong, W., and Kiehler, K. (2012). "Can Psychopathic Offenders Discern Moral Wrongs? A New Look at the Moral/Conventional Distinction." *Journal of Abnormal Psychology* 121: 484–97.

American Law Institute. (1981). *Model Penal Code*, ed. M. Dubber (New York: Foundation Press, 2002).

Aristotle. (*NE*) *Nicomachean Ethics*. Trans. T. Irwin (Indianapolis, IN: Hackett, 1985).

Bedau, H. and Kelly, E. (2010). "Punishment." *The Stanford Encyclopedia of Philosophy* <http://plato.stanford.edu/archives/spr2010/entries/punishment>.

Blair, J., Mitchell, D., and Blair, K. (2005). *The Psychopath: Emotion and the Brain* (Oxford: Blackwell).

Brink, D. (2004). "Immaturity, Normative Competence, and Juvenile Transfer: How (Not) to Punish Minors for Major Crimes." *Texas Law Review* 82: 1555–85.

Clarke, R. (2009). "Dispositions, Abilities to Act, and Free Will: The New Dispositionalism." *Mind* 118: 323–51.

Coates, D.J. and Swenson, P. (forthcoming). "Reasons-Responsiveness and Degrees of Responsibility." *Philosophical Studies*.

Damasio, A. (1994). *Descartes' Error: Emotion, Reason, and the Human Brain* (New York: Putnam).

Dressler, J. (2009). *Understanding Criminal Law*, 5th edn. (Boston: LexisNexis).

Duff, A. (2008). "Legal Punishment." *The Stanford Encyclopedia of Philosophy* <http://plato.stanford.edu/archives/fall2008/entries/legal-punishment>.

Fara, M. (2008). "Masked Abilities and Compatibilism." *Mind* 117: 843–65.

Fine, C. and Kennett, J. (2004). "Mental Impairment, Moral Understanding, and Criminal Responsibility." *International Journal of Law and Psychiatry* 27: 425–43.

Fischer, J. and Ravizza, M. (1998). *Responsibility and Control* (New York: Cambridge University Press).

Frankfurt, H. (1969). "Alternate Possibilities and Moral Responsibility," reprinted in Frankfurt (1988).

—— (1988). *The Importance of What We Care About* (New York: Cambridge University Press).

Hart, H L. A. (1957). "Legal Responsibility and Excuses," reprinted in Hart (1968).

—— (1961). "Negligence, Mens Rea, and Criminal Responsibility," reprinted in Hart (1968).

—— (1968). *Punishment and Responsibility* (Oxford: Clarendon Press).

Kant, I. (1788). *Critique of Practical Reason.* Trans. L.W. Beck (Indianapolis, IN: Bobbs-Merrill, 1956) (Prussian Academy pagination).

Kripke, S. (1982). *Wittgenstein on Rules and Private Language* (Cambridge, MA: Harvard University Press).

McKenna, M. (2001). "Review of Fischer and Ravizza, *Responsibility and Control*." *Journal of Philosophy* 98: 93–100.

—— (2012). "Reasons-Responsiveness, Agents, and Mechanisms" (this volume).

Mele, A. (1990). "Irresistible Desires." *Noûs* 24: 455–72.

Moore, M. (1997). *Placing Blame* (Oxford: Clarendon Press).

Morse, S. (1984). "Undiminished Confusion in Diminished Capacity." *The Journal of Criminal Law and Criminology* 75: 1–55.

—— (1994). "Culpability and Control." *University of Pennsylvania Law Review* 142: 158–660.

—— (2002). "Uncontrollable Urges and Irrational People." *Virginia Law Review* 88: 1025–78.

—— (2003). "Diminished Rationality, Diminished Responsibility." *Ohio State Journal of Criminal Law* 1: 289–308.

Nelkin, D. (2011). *Making Sense of Freedom and Responsibility* (Oxford: Clarendon Press).

Pereboom, D. (2006). "Reasons-Responsiveness, Alternative Possibilities, and Manipulation Arguments against Compatibilism." *Philosophical Books* 47: 198–212.

—— (2012). "Freedom and Punishment." In *The Future of Punishment*, ed. T. Nadelhoffer (Oxford: Clarendon Press).

Scanlon, T. M. (2008). *Moral Dimensions: Permissibility, Meaning, and Blame* (Cambridge, MA: Harvard University Press).

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

Smith, M. (2003). "Rational Capacities, or How to Distinguish Recklessness, Weakness, and Compulsion." In *Weakness of Will and Practical Irrationality*, eds. S. Stroud and C. Tappolet (Oxford: Clarendon Press).

Strawson, P. F. (1962). "Freedom and Resentment," reprinted in Watson (1982).

Todd, P. and Tognazzini, N. (2008). "A Problem for Guidance Control." *Philosophical Quarterly* 58: 685–92.

Vihvelin, K. (2004). "Free Will Demystified: A Dispositional Account." *Philosophical Topics* 32, 427–50.

Wallace, R. J. (1994). *Responsibility and the Moral Sentiments.* (Cambridge, MA: Harvard University Press).

Watson, G., (ed.) (1982). *Free Will* (New York: Oxford University Press).

—— (1987). "Responsibility and the Limits of Evil," reprinted in Watson (2004).

—— (1996). "Two Faces of Responsibility," reprinted in Watson (2004).

—— (2001). "Reasons and Responsibility," reprinted in Watson (2004).

—— (2004). *Agency and Answerability* (New York: Oxford University Press).

—— (2011). "The Trouble with Psychopaths." In *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon* (New York: Oxford University Press).

Williams, B. (1976). "Moral Luck," reprinted in Williams (1981).

—— (1981). *Moral Luck* (Cambridge: Cambridge University Press).

Wolf, S. (1990). *Freedom Within Reason* (New York: Oxford University Press).

This page intentionally left blank

Index