OXFORD

oxford studies in normative ethics

volume 3

Oxford Studies in Normative Ethics

Oxford Studies in Normative Ethics

Volume 3

EDITED BY

MARK TIMMONS

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

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2013938356

ISBN 978–0–19–968591–2

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

Contents

Acknowledgments *vii*
List of Contributors *viii*

Introduction 1
MARK TIMMONS

1 Supererogation and Virtue 13
ROGER CRISP

2 The Weight of Moral Reasons 35
RALPH WEDGWOOD

3 Scanlon's Promising Proposal and the Right Kind of Reasons to Believe 59
MARK VAN ROOJEN

4 Evil Achievements and the Principle of Recursion 79
GWEN BRADFORD

5 Self-Ownership and the Conflation Problem 98
DAVID SOBEL

6 Consequentializing and Deontologizing: Clogging the Consequentialist Vacuum 123
PAUL HURLEY

7 On Criminal and Moral Responsibility 154
DAVID SHOEMAKER

8 Consequentialism, Cognitive Limitations, and Moral Theory 179
DALE DORSEY

9 They Can't Take That Away From Me: Restricting the Reach of Morality's Demands 203
Sarah Stroud

10 What We Know and What We Owe 235
Vanessa Carbonell

11 Objective Double Effect and the Avoidance of Narcissism 260
Howard Nye

12 Some Intellectual Aspects of the Cardinal Virtues 287
Paul Bloomfield

Index 315

Acknowledgments

Versions of the articles in this collection were presented at the third annual Arizona Workshop in Normative Ethics that took place in Tucson on 5–7 January 2012. I would like to thank the Freedom Center and the Department of Philosophy at the University of Arizona for their continuing generous financial support of the workshop. Thanks to Brandon Warmke who assisted with workshop details, and to Sarah Raskoff who created the volume's index. I would also like to express my sincere thanks to the following philosophers for serving as *de facto* program referees: Matt Bedke, Christian Coons, Julia Driver, William FitzPatrick, Doug Portmore, and Daniel Star. Drew Schroeder and a referee who wishes to remain anonymous each composed an excellent set of comments on the penultimate versions, and their incisive comments and suggestions contributed importantly to the excellence of this volume's essays. Thanks finally to Peter Momtchiloff, my editor at Oxford University Press, for his support and good advice.

Mark Timmons
Tucson, Arizona

List of Contributors

Paul Bloomfield Professor of Philosophy at the University of Connecticut

Gwen Bradford Assistant Professor of Philosophy at Rice University

Vanessa Carbonell Assistant Professor of Philosophy at the University of Cincinnati

Roger Crisp Uehiro Fellow and Tutor in Philosophy at St Anne's College, Oxford and Professor of Moral Philosophy at the University of Oxford

Dale Dorsey Associate Professor of Philosophy at the University of Kansas

Paul Hurley Sexton Professor of Philosophy at Claremont McKenna College

Howard Nye Assistant Professor of Philosophy at University of Alberta

David Shoemaker Associate Professor in the Department of Philosophy & Murphy Institute at Tulane University

David Sobel Guttag Professor of Ethics and Political Philosophy at Syracuse University

Sarah Stroud Associate Professor of Philosophy at McGill University

Mark van Roojen Professor of Philosophy at the University of Nebraska—Lincoln

Ralph Wedgwood Professor of Philosophy at the University of Southern California

Introduction

MARK TIMMONS

This is the third volume of *Oxford Studies in Normative Ethics*, an annual whose aim is to bring together cutting-edge articles on topics in normative ethical theory. I am pleased to present twelve articles that advance in various ways our understanding of such topics as supererogation, why be moral, promising, the value of achievements, self-ownership, consequentializing moral theories, moral and criminal responsibility, the project of moral theory and cognitive limitations, the extent of the demands of morality, the bearing of one's knowledge on one's obligations, the principle of double effect, and the intellectual components of cardinal moral virtues. What follows is a brief overview of the volume's essays.

In his 1958 seminal article "Saints and Heroes," J. O. Urmson argued that moral theory ought to recognize the commonsense moral category of the supererogatory—described as acts that are not one's duty but whose performance, when appropriately motivated, is morally best or virtuous. But is it a desideratum of a moral theory that it recognize the supererogatory? This is the question that **Roger Crisp** asks in his "Supererogation and Virtue." As Crisp notes, the category in question only became part of commonsense moral thinking through Catholicism, and so one may wonder whether Urmson's claim that any adequate moral theory must allow for acts of supererogation should be accepted. Crisp's view is that we should not follow Urmson on this matter; that the category of the supererogatory must be defended on its own merits. He proceeds to argue that Aristotle's conception of virtue does not make room for the supererogatory, given the tight connection in Aristotle's thought between duty and virtue. So those attracted to an Aristotelian conception of morality should not recognize the supererogatory. Urmson, approaching the issue from a broadly utilitarian perspective, defended the category, offering five reasons for accommodating such actions in one's moral theorizing. But Crisp argues that Urmson's defense is inadequate. One lesson bearing on moral theory that Crisp draws from his reflections on supererogation is that theorists ought not take at face value moral concepts that have found their way into our conceptual system. Moral theorists, he thinks, ought to strive for conceptual parsimony, making normative concepts earn their keep.

A perennial philosophical question in ethics is "Why be moral?", which **Ralph Wedgwood** in "The Weight of Moral Reasons" proposes to tackle by first interpreting the question as about whether the two following claims are true. (1) All normal human beings are subject to non-trivial moral requirements; and (2) whenever any agent is subject to a non-trivial moral requirement, that agent has an overriding or conclusive reason for conforming to it. In developing his proposal for defending the truth of these claims, Wedgwood employs a value-based conception of reasons for action, according to which (roughly), a fact counts as a reason for an individual to Φ iff the fact in question entails that Φ-ing is in an appropriate way a *good thing for that individual to do.* Assuming there are moral values, then on the value-based conception of reasons, one has a *moral* reason to Φ (in cases where some courses of action that are available to one would, if performed, instantiate these moral values to different degrees) iff Φ-ing is better, with respect to one of the moral values bearing on action, than the relevant benchmark. But there being such moral reasons does not show that they ground non-trivial moral requirements that are, in addition, overriding—the two crucial elements for addressing the two questions about the status of morality. Explaining why it is that *whenever* one is subject to a moral requirement one has an overriding reason to conform to it, Wedgwood argues, can be handled by distinguishing overriding moral reasons from merely sufficient moral reasons of the sort associated with supererogation: moral requirements just are those acts for which there are *overriding* moral reasons to perform. This addresses the second of the two claims about the status of morality, but according to Wedgwood it is the claim that normal human beings are subject to non-trivial moral requirements that is the more challenging claim to defend. A defense of this claim will have to explain why there are any moral reasons that override all countervailing non-moral reasons for action. Wedgwood's proposal for defending this claim involves embracing a "weighing" model of reasons and then defending the claim that when moral reasons override non-moral reasons, they do so by outweighing them. How, then, can one understand the weight of a reason for action? In response, the idea is that moral values "subsume" other non-moral values pertaining to some choice situation which then add up to a "bigger" value, and so, according to Wedgwood's tentative proposal "the fact that moral values are 'big values' in this way is at least part of what explains why moral reasons are often so weighty that they outweigh all countervailing non-moral reasons." In defending this proposal Wedgwood offers it as at least a partial explanation of why normal human beings are often subject to non-trivial moral requirements.

According to the social practice account of the duty to keep one's promises, this particular duty is explained by one's duty to comply with socially beneficial or just practices, among which is the practice of promising. Such views have trouble explaining the special "directedness" of the duty to keep one's promises; breaking a promise to an individual without a justifying reason wrongs some particular party—the would-be recipient of the promised action. T. M. Scanlon's (1990) own account of the duty to keep one's promises is not a social-practice account. On his view, the duty in question is grounded by a complex principle F that involves (in part) the promisee coming to be assured by the acts of promisor that the promisor will perform the actions being promised. Unlike the social-practice view, Scanlon's account accommodates straightforwardly the special directedness of the promising relation. However, critics have argued that this "assurance" account faces a debilitating circularity worry. Part of the ground of the obligation to keep one's promises and thus a promisor's moral reason to keep the promise is the assurance that the promisor's act of promising creates in the promisee. However, it would appear that this very assurance is what is provided by the duty to keep the promise. So, it looks as if either the assurance (a belief) needed to ground the duty is grounded in wishful thinking (which arguably is the wrong kind of reason for belief), or there must be some reason of the right kind—a reason bearing on the truth of the belief in question—that can serve to justify the promisee's belief in a way that does not depend on the promisor being put under the obligation to keep her promises. Accomplishing this would allow this rationally grounded belief to play a role in explaining the duty to keep one's promises according to Scanlon's principle F. In "Scanlon's Promising Proposal and the Right Kind of Reasons to Believe," **Mark van Roojen** argues that a way out of the apparent circularity and thus a way of understanding the kind of reason that can justify a promisee's assurance in the promisor's compliance with the promise is to take a cue from certain cooperation games and their real-life counterparts. The basic idea applied to the case of promising is that both parties share a common goal: namely, that the promisor wishes to create an expectation in the promisee that the promisee welcomes. The promisor can create this expectation by deciding to do whatever she intends to promise and communicating this to the promisee, giving the promisee some reason to think that the promisor will follow through. The rationally created expectation thus puts the promisor in the domain of Scanlon's principle F, which thereby generates a duty to keep one's promises and thus a moral reason to comply. Although van Roojen's proposal does not turn Scanlon's view into a social-practice account, it is similar to such accounts in one way: it appeals to a reason on the part of the promisor, grounded in a norm of veracity, to comply with the promise that is not itself grounded in Scanlon's principle of promising.

The Principle of Recursion in value theory holds that pursuit of a good is itself intrinsically good, while pursuit of a bad is intrinsically bad. But there are various ways to understand how this principle works. In "Evil Achievements and the Principle of Recursion," **Gwen Bradford** considers the value of evil achievements as a way of exploring how exactly the recursion principle is to be understood. With regard to the value of achievement, Bradford accepts the Process Thesis, according to which the process involved in any achievement is a source of some positive intrinsic value. But in cases of evil achievements where the product has a negative value, the principle of recursion seems to imply that the process itself is evil. If so, then the principle is in tension with the Process Thesis. In order to reconcile the recursion principle with the Process Thesis, Bradford considers various interpretations of the principle, arguing that the most plausible of them construes the principle as an instance of an organic unity—according to which the value of a whole differs from the sum of the values of its parts. On this construal, then, recursion governs the value of the whole—the process and the product as a unit—but does not alter the positive value of the process involved even in cases of evil achievements. In developing this interpretation (including use of Moore's distinction between the value of a complex *as a whole* and its value *on the whole*), Bradford considers the apparently problematic cases of unsuccessful attempts to achieve evil ends, arguing that the organic unity understanding of the recursion principle can handle such cases by appeal to the manner in which unrealized intentional objects of one's efforts can and do shape the value of wholes. Although the organic-unities interpretation of the Principle of Recursion is more complex than rival interpretations (not mentioned here), Bradford argues that it is superior to them in accommodating a wide range of intuitions about the value of achievements.

One foundation for a non-consequentialist, deontological moral theory is the libertarian self-ownership thesis, according to which, individuals enjoy powerful ownership rights over their bodies and their property, as well as rights against certain other forms of coercion that strongly protect them against interference on behalf of promoting social good. As **David Sobel** argues in his "Self-Ownership and the Conflation Problem," traditional understandings of the self-ownership thesis generate what he calls the conflation problem—the problem of treating all property infringements as equally serious so that even trivial infringements of one's property rights for the sake of significant social gains are impermissible. My trivial pollution, for example, is impermissible if it runs the risk of causing someone a very small skin irritation. Such examples can be multiplied easily. The conflation problem thus leads to unacceptable limitations on one's liberty. After exploring various attempts to deal with this problem, Sobel argues that the most attractive solution is to allow that property infringements vary in their importance and thus one is owed different levels of protection against them. This requires rejection of the All Infringements are Equal (and seriously so) thesis that is characteristic of the traditional self-ownership view. Sobel's proposal allows that trivial rights infringements in return for significant social gains may be permissible, including, for example, redistributive takings of one's property. Of course, this proposal requires that one have a plausible theory of value that can explain differences in the significance of infringements of property rights while remaining as faithful as possible to traditional libertarian commitments. So-called objective measures are not sensitive to an individual's own assessment of the importance of her rights and so, as Sobel points out, are at odds with the self-ownership view. Embracing a subjectivist theory, which ties considerations of importance to an individual's preferences, also has certain costs, including the vindication of restrictions on certain classes of actions such as freedom of conscience. Despite this cost and some others, Sobel maintains that the most plausible understanding of the rights of property owners is his Value-Sensitive Self-Ownership View.

The idea that any plausible moral theory can be formulated as a form of consequentialism—that all such theories can be consequentialized—is grounded in the idea that all morally relevant values, including those recognized by traditional opponents of consequentialism, can be included in a best-to-worst ranking of possible outcomes of actions and practices. If consequentializing succeeds, then what has traditionally been taken to be debates in moral theory between consequentialists and non-consequentialists are really best understood as debates within the consequentialist framework. **Paul Hurley**'s "Consequentializing and Deontologizing: Clogging the Consequentialist Vacuum," argues that just as all plausible moral theories can be consequentialized, so can they also be deontologized—represented as alternative forms of deontology. If so, then consequentializing cannot itself provide support for consequentialism. This suggests that the real work in arguing for a consequentialist moral framework is being carried by what are taken to be distinct advantages of consequentialism. In particular, according to the Compelling Idea it is always permissible to do what is best, and according to the Explanatory Thought a proper explanation of the deontic status of actions will appeal to a fundamental account of intrinsic value. These attractive theses arguably do constrain plausible moral theories of right conduct, and one might suppose that they are uniquely captured by consequentialist frameworks. However, Hurley argues that it is only when the Compelling Idea and the Explanatory Thought are wedded to a particular (and controversial) evaluative conception, according to which the deontic status of action is determined by the comparative value of outcomes of actions, that these two theses provide support for a consequentialist framework. But to interpret these two intuitively attractive theses as committed to this particular evaluative framework is to beg the question against deontological views that work with an alternative conception of value. The idea is that an essentially deontological framework may be understood as taking the deontic status of actions to be fully explained in terms of the value of actions and reasons for actions as good, better, and best. Considerations that determine the value of outcomes of actions are reflected in good reasons that are relevant to determining the evaluative statuses of actions. In this way, any plausible moral theory, including forms of consequentialism, may be formulated as a version of deontology that accommodates both the Compelling Idea and the Explanatory Thought, understood generically.

So, within the evaluative framework favoured by consequentialists, all plausible non-consequentialist moral theories can be consequentialized, while within the evaluative framework favoured by deontologists, all plausible non-deontological moral theories can be deontologized. If this is right, then the projects of consequentializing and deontologizing do not themselves provide non-question-begging support for either sort of view. One upshot of all this, according to Hurley, is that debates among competing moral theories are best understood as debates over competing conceptions of value, and resolving these debates will likely turn on resolving controversies over the nature of normative "ought" claims, the nature of practical reasons, and the nature of desire.

According to the Standard View, (1) criminal responsibility entails moral responsibility and (2) the elements that figure in criminal responsibility are structurally and functionally analogous to those that figure in moral responsibility, the main difference between them being a matter of the contents of legal and moral norms. **David Shoemaker** in "On Criminal and Moral Responsibility" critically examines this view. He distinguishes three conceptions of responsibility—attributability, answerability, and accountability—and argues with respect to each of them that the Standard View, if not false, at least overly simplifies the relations among these facets of responsibility. For instance, according to the attributability conception, to be responsible for an action (or attitude) is for it to flow from one's ends, commitments, or cares, while accountability concerns being responsible in the sense that one is appropriately subject to sanctions or rewards. Within ordinary morality an agent's being accountable for an action requires that it be attributatible to that agent. However, in criminal law this connection is broken in the case of strict liability laws. One might, of course, preserve the connection by eliminating strict liability laws, or explore other ways in which to preserve the connection within the law. But Shoemaker argues that however one deals with this issue, the fact remains that the relation between criminal and moral accountability is more complex than the Standard View recognizes. The answerability conception of responsibility concerns our interest in an agent's motivating reasons for performing some action. Here, Shoemaker finds important disanalogies between moral answerability and criminal answerability. For example, the former is primarily concerned about an agent's quality of will and what it reveals about that person's attitude toward those to whom he or she is answerable, while the latter concerns the different issue of how one's motivational reasons compare to the normative standards governing the action in question. Again, moral accountability is sensitive to the reasons an agent has for performing or failing to perform actions, while criminal accountability is concerned only with whether the action simply conforms to the relevant legal demands. Thus, criminal accountability does not entail moral accountability, contrary to the Standard View. Shoemaker's conclusion, based on these observations, is that the Standard view is mistaken.

Objective-act consequentialism maintains that the deontic status of an action is determined by the value of the actual consequences of the action (the consequences that would be brought about by the action were it to be performed) compared to the value of the actual consequences of alternative actions open to the agent at the time. Of course, one's ignorance of the actual consequences of one's actions hinders one's capacity to know the moral status of alternative courses of action. But recently it has been noted that the problem of cognitive limitations presents more serious problems. Agents will often find themselves in circumstances where it is possible for them to perform the action that objective-act consequentialism requires but simply will not perform the action in question owing to cognitive limitations, even in cases where the agent knows that the action in question would, if performed, produce better consequences than any alternative action she could perform. The root of the problem, according to **Dale Dorsey** in his "Consequentialism, Cognitive Limitations, and Moral Theory," concerns the traditional understanding of moral requirability—of the set of actions that are candidates for being morally required. According to the traditional understanding, actions that one can perform on some occasion are candidates for being morally required. But this understanding of requirability seems too lax. For example, although strictly speaking one can, say, write down the sequence of words that express a cure for cancer, and believe that from among the actions one could perform, writing out the cure for cancer would produce far better consequences than any alternative action one might perform instead, one simply has no clue what the cure is. But the traditional conception of requirability, together with objective-act utilitarianism, yields the very counter-intuitive verdict that one is morally required to do (for example, write down the cure for cancer) what one in fact will not do. Furthermore, the problem of moral requirability generalizes to all moral theories that take the consequences of actions into account in determining the deontic status of actions, and so is not strictly limited to forms of consequentialist moral theory. The task, then, and the one that Dorsey tackles in his contribution, is to devise a principled account of moral requirability that plausibly narrows the range of actions that are candidates for being morally required. Dorsey's proposal is what he calls the "agency view," according to which (roughly) an action is morally requirable for an agent at a time if and only if the agent can perform that action *as an agent*—on the basis of one's capacity as a deliberative agent. After spelling out his agency view, Dorsey proceeds to defend it against a variety of possible objections, arguing that, despite certain seemingly counterintuitive implications, the view deserves to be taken seriously as an account of requirability.

A seemingly plausible principle of beneficence requires that one respond (as one can) to those in need. As is well known, this seemingly innocuous principle arguably imposes very demanding requirements on those in a position to help others—so demanding, that it threatens the pursuit of one's personal projects, enjoyments, as well as the care and attention one is inclined to show for one's family and friends. **Sarah Stroud** considers proposals that would impose a principled limit on the demands of beneficence grounded in what she calls "they can't take that away from me" arguments—the namesake of her contribution. She begins by locating possible sources of the extreme demands that beneficence may impose, including: (1) agent-neutral maximizing versions of consequentialism, (2) Peter Singer's (1972) agent-neutral "prevention" principle which, although weaker than the consequentialism mentioned in (1) is still extremely imposing in its moral demands, and (3) Garrett Cullity's (2004) even more modest proposal grounded in seemingly uncontroversial rescue cases, which he then argues can be applied iteratively to generate extreme demands. Stroud then turns to versions of the "can't take that away from me" argument that we find in the writings of Bernard Williams (1973, 1976) and Barbara Herman (2001), before turning to Cullity's own, seemingly superior version which rests on the appealing idea that the very things that beneficence can require entail limits on what beneficence cannot require. The intended upshot of Cullity's argument is that beneficence does not, after all, impose extreme demands. But Stroud raises a number of worries about Cullity's attempt to deflect the extreme demands generated by requirements of beneficence, and in general voices worries about the prospects of finding a fully defensible form of the "can't take that away from me" argument. She concludes by suggesting that resisting the extreme demands that are seemingly imposed by a principle of beneficence may require showing that arguments appearing to generate extreme demands start off with assumptions that upon examination are dubious. She illustrates the point in relation to Cullity's initially compelling "life-saving analogy," featured in his argument from beneficence to extreme demands, by noting an important scope ambiguity in the argument that arguably undermines its force.

How does one's knowledge, skill, and expertise bear on one's moral obligations? This is the question addressed by **Vanessa Carbonell** in "What We Know and What We Owe." Carbonell approaches this question by focusing on a puzzle that arises from three commonsense principles about how to live our lives. According to the Life Path Freedom principle, one is not morally required to pursue the morally best career that is available, while according to the Quit Any Time principle, all else equal, one is morally permitted to quit a career or life path, thus changing course at any time. These principles presumably capture commonsense ideas about one's freedom in living one's life according to one's own plans. Both principles seem to be well motivated, but the Quit any Time Principle is in tension with the Burden of Expertise Principle: one has certain moral obligations in virtue of possessing certain knowledge, skills, or experience. After all, if once one has advanced far enough in a career or life path to have acquired certain knowledge, skills, or expertise that (together with background moral considerations) generate moral obligations, how can it be that one may avoid such obligations by just quitting one's career or changing one's life path at *any time*? After explaining how knowledge (skill, and expertise) create obligations, Carbonell proceeds to resolve the puzzle in question by rejecting the Quit Any Time principle in favour of a more nuanced principle—Quitting Without Blame—that puts restrictions on when, given the obligations created by one's chosen career or life path, one is morally permitted to give up one's career or change one's life-path. After defending her proposal, Carbonell examines the questions of whether prospective knowledge-based obligations ground obligations that one now has to pursue courses of action through which one would acquire such knowledge, and when knowledge-based obligations do take effect.

According to "subjective" interpretations of the doctrine of double effect (DDE), there are stronger moral reasons against causing or allowing harms than there are against causing or allowing harms that one foresees but does not intend—so-called side effects. "Objective" interpretations of this doctrine maintain that the differential strength in moral reasons between causing or allowing harm and causing or allowing harms that are merely foreseen is explained in terms of the objective explanatory relationship between the effects one's conduct has on those it harms and the effects one's conduct has on those it benefits. Critics of subjective interpretations argue the deontic status of action does not depend on an agent's intentions in the manner expressed by such versions of DDE, while objective versions have often been dismissed by some as having absurd implications. **Howard Nye**, in "Objective Double Effect and the Avoidance of Narcissism," defends an objective version. As he explains, one of the motivations for embracing DDE is to defend deontological constraints while avoiding the problem of "dirty hands." However, subjective versions of this doctrine, while able to avoid dirty hands, are subject to the problem that Nye calls the "dirty heart" objection to the effect that because these versions understand deontological constraints as importantly concerned about one's intentions in acting, "they embody narcissistic obsessions with our personal purity of heart." After responding to standard objections to objective interpretations of DDE, Nye proceeds to defend a version of the view meant to capture the denial of the Machiavellian claim that the end justifies the means. Stated positively, Nye's proposal is that when an act (including omissions) will benefit some, but only by imposing harms on others, the benefits do not count as reasons to perform that action in the same way as do reasons against imposing the harms; the force of the reasons to benefit are weakened by the fact that the action in question will impose the harms in question.

In our final essay, "Some Intellectual Aspects of the Cardinal Virtues," **Paul Bloomfield** examines the intellectual elements of the cardinal moral virtues of courage, temperance, and justice, which he uses as a basis to explore the relation between these virtues and *phronesis*, as well as the much-discussed unity of the virtues thesis. According to Bloomfield, it is a mistake, for instance, to think of courage as merely a matter of being able to deal with fear appropriately, as perhaps the folk conception of courage allows. A truly courageous individual must know the difference between what is truly fearful and what is not, and thus possess certain axiological knowledge. But also, an even more purely intellectual aspect of courage concerns knowledge of risk—of knowing when, for example, tactics or missions in battle or in a courtroom are possible, and when not. Similar observations apply to the virtues of temperance and justice. Temperance, on Bloomfield's reading, requires both the axiological knowledge of what is truly valuable in life, as well as knowing how and to what degree one's personal affections and desires influence one's decision-making. Justice, too, involves an intellectual component of knowing how to discriminate among cases calling for judgment, so that truly like cases are judged alike. Reflection on the intellectual components of these virtues helps reveal the complex manner in which they are properly related to *phronesis* (wisdom) and to the unity of virtues thesis. With regard to their unity, Bloomfield defends a "limited" unity thesis, according to which having the virtue of courage, for instance, implies having it in all domains of life, and rules out having the kinds of vice associated with temperance and justice. However, being a fully courageous person does not imply having full possession of these other two moral virtues. Just as an expert electrician must have some minimal competence in carpentry and plumbing, she need not have the kind of expertise of the carpenter and plumber that results only from the extensive experience and thus knowledge gained in the practice of these latter skills.

REFERENCES

Cullity, Garrett (2004). *The Moral Demands of Affluence* (Oxford: Oxford University Press).

Herman, Barbara (2001). "The Scope of Moral Requirement," *Philosophy and Public Affairs* 30: 227–56.

Scanlon, T. M. (1990). "Promises and Practices," *Philosophy and Public Affairs* 19: 199–226.

Singer, P. (1972). "Famine, Affluence, and Morality," *Philosophy & Public Affairs* 1: 229–43.

Urmson, J. O. (1958). "Saints and Heroes," in A. I. Melden (ed.), *Essays in Moral Philosophy* (Seattle and London: University of Washington Press): 198–216.

William, Bernard (1973). "A Critique of Utilitarianism," in J. J. C. Smart and Bernard Williams, *Utilitatianism: For and Against* (Cambridge: Cambridge University Press).

—— (1976). "Persons, Character, and Morality," in Bernard Williams, *Moral Luck* (Cambridge: Cambridge University Press.)

1

Supererogation and Virtue

ROGER CRISP

1.1 AN IMMORAL PARADOX?

Consider the following case, from a seminal paper by Urmson:

> *Grenade.* We may imagine a squad of soldiers to be practicing the throwing of live hand grenades; a grenade slips from the hand of one of them and rolls on the ground near the squad; one of them sacrifices his life by throwing himself on the grenade and protecting his comrades with his own body. (Urmson 1958: 202)

Urmson says it is quite clear that, had the soldier not thrown himself on the grenade, he would not have failed in his duty, and that no one could have said to him: 'You ought to have thrown yourself on that grenade'. The possibility of supererogation—the performance of an action which is morally admirable and yet beyond the call of duty—is a central component of common-sense morality, the set of moral principles that most of us are brought up to live by. And yet its presence might seem to create a puzzle for common sense, as Henry Sidgwick notes:

> Certainly we should agree that a truly moral man cannot say to himself, 'This is the best thing on the whole for me to do, but yet it is not my duty to do it though it is in my power': this would seem to common sense an immoral paradox. (Sidgwick 1907: 220)

The context of the passage—a discussion of the relation between virtue and duty—suggests that by 'the best thing on the whole' Sidgwick means 'morally best' or 'virtuous'. Sidgwick avoids the paradox by distinguishing between, on the one hand, the notion of duty, and on the other the notion of what we ought or ought not to be blamed for. Although virtue does indeed involve doing what is ordinarily considered beyond duty, it would be inefficient to blame everyone who fails to live up to this ideal:

> [W]e think that moral progress will on the whole be best promoted by our praising acts that are above the level of ordinary practice, and confining our censure…to acts that fall clearly below this standard.

Sidgwick's solution, however, is likely to convince only those who accept that common sense can be 'unconsciously utilitarian' (Sidgwick 1907: 424, 454). On the face of it, common sense does not (even 'unconsciously') view supererogation as a mere device to allow us to put blame to its best use. So, if Sidgwick is right to say that common sense will recognize the immoral paradox, we have here another of those tensions within common-sense morality which Sidgwick is elsewhere so effective at bringing out. I suspect, however, that many people would not recognize the paradox in the first place, allowing that there is no duty to be virtuous and that the virtuous go beyond duty. This raises the question central to this essay: whether a theory of virtue should incorporate the common-sense view or adopt Sidgwick's more demanding ideal.[1]

1.2 SUPEREROGATION

How should we understand the common-sense conception of supererogation? Straightforward parsimony requires us to keep our account as simple as possible. I suggest, then, that for an act to be supererogatory, it must meet the following three conditions:

(a) *Non-duty.* It must not be a moral duty.[2]
(b) *Moral value.* It must be morally more valuable than its non-performance.
(c) *Appropriate motivation.* It must be appropriately motivated.[3]

[1] For the view that virtue ethics can incorporate supererogation through the notion of a virtuous ideal observer, see Kawall (2009).

[2] See Chisholm and Sosa (1966: 327).

[3] In a broader sense, an act may be described as supererogatory if it is such that it *would* be morally valuable if appropriately motivated. But my concern will be with that subset of such acts which are appropriately motivated. See Montague (1989: 100–11).

The non-duty condition (a) might seem close to analytic, taking its notion of duty from the idea of going beyond duty. Urmson, as we saw, claims that, had the soldier decided not to sacrifice himself, no one could have told him that he ought to have done so. But it might be said that there are 'weaker' senses of 'ought' or 'duty' which would permit this.[4] One can imagine a soldier who stood back while another sacrificed himself later feeling ashamed, even blaming himself ('It ought to have been me', 'I should have done it', 'I ought to have done it'), and yet believing that he has not violated any strict moral duty. If such weaker senses are available and coherent, this throws into doubt the suggestion that supererogatory actions must be entirely optional and at the discretion of the agent, in the same sense that it is entirely up to me whether I eat my last sweet now or save it until later.[5] But whether there are such senses is a matter for discussion elsewhere, since it does seem that the common-sense conception of superogation includes condition (a).

Nor need we commit ourselves to the view that that there is a *conclusive reason* to perform the supererogatory act which can be stated in the form of an 'ought' claim.[6] In at least many cases, common sense will allow only that there is *a* reason to go beyond duty, not that failure to do so is a failure to respond properly to reason. Indeed, the agent need not even have a reason for *not* acting on the supererogatory reason. Consider Horgan and Timmons' case of Olivia.

Olivia's offer. Olivia and her husband Stan have recently moved to St Louis, each having accepted an academic appointment at one of the local universities. During their first week in their new home, Olivia attends a block party organized by one of their new neighbours where she meets a recently widowed

[4] See, for example, Feinberg (1961: 276–7); Forrester (1975: 219); Dancy (1988: 175). Cf. Trianosky's suggestion that we need to distinguish between deontic and aretaic judgements to explain why we feel the need to offer excuses for not performing supererogatory actions (1986: 27–30; see also Hale 1991: 273); Mellema's 'quasi-supererogatory' actions, which are non-obligatory but blameworthy to omit (1991: ch. 5); the distinction mentioned by Johnson between two senses of 'right': right as fully adequate, and right as morally excellent (2003: 825); Markovits's suggestion that supererogatory actions may be obligatory and yet such that it would be inappropriate for us to demand them of others (2011: 12).

[5] See, for example, Urmson (1958: 203); McNamara (1996: 42); Zimmerman (1996: 236).

[6] See, for example, Raz (1975: 164). Cf. Portmore's suggestion that a supererogatory act is one we have most moral reason to do, while we have most reason overall to act otherwise (2003: 307). As Postow (2005: 246) pointed out, this view makes supererogatory actions in an important sense irrational, and Portmore himself later resiled from the position for that reason (2008: 382, n. 21).

> woman, Mary, a neighbour who lives a few doors down from Olivia and Stan. In conversation, Olivia learns that Mary lost her husband to cancer after forty-eight years of marriage. She also learns that Mary is an avid baseball fan and that she and her husband used to regularly attend Cardinals games. But without anyone to go with, she does not go any more. The next day, it occurs to Olivia that it would be a nice gesture to offer to go to a Cardinals game with Mary, though she herself has no particular interest in the game. But she thinks: 'Here is a chance to do something nice for someone, and the fall semester does not begin for another couple of weeks. Why not?' She calls Mary, who is delighted by the invitation, and they end up going to a game. (2010: 47)

As Horgan and Timmons point out (2010: 48), Olivia does not take herself to need any excuse if she decides not to extend the invitation. The reasons to invite Mary are *favouring* rather than *requiring* ones, in any sense of 'requiring' (morally, rationally, or whatever). This case implies that it would be a mistake to include a further 'sacrifice' condition in our account. Supererogation, though often costly, can be cost-free: Olivia may even know that she will enjoy the game.

I have put condition (b) in terms of moral value to avoid questions concerning praiseworthiness and blameworthiness such as those raised by Sidgwick's suggested bypass for the immoral paradox. 'Praiseworthy' has at least two distinct senses: 'deserving of praise' and 'worth praising' (i.e. 'such that there is a reason to praise'). These two senses can come apart. We can imagine cases in which someone has done something deserving of praise, but there is a reason—perhaps a conclusive reason—not to praise them (because they are excruciatingly shy, say, or because they will be harmed by some other person envious of the attention they are receiving), as well as cases where there is a reason to praise something entirely undeserving of praise (such as the banal etchings of someone who has taken you hostage). The important sense in understanding supererogation is the first, and this is most plausibly seen as correlative with moral value. In a standard case of supererogation, we praise (to speak strictly) neither the supererogatory act itself, considered independently of the agent, nor the agent *qua* agent, but the agent for doing what she does. In *Grenade*, as Urmson notes, the soldier who sacrifices himself shows himself to be 'superior in some way' to his comrades. That is to say, he is morally superior in acting in the way he does; perhaps in other respects he was an utter cad, and overall inferior to his comrades. Another advantage of avoiding talk of blame and blameworthiness here is that we need not detain ourselves with the question whether, according to common sense, blame is appropriate only for a failure to perform a duty, and not for a failure to perform a supererogatory action.

This brings me to the final condition, concerning appropriate motivation. Consider *Grenade* again. Urmson is, implicitly, asking us to assume that all is as it seems. We may revise our judgement if we learn that the soldier believed that he alone of his comrades knew the grenades to be dummies, and was seeking their admiration for his apparent willingness to sacrifice himself.

1.3 A BRIEF HISTORY OF SUPEREROGATION

Urmson's article was an attempt to dislodge what he saw as the standard threefold philosophical classification of actions into duties, permitted actions, and wrong actions, which he saw as 'inadequate to the facts of morality' (1958: 199). Though it is not clear, I suspect Urmson thought that, because the justification for believing in supererogation is so strong, any moral theory that ignores it will itself be inadequate.

On the face of it, Urmson's position might be said to be somewhat parochial. Imagine that some moral theory rested on the view that burning children alive causes them no pain. That view is so inconsistent with facts that have been, and will always be, acknowledged that the theory is almost certainly to be rejected. But supererogation is not like that. It became part of our conceptual scheme only quite recently, through Christianity. In brief, the story appears to be roughly as follows.[7] In the Gospel of St Matthew, we read the following account of a rich young man's meeting with Jesus:

> [**16**] And, behold, one came and said unto him, Good Master, what good thing shall I do, that I may have eternal life?
> [**17**] And he said unto him, Why callest thou me good? there is none good but one, that is, God: but if thou wilt enter into life, keep the commandments.

[7] See, for example, Heyd (1982: 13–29); Dentsoras (2011).

> [**18**] He saith unto him, Which? Jesus said, Thou shalt do no murder, Thou shalt not commit adultery, Thou shalt not steal, Thou shalt not bear false witness,
> [**19**] Honour thy father and thy mother: and, Thou shalt love thy neighbour as thyself.
> [**20**] The young man saith unto him, All these things have I kept from my youth up: what lack I yet?
> [**21**] Jesus said unto him, If thou wilt be perfect, go and sell that thou hast, and give to the poor, and thou shalt have treasure in heaven: and come and follow me. (*Matthew* 19: 16–21, King James Version)

In the third and fourth centuries, the Church Fathers used this passage to justify the poverty and obedience required for the monastic life.[8] Then, in the works of Sts Ambrose and Augustine, we find for the first time a distinction between 'precepts' (requirements) and 'counsels' (recommendations). One significant motivation for the further development of the doctrine of supererogation was the role it could play in the institution of 'Indulgences', which involved the idea that the merit of Jesus and the Saints was stored in what came to be known as the 'Spiritual Treasury of the Church'.[9] This merit was available from the Pope as compensation for taking part in a Crusade, or for purchase as a way to remit the penalties of sin (originally at least the sinner also had genuinely to repent). The doctrine was developed most carefully and influentially by Aquinas, who contrasted the 'Old Law' of Bondage based on the Ten Commandments with the 'New Law' of Liberty.[10] The New Law not only states what is *required* for salvation, but also *recommends* certain ways of achieving it more effectively or quickly (viz., chastity, poverty, and obedience).[11]

[8] Chastity was justified with reference to a slightly earlier passage in the same chapter:

> [**10**] His disciples say unto him, If the case of the man be so with his wife, it is not good to marry.[**11**] But he said unto them, All men cannot receive this saying, save they to whom it is given.[**12**] For there are some eunuchs, which were so born from their mother's womb: and there are some eunuchs, which were made eunuchs of men: and there be eunuchs, which have made themselves eunuchs for the kingdom of heaven's sake. He that is able to receive it, let him receive it.

[9] See Jardine (1996: 335–6). (Thanks to Julia Annas for this reference.)

[10] See, for example, Aquinas (1947: I-II Q. 107, Art. 1; Q. 108, Art. 4).

[11] See Irwin (2007: sect. 340).

The doctrine of supererogation, and the institution of indulgences, were major targets of criticism during the Reformation.[12] According to Luther, even the Saints, in their most perfect work, do no more than is required, and hence they have no superfluous merit to help those who are 'lazy' (1957: 213, 215), while Calvin suggests that righteousness is a matter of grace, not works, and, since God is anyway entitled to everything we are and possess, there can be no merit for us in what we are or do (1960: 3.14.3, 15).

Indulgences are no longer for sale within the Catholic church, but the doctrine of supererogation remains a part both of Catholic theology and of common-sense morality. One hypothesis is that the Reformation and post-Reformation theologians were primarily out to unseat the institution of indulgences, and left supererogation alone once they had succeeded in that. Or, of course, it may be that this interesting twist in the history of ethical thought, largely springing from interpretation of one or two passages in the Bible, enabled us for the first time to understand a central aspect of morality which had remained hidden from the ancients, including Socrates, Plato, Aristotle, and all of the pre-Christian Hellenistic philosophers. But my point against Urmson does not depend on any such hypothesis. What I want to stress is that the apparent 'moral facts' have changed over time, so that whether some moral theory does or does not countenance supererogation cannot plausibly count for or against it. The idea has to be defended on its own merits. Of course, Urmson and others have provided such defences, and I shall come to them below. But I want first to explain the Aristotelian conception of virtue, which, like Sidgwick's, is inconsistent with supererogation as I have outlined it above. We can then consider whether those attracted to an Aristotelian account of the virtues should be inclined to amend it in response to post-Aristotelian arguments in favour of supererogation.

1.4 ARISTOTELIAN VIRTUE

Aristotle noticed that human life could be seen as consisting in several different 'spheres', involving certain characteristic feelings or actions. One set of spheres concerns the *pathē*—feelings, broadly construed to include emotions

[12] Much of the criticism was directed at Tertullian rather than Aquinas; see Trianosky (1998: sect. 1).

as well as pleasure and pain. All human beings who survive birth for a non-trivial period will experience, for example, fear, confidence, appetite, anger, pity, and pleasure and pain. Virtue consists in having these feelings 'at the right time, about the right things towards the right people, for the right end, and in the right way'—this is 'the mean, and best' (Aristotle 1894: 1106b21–2).[13] The mean lies between two extremes—that is, between two ways of going wrong. Consider anger (1894: 4.5). The virtue of even temper will consist in feeling anger at the right time, about the right things, and so on. There is a suite of all-too-common excessive vices: feeling anger at the wrong time, about the wrong things, for too long, too quickly, and so on. But there is also a set of deficient vices leading one to fail to feel anger at the right time, about the right things, and so on. As Aristotle puts it, employing his standard analogy of the virtuous person with the archer: '[O]ne can miss the mark in many ways...but one can get things right in only one' (1894: 1106b28–31).

The same analysis applies to actions. Consider money, the control of which again will concern nearly everyone at certain periods of their life. Here the virtue is generosity, and consists in giving away money at the right time, to the right people, in the right amounts, and so on (1894: 4.1).[14] The excessive vice will be wastefulness, and the deficient stinginess. Further, as Aristotle notes (1894: 1121a30–32), excessive and deficient vices are commonly found together. If you waste your money, then you will be unable to give when you should and hence will end up being stingy.

Aristotle's account does not require complete success of the virtuous person. A generous person may spend more or less than is right or noble on certain occasions (and hence be—moderately—pained by what he has done) (1894: 1121a1–2). But there is no room in this account for going *beyond* duty. The doctrine of the mean itself is couched in the language of duty. The virtuous person will become angry *hote dei*—when it is his duty to do so.[15] If we take the three recommendations of the New Law, then, the Aristotelian view would be Protestant rather than Catholic. If poverty, obedience, and chastity are part of the morally ideal life, then they are your

[13] Translations are mine.

[14] Aristotle puts the taking of money also within the sphere of generosity; it would perhaps have been wiser to create an independent virtue to govern this activity, though the two are of course closely related.

[15] Liddell and Scott (1940: *s.v.*) use notions such as 'must' and 'ought', but I take it that in this context it is clear that what is meant is the same as 'duty'. See, for example, Woodhouse (1910: *s.v.*). It might be claimed that Aristotle distinguishes between strong and weak senses of duty (see Section 1.2); that, for example, he accepts a teleological conception of duty

duty. The same applies for other sacrifices, including of life itself, as we see in the eighth chapter of book 9 of Aristotle's *Ethics*:

> It is true also of the good person that he does a great deal for his friends and his country, and will die for them if he must; he will sacrifice money, honours, and in general the goods for which people compete, procuring for himself what is noble. (1894: 1169a18–20)

The virtuous person is not, on Aristotle's account, capable of full self-sacrifice—that is, giving up for the sake of others what is best for himself. But what matters here is the link between duty and virtue. Though Aristotle does not say so explicitly, it is clear that the virtuous person's dying for others is not supererogatory; it is virtuous, and what is virtuous is a matter of doing one's duty. This, then, is how he would understand *Grenade*. It is the duty of each soldier to throw himself on the grenade; and by the time another has done so, it is too late. Similarly with generosity. If the virtuous person makes large donations to Oxfam, for example, rather than spending the money on luxuries for himself, we can conclude that this is his duty. Virtue is itself an excess of a kind: you cannot go beyond it (1894: 1107a7–8).

Sidgwick's immoral paradox arises when an allegedly virtuous person allows that some course of action is best—'morally' best—but denies that it is her duty to take it. This represents a failure of what has become known in the literature as the 'good-ought tie-up' (Heyd 1982: 4).[16] On the face of it, however, this alleged tie-up might seem a rather shaky foundation for an anti-supererogationist position. Consider, for example, those many deontological views which allow some state of affairs to be the best, but deny that we ought to bring it about, perhaps because it would involve injustice or violate some other side-constraint. Aristotle's position is rather different. Towards the end of the *Ethics*, Aristotle

as necessity, according to which the performance of certain basic duties (or 'necessities') is required for moral respectability (or 'rectitude' (*honestas*), as Aquinas puts it ((1947: II-II Q. 80, Art. 1); thanks to Terry Irwin for the reference)), and that of other less basic duties for *greater* respectability (*maiorem honestatem*). But there is no textual evidence of any such distinction in Aristotle. On the most natural reading of Aristotle, virtue consists in the performance of duties which are categorical and non-teleological; see the discussion of fittingness in the main text following.

[16] See Dancy (1988: 180): 'the problem of supererogation just is the problem of how moral thought can be evaluative without being prescriptive'.

compares the life of practical virtue with that of contemplation, saying of the former:

> [W]e do just actions, courageous actions, and the other actions in accordance with the virtues, in relation to each other, in contracts, services, and actions of all kinds, and in feelings as well, maintaining what is fitting for each. (1894: 1178a10–13)

Exactly how to understand the phrase 'what is fitting (*to prepon*) for each' is not clear. It can be taken to refer back to 'each other', but this seems to me unlikely, given the distance between the two phrases in the Greek. Alternatively, Aristotle may be suggesting that we maintain what is fitting for each of our actions and feelings, or what is fitting for each of us in relation to actions and feelings. Whichever of these Aristotle means, the idea is similar, and can be elucidated by reference to the doctrine of the mean. In each sphere of our lives, there will be actions and feelings that are fitting or appropriate to the situations in which we find ourselves. That fittingness is in a sense foundational, but that it is not to say that it cannot be elucidated by reference to the particulars (*ta kath'hekasta*) (1894: 1109b22–3 and *passim*). 'Why did you give money, and in that amount, to that person?' 'She is promoting a worthy cause, which helps prevent blindness in developing countries. I have already committed a good deal of my income this month to a bridging loan for a friend buying a house, so I could not reasonably have given more.' At this point, a person with any degree of moral insight, or *phronēsis*, will probably be satisfied. And there is little more that can be said—without seeking to explain a good deal of ethical theory—to the egoist who can see nothing fitting in helping those in need.[17]

The notions of fittingness and duty in Aristotle are closely related. The mean at which we should aim can be described either as what it is fitting or as one's duty. In other words, what we find in Aristotle is not so much a good-ought tie-up as a fittingness-duty tie-up. Fittingness is not a good, but rather a relation between circumstances and actions or feelings. Virtue is then conceptually downstream from both fittingness and duty, and is characterized in terms of a disposition to perform duties to act and to feel. So this gives us a duty-virtue tie-up. Finally,

[17] See McDowell (1980: 369–72).

moral value (*to kalon*) supervenes on exercising the virtues, which is acting and feeling as one has a duty to act and to feel, which is to act and to feel as befits the circumstances one is in. So here we have a four-way tie-up: moral value-virtue-duty-fittingness. Let me call this a *fittingness theory* of ethics. We might be reminded here of C. D. Broad's view of rightness.

> It seems to me that, when I speak of anything as 'right', I am always thinking of it as a factor in a certain wider total situation, and that I mean that it is 'appropriately' or 'fittingly' related to the rest of this situation. When I speak of anything as 'wrong' I am thinking of it as 'inappropriately' or 'unfittingly' related to the rest of the situation...What I have just asserted is not, and does not pretend to be, an analytical *definition* of 'right' and 'wrong'. It does bring out their relational character, and it correlates them with certain other notions. But the kind of appropriateness and inappropriateness which is implied in the notions of 'right' and 'wrong' is, so far as I can see, specific and unanalysable.[18]

The various components of the Aristotelian account described above—the three tie-ups I have mentioned—seem to me independently plausible, and their coherence adds credibility to the account as a whole. Your duty is to do what is fitting in the circumstances you are in; to act in a way that is not your duty will be to do what is unfitting and inappropriate. Virtue could not plausibly consist in acting or feeling in an unfitting way, and it is hard to see how acting or feeling fittingly could be anything other than morally valuable, and failure to do so, or acting and feeling unfittingly, could be anything other than vicious and morally disvaluable or 'disgraceful' (*aischros*).

Aristotle might allow that in certain cases it is roughly equally fitting for you to help some other person or to do something for yourself. But then your duty will be to do one or other of these things; helping the other person will not be 'going beyond' duty. What about *Grenade*? Here, I think, Aristotle will deny that it can be fitting to stand by while someone else sacrifices their life for you. If one is unsure about what is fitting in any case, one can examine the options for their moral value, because of the (indirect) tie-up between moral value and fittingness (via

[18] Broad (1930: 164–5), cited approvingly, and with reference to Clarke, in Ross (1939: 51–5). Philip Stratton-Lake rightly notes how close an account of duty in terms of fittingness is to an account in terms of reasons (2002: xxxv).

virtue and duty). So not sacrificing yourself in such a case is vicious (it is cowardly, and probably violates obligations of friendship to one's comrades).[19]

It could be, however, that there are arguments for supererogation that count strongly against the Aristotelian conception. Let me now discuss those offered in Urmson's paper.

1.5 ARGUMENTS FOR SUPEREROGATION

Urmson (1958: 211–14) offers five reasons in favour of making room for supererogation in one's ethical theory.[20] Some of these have been criticized or developed further in the literature inspired by his article.[21] Let me deal with each in turn.

1. It is important to give a special status of urgency, and to exert exceptional pressure, in those matters in which compliance with the demands of morality by all is indispensable.

One of the problems with Urmson's account is that, by this stage in his paper, he has moved away from the question of which is the correct moral theory to that of which 'moral code will best serve human needs' (211). This explains how, on the one hand, he can (correctly) point out that traditional utilitarianism is inconsistent with the idea of supererogation (206–7), and on the other claim that utilitarianism is best able to 'accommodate the facts' of supererogation (208, 214–15). According to utilitarianism, if we make a few simple assumptions, the sacrificial soldier in *Grenade* was only doing his duty. But, in utilitarian terms, it would be counter-productive to use the sanctions of morality to enforce such duties.[22]

[19] In an appendix, I provide a response to Heyd's suggestion that supererogation, or elements of supererogation, are present in Aristotle's ethics.

[20] Although this essay is not advocating 'virtue ethics', I take it that any plausible account of ethical or moral virtue will link virtuous traits in some way with what we ought ethically to do and hence with ethical theory.

[21] All otherwise unattributed page references in this section are to Urmson (1958).

[22] As we saw, this is close to Sidgwick's view.

Whether it would be counter-productive is largely an empirical matter, of more interest to those working in applied rather than theoretical ethics.[23] Certainly when it comes to helping others, it seems to me highly implausible to think that our currently lax moral code is the most efficient. Nor does it seem to me obvious that it would be self-defeating to seek to inculcate in soldiers a common-sense moral principle (such as *dulce et decorum* (fitting) *est pro patria mori*) so as to promote self-sacrifice. Indeed, it might even be in the interest of soldiers themselves.[24]

If we detach Urmson's suggestion from its utilitarian context, however, we can ask whether morality itself should be seen as incorporating supererogation as a reflection of the urgency of certain duties. If we allow duties to be placed on a scale from more to less urgent, it is not clear why there is any need for the kind of threshold implicit within the doctrine of supererogation.[25] It may be that the duty to sacrifice oneself for one's comrades is much weaker and hence less urgent than one's duty not to murder them. But there seems no immediate reason why we should think of it as no duty at all.

> 2. If we are to exact basic duties like debts, and censure failure, such duties must be, in ordinary circumstances, within the capacity of the ordinary man.

Urmson's point here is one about the effects of applying the Aristotelian view of duty in cases such as *Grenade*. He suggests that, if certain highly demanding courses of action of which only some small subset of people were capable were said to be matters of duty, 'duty would seem to be something high and unattainable, and not for "the likes of us" '.

Again, this is a point concerning not ethical theory, but the best moral code. So it is open to an Aristotelian to accept Urmson's point, and, again following Sidgwick, to recommend against the advocacy of especially stringent duties which apply only in exceptional cases. But again it is not clear that talk of such duties, especially if they are rare, will have the 'inevitable' consequence of bringing the very notion of duty into disrepute. Urmson's comparison with prohibition laws suggests

[23] See Clark (1978–9: 27).

[24] Note that Urmson's original example concerns practising with grenades. Tony Coady has suggested to me that this may be because Urmson felt that in the field each soldier would have had an obligation to sacrifice himself.

[25] See New (1974: 180–1); Attfield (1979: 482).

that any moral duty must apply to all. But the Aristotelian can make it clear that some duties can be performed only by exceptional people, so those who are psychologically incapable of performing them should not feel that they have failed in their duty. These duties would not be 'for the likes of us'; but plenty of other duties would be.

> 3. A moral code, if it is to be a code, must be formulable, and if it is to be a code to be observed it must be formulable in rules of manageable complexity.

Urmson's thought here is that certain moral rules—such as 'Do not murder' or 'Do not break promises'—are easily formulated and applied, whereas it would be 'absurd' to try to formulate some complicated rule to determine when it is or is not one's duty to 'go off and nurse lepers'. Urmson mentions also two other examples of supererogation: excusing debts, and nursing sick neighbours.

Aristotle himself would no doubt be somewhat doubtful about whether the application of rules concerning murder and promises is as clear as Urmson claims. Is it murder when non-combatants are killed as a foreseen consequence of some offensive action taken in a just war? Should I keep a promise if its fulfilment will harm the promisee? Even here some degree of *phronēsis* is called for. Nor should we expect a moral theory to deliver separate principles covering all the different ways in which human beings can benefit one another. There seems no reason to think that a simple principle such as 'Help others when you can', understood as a principle of duty, could not be applied effectively by most people.[26] Indeed, common-sense morality itself appears to include such a principle, which raises a difficulty peculiar to it of distinguishing between cases in which that principle applies and cases of supererogation. This difficulty will be mirrored in any moral theory that includes any such principle of beneficence alongside supererogation, and it will put pressure on such a theory to reinterpret cases of apparent supererogation as cases of discretion governed by imperfect duties.

> 4. It is part of the notion of duty that we have a right to demand compliance from others even when we are interested parties...A line must be drawn

[26] See New (1974: 182); Attfield (1979: 483); Heyd (1982: 167).

> between what we can expect and demand from others and what we can merely hope for and receive with gratitude when we get it; duty falls on one side of this line, and other acts with moral value on the other, and rightly so.

Urmson uses promise-keeping as an example of something we can demand from others, claiming that we cannot demand of some stranger that she tend us when we are ill or offer us a cigarette when we have run out, though either form of assistance would, as supererogatory, have moral value. Note that Urmson is not here making a point about demands themselves. That would be for him to do little more than reiterate the common-sense view. Rather, he is suggesting that it is a conceptual truth that we have a right to demand from others that they comply with their duties.

First, it seems to me clear that this is not a *conceptual* truth (so it is not 'part of the notion of duty'). Second, it seems *fairly* clear that common sense does not take this view of duty. Sometimes, for example, it would be just plain rude to demand thanks from one's beneficiary. Just as blameworthiness in principle can come apart from blameworthiness in practice, so the same may be true of whether one has a right to demand compliance. There may be cases—such as that involving the cigarette, perhaps—where, because your potential benefactor has a duty to help you, you have *a* reason to demand compliance, but overall stronger reason (resting on the workability of a real-life moral system, or indeed the sheer unlikelihood of success or the non-negligible probability of causing harm to oneself) not to do so. Further, the very notion of 'imperfect duty' relies on the thought that one can have general duties of, say, benevolence, without any particular individual's being in a position to demand any benefit. So lack of a right to demand the performance of some action which in fact constitutes (perhaps only partial) fulfilment of some duty is no evidence of supererogation.

Heyd provides what he takes to be a further argument for Urmson's fourth claim (1982: 167, 172–5). This argument is based on the notion that individuals have a right to autonomy—to pursue ends of their own choosing, to satisfy their desires, and to seek to realize personal ideals. Heyd sees it as 'analogous to the widely accepted belief that individual persons should not be sacrificed for the promotion of overall happiness'.

Although Aristotle himself does not place any non-instrumental value on the kind of autonomy underlying the right alleged by Heyd, nothing prevents an Aristotelian allowing such a right. Indeed, it may be that this right will itself play a role in setting the limits of moral demands. Let us say that generosity requires the virtuous agent to give roughly 5% of her income to charity. An Aristotelian view including a right to autonomy might then allow that agent discretion as to how to use the rest of her salary. She might of course choose to give more than 5%, and though this would be permitted it would not be especially morally valuable. If it is said clearly to have such value, then the Aristotelian will claim that the level of required sacrifice has been set too low. At some point, however, it will give out, and any sacrifice beyond that might, plausibly enough, be seen as the result of a vice such as imprudence or 'wastefulness'. The agential discretion, then, that may be required for the right to autonomy need not be that provided by the doctrine of supererogation.

> 5. In the case of basic moral duties we act to some extent under constraint... But free choice of the better course of action is always preferable to action under pressure, even when the pressure is but moral... [T]here is something horrifying in the thought of pressure being brought on [someone] to perform an act of heroism.

Urmson's first point here may be taken to concern motivation. If so, the Aristotelian will have no difficulty with it. Aristotle himself does not suggest that virtuous people will feel constrained by duty to act virtuously; rather, what motivates them will be the idea of the moral value supervening on the virtuous performance of duty. If Urmson is rather making a moral theoretical point, it seems merely to reassert the doctrine of supererogation, according to which there is special moral value in going beyond duty. There is no reason why the same degree of moral value might not be available within the Aristotelian account for the performance of duty.

Urmson's second point concerns the morality of pressurizing others into heroic action. Here we might appeal to a principle according to which it is wrong on certain occasions to put moral pressure on individuals. Indeed, this is central to the virtue of tolerance. And there is nothing to prevent the incorporation of such a principle into a moral theory, according to which, though each soldier in *Grenade* has a duty to sacrifice himself, it would be wrong to put moral pressure on any of them to perform that duty.

Heyd offers an argument he takes to be developing the themes of Urmson's fifth point (1982: 167, 175–8).[27] According to this argument, supererogation, unlike the morality of duty, allows for the *value* (as opposed to the right) of freely expressing one's individual preferences, including partiality towards others chosen by the agent herself. Again, however, Heyd is making undue assumptions about the morality of duty. A morality that limits the demands on agents can make room for the value of autonomy so understood. Indeed, it could be argued that it makes more room for it than a morality which allows for supererogation as entirely optional. For in a limited morality of duty, the agent need feel unconstrained by any thought that choosing certain actions will result in a loss of moral value. It really is up to her how to act—morality takes no view on the matter.

1.6 PARSIMONY IN ETHICAL THEORY

I have argued that, as far as supererogation is concerned, the classical conception of the relation between virtue and duty, in particular that Aristotelian version of it grounded in the notion of fittingness, is superior to the Catholic conception of that relation as allowing for supererogation—a conception which has found its way into our common-sense morality.

We may also draw a wider lesson. The doctrine of supererogation is a remarkable example of how contingent cultural change can deeply affect our moral intuitions, causing us to see as obvious certain apparent 'moral facts' which were previously unacknowledged. This reinforces the case for parsimony in normative theory. Rather than taking moral concepts at face value, as items already guaranteed a place in our conception of the normative realm, we should begin by assuming an empty world and then populate it only so far as is necessary, beginning with fundamental and carefully constructed questions. Consider a case in which at some very small cost to yourself you can bring an end to the

[27] See also Clark (1978–9: 29–30); Nagel (1986: 202–4); Scheffler (1992: 128).

severe suffering experienced by some very young child. One question you might ask here is what it would be fitting for you to do in this situation. That question can be answered without reference to duty, virtue, moral value, wrongness, blameworthiness, or indeed any other moral concepts. It may be that, after careful reflection on both the concepts we are using and those we are putting to one side,[28] all practical normative questions can be answered within the same, spare conceptual scheme. Of course, it may not. But it would be fitting for us at least to try, before flooding the normative world with historically and culturally freighted concepts that, like the doctrine of supererogation, may lead us into illusion and confusion.

APPENDIX: HEYD ON ARISTOTLE

In his important book *Supererogation* (1982) David Heyd claims that classical ethics—including the ethics of Aristotle—can be described as superogatory in a 'secondary' sense, since 'it is not based on duty' (38).[29] I hope I have said enough above to show that this is a somewhat misleading claim. The idea of duty is close to foundational in Aristotle's account of virtue and moral value.[30]

Heyd focuses first on Aristotelian 'particular' justice (43), suggesting that Aristotle will allow that one can take a smaller share than one's due or make a larger compensation than is required, and that this is clearly supererogatory. As Heyd admits, Aristotle does not in fact make this claim explicitly; but, he says, it is compatible with his view. How

[28] Those put to one side should not be forgotten. They should be checked every now and then for signs of life, and even when clearly defunct ('chastity', for example) remembered as part of the history of our living conceptual scheme.

[29] All page references in this appendix are to Heyd (1982). At 41–2, Heyd cites some passages of Seneca on gratitude which are said to show 'Seneca's awareness of the intrinsic value of actions which are not governed by principles of strict obligation'. These passages seem to me to concern motivation rather than grounding principles of obligation. What is valued is the willingness of a beneficiary to return the favour without being compelled by any sense of duty. This is not to say that she has no duty. (Heyd nowhere makes clear how his view of Seneca can be made consistent with his claim in ch. 1 that supererogation is a Christian doctrine.)

[30] It is sometimes claimed that Aristotle's notion of 'superhuman virtue, a virtue heroic and godlike' (1894: 1144a19–20) provides conceptual room within his theory for the notion of supererogation. But that the virtue of heroes and gods is beyond the virtue of ordinary humans does not show that those humans themselves can go beyond human virtue. For an argument that Kant's notion of imperfect virtue can allow for superlative virtue without supererogation, see Sherman (1997: 350–61).

justice fits into the doctrine of the mean is a difficult question, and one Aristotle wrestles with in a rather unsatisfactory way. But the statement of the doctrine of the mean in *Nicomachean Ethics* (*EN*) 2.6 strongly suggests that he would see both of these actions as vicious. 'One's due' is what one has a duty to take; and justice requires one compensate to the level duty requires, and no more.

Later (46), Heyd returns to the idea of taking less than one's due, suggesting that Aristotle's claim in *EN* 5.11 that no one can be voluntarily treated unjustly shows that acting in this way would be to go beyond justice without being unjust. But in 5.11 Aristotle carefully draws a distinction between voluntarily treating oneself unjustly and voluntarily acting unjustly in such a way that one suffers. A suicide is acting unjustly, but towards the city, not himself. And the same could be true of the person who takes less than her share.

Heyd mentions the distinction Aristotle draws in the *Eudemian Ethics* (*EE*) (1235a) between 'private justice' and 'justice towards others'. Private justice is 'practised to friends' and 'depends on ourselves alone', and 'allows—indeed recommends—surpassing the requirements of justice in the legal sense'. But that one sphere of life involves a special form of justice which is more demanding than that in another sphere is not supererogation. Heyd also cites *EN* 1155a, where Aristotle claims that law-givers are more concerned to promote friendship than justice, 'implying that a system of principles of distributive and rectifying justice is no more than a necessary evil, and that ideally, if everyone behaved supererogatorily... there would be no need for distributive and rectifying principles' (44). But here we have merely another case of priority among the virtues: it is better to have a community governed by the virtue of friendship than by the virtue of justice. This is not to say that friendship itself is not to be understood in terms of doing one's duty. The same goes for the notion of 'moral' friendship, distinguished from 'civic' friendship at *EE* 1242bff. Civic friendship is indeed contractual, so contractual demands cannot be made within moral friendships. But that is not to say that other, non-contractual duties are not in play.

Heyd then moves to a discussion of 'beneficence'. The single textual reference he gives—*EN* 1167b-1168a—suggests he is thinking not of the virtue of beneficence, but of the beneficial activities carried on within the sphere of friendship. Heyd claims that the gratitude felt by a recipient of such activity 'is an acknowledgement of the supererogatory, gratuitous nature of the beneficent act' (45). I cannot find this view in the passage Heyd cites. The only mention of gratitude is in the statement that it is expected by benefactors, while beneficiaries are reluctant to to feel it.

Heyd argues that beneficence is supererogatory because it requires more than particular justice. But we have already seen that the virtue of one sphere of life can be more demanding than that of another without any dependence on the idea of supererogation. As Heyd goes on to admit, if we take justice in its 'general' sense, equivalent to 'complete virtue', beneficent acts are 'simply just' (or, as we might put it, simply a matter of duty) (46). He concludes his discussion of Aristotelian beneficence with the suggestion that, although there are limits to beneficence according to Aristotle, these restrictions 'do not mean that beneficent acts are not supererogatory'. That is true; but neither does it mean that they are. Nor does it follow (Heyd 47–8) that, because it is 'more characteristic of virtue to do good than to have good done to one . . . and to do what is noble than not to do what is base' (*EN* 1119b-1120a), we should understand Aristotelian liberality or generosity as a matter of supererogation.

In *EN* 5.10, Aristotle distinguishes between 'equity' and justice. Equity is that virtue which enables judges to make decisions about particular cases in which the law is not clear. The equitable, that is to say, is in a sense 'superior' to justice, since it completes it. Heyd first suggests (47) that this notion of the equitable is analogous to his own conception of supererogation, since it respects his 'continuity condition', according to which 'supererogation should be characterized as realizing *more* of the same type of value attached to obligatory action' (5). But, as Heyd himself goes on to admit (47–8), this analogy provides no reason to interpret equity as a matter of supererogation. It is another case of two independent virtues, both to be characterized in terms of duty, and one of which 'trumps' another in terms of moral value.

In conclusion, it is worth explicitly noting Heyd's apparent retreat at the end of his chapter on classical ethics:

> It should . . . be noted that unlike our notion of supererogation, which attaches special moral value to the very act of going beyond duty, classical morality does not view virtuous action as meritorious for that reason . . . [A]lthough beneficent, equitable, and virtuous actions in general meet the condition of *continuity*, they cannot be treated as supererogatory in the strict sense of the word, because they do not fulfil the condition of *correlativity*. (48)

That condition Heyd outlines as follows:

> Correlativity means that acts of supererogation derive their special value from their being 'more than duty requires'; i.e. they have meaning only relatively to obligatory action. (5)

The continuity condition is equivalent to the condition that there is a single species of moral value. Since this condition may be met equally well by theories that deny as by those that accept supererogation, its being met by some theory provides no evidence either way on the room within that theory for supererogation. And since the correlativity condition must be met by any theory that allows for genuine supererogation, the fact that Aristotle's theory and indeed other pre-Christian theories—as Heyd admits—fail to meet it is conclusive evidence that they cannot allow for supererogation.[31]

REFERENCES

Aquinas, T. (1947) *Summa Theologica*, tr. Fathers of the Dominican Province, 3 vols. (New York: Benziger Bros.).

Aristotle (1894) *Ethica Nicomachea*, ed. I. Bywater (Oxford: Clarendon Press).

Attfield, R. (1979) 'Supererogation and Double Standards', *Mind* 99: 481–99.

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

Calvin, J. (1960) *Institutes of the Christian Religion*, ed. J. McNeill, tr. F. Battles (Philadelphia, PA: Westminster Press).

Chisholm, R. and Sosa, E. (1966) 'Intrinsic Preferability and the Problem of Supererogation', *Synthese* 16: 321–31.

Clark, M. (1978–9) 'The Meritorious and the Mandatory', *Proceedings of the Aristotelian Society* 79: 23–33.

Dancy, J. (1988) 'Supererogation and Moral Realism', in J. Dancy *et al.* (ed.), *Language, Duty, and Value* (Stanford: Stanford University Press): 169–88.

Dentsoras, D. (2011) 'The Birth of Supererogation', unpub. TS.

Feinberg, J. (1961) 'Supererogation and Rules', *Ethics* 71: 276–88.

Forrester, M. (1975) 'Some Remarks on Obligation, Permission, and Supererogation', *Ethics* 85: 219–26.

Hale, S. (1991) 'Against Supererogation', *American Philosophical Quarterly* 28: 273–85.

[31] For help with previous drafts, I am grateful to Julia Annas, Paul Bloomfield, Terry Irwin, Amelie Rorty, Daniel Schwartz, Drew Schroeder, Daniel Star, John Tasioulas, David Wiggins, and an anonymous referee; and to audiences at the Arizona Workshop on Normative Ethics, the Balliol Cerberus Society, and the Jurisprudence @ UCL Colloquium in Legal and Social Philosophy. The first draft of the essay was completed during my tenure of the Findlay Visiting Professorship in the Department of Philosophy, Boston University. I would like to thank the department for its intellectual, social, and practical support.

Heyd, D. (1982) *Supererogation: Its Status in Ethical Theory* (Cambridge: Cambridge University Press).

Horgan, T. and Timmons, M. (2010), 'Untying a Knot from the Inside Out: Reflections on the "Paradox" of Supererogation', *Social Philosophy & Policy* 27: 29–63.

Irwin, T. (2007) *The Development of Ethics*, vol. 1 (Oxford: Oxford University Press).

Jardine, L. (1996) *Worldly Goods: A New History of the Renaissance* (London: Macmillan).

Johnson, R. (2003) 'Virtue and Right', *Ethics* 113: 810–34.

Kawall, J. (2009) 'Virtue Theory, Ideal Observers, and the Supererogatory', *Philosophical Studies* 146: 179–96.

Liddell, H. and Scott, R. (1940) *Greek–English Lexicon*, 9th edn. (Oxford: Clarendon Press).

Luther, M. (1957) 'Explanations of the Ninety-Five Theses', in *Works*, ed. H. Grimm, vol. 31 (Philadelphia: Muhlenberg Press).

Markovits, J. (2012) 'Saints, Heroes, Sages, and Villains', *Philosophical Studies* 158: 289–311.

McDowell, J. (1980) 'The Role of *Eudaimonia* in Aristotle's *Ethics*', in *Essays on Aristotle's Ethics*, ed. A. Rorty (Berkeley and Los Angeles, CA: University of California Press): 359–76.

McNamara, P. (1996) 'Making Room for Going Beyond the Call', *Mind* 105: 415–50.

Mellema, G. (1991) *Beyond the Call of Duty: Supererogation, Obligation, and Offence* (Albany, NY: State University of New York Press).

Montague, P. (1989) 'Acts, Agents, and Supererogation', *American Philosophical Quarterly* 26: 100–11.

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

New, C. (1974) 'Saints, Heroes, and Utilitarians', *Philosophy* 49: 179–89.

Portmore, D. (2003) 'Person-Relative Consequentialism, Agent-Centered Options, and Supererogation', *Ethics* 113: 303–32.

—— (2008) 'Are Moral Reasons Morally Overriding?', *Ethical Theory and Moral Practice* 11: 369–88.

Postow, B. (2005) 'Supererogation Again', *Journal of Value Inquiry* 39: 245–53.

Raz, J. (1975) 'Permissions and Supererogation', *American Philosophical Quarterly* 12: 161–8.

Ross, W. D. (1939) *Foundations of Ethics* (Oxford: Clarendon Press).

Scheffler, S. (1992) *Human Morality* (New York, NY: Oxford University Press).

Sherman, N. (1997) *Making a Necessity of Virtue: Aristotle and Kant on Virtue* (Cambridge: Cambridge University Press).

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

Stratton-Lake, P. (2002) 'Introduction' to W. D. Ross, *The Right and the Good* (Oxford: Clarendon Press).

Trianosky, G. (1986) 'Supererogation, Wrongdoing, and Vice: On the Autonomy of the Ethics of Virtue', *Journal of Philosophy* 83: 26–40.

—— (1998) 'Supererogation', in E. Craig (ed.), *Routledge Encyclopedia of Philosophy* vol. 9 (London: Routledge): 232.

Urmson, J. (1958) 'Saints and Heroes', in A. I. Melden (ed.), *Essays in Moral Philosophy* (Seattle, WA, and London: University of Washington Press): 198–216.

Woodhouse, S. (1910) *English–Greek Dictionary* (London: George Routledge and Sons).

Zimmerman, M. (1996) *The Concept of Moral Obligation* (Cambridge: Cambridge University Press).

2

The Weight of Moral Reasons

RALPH WEDGWOOD

Many thinkers have held that one of the central issues of moral philosophy is a question of the *reasons*—if any—that we have for conforming to moral requirements. In effect, this is the question "Why be moral?" Many different answers to this question have been attempted—including Humean, Hobbesian, and Kantian answers, among others.

This essay aims to develop a different sort of answer—an answer that is situated within the framework of a broadly *value-based* conception of reasons for action, of roughly the sort that has been advocated by Joseph Raz (1999a).[1] According to such value-based conceptions, every reason for action corresponds to a fact about how the available options instantiate some appropriate *value*. Roughly, according to these conceptions, you have a reason to φ if and only if the option of φ-ing is available to you, and is in an appropriate way a *good thing to do*.

First, however, we have to clarify what exactly our central question *means*. This will help us to understand what would count as an adequate way of answering it.

2.1 WHAT IS THE QUESTION?

Not all moral philosophers accept that there is a good question to answer here. For example, the eighteenth-century British moralist Richard Price (1787, 180) writes:

> To ask, why are we obliged to practise virtue, to abstain from what is wicked, or perform what is just, is the very same as to ask, why we are *obliged* to do what

[1] For a more detailed version of this approach to reasons for action, see Wedgwood (2009a).

we are *obliged* to do?—It is not possible to avoid wondering at those, who have so unaccountably embarrassed themselves, on a subject that one would think was attended with no difficulty.

In view of how many philosophers have been seriously perplexed by this question, Price's interpretation of the question is too uncharitable to be credible. In what follows, I shall interpret the question in the following way. First, let me introduce the idea of a "non-trivial moral requirement." Roughly, a requirement that you are subject to counts as *non-trivial* if and only if it is possible for you not to conform to it, and it would intelligible for you to be tempted not to. (I shall say more about such "non-trivial" requirements later.)

With this notion of non-trivial requirements in hand, we can now interpret our central question as concerned with the following two propositions:

i. All normal human beings are subject to *non-trivial moral requirements*.
ii. Whenever any agent is subject to a non-trivial moral requirement, that agent has *overriding* or *conclusive* reasons for conforming to it.

Our central question is, in effect: What explains *why* these two propositions are true?

Some philosophers might be tempted to agree with H. A. Prichard (1912, 21) that this question is "improper"—on the grounds that these two propositions are simply utterly primitive truths, which cannot be explained on the basis of any other truths whatsoever. However, it is not clear that there is any compelling argument for the conclusion that these propositions are primitive truths of this sort.

One famous point that Prichard (1912, 23) made in the course of arguing for this conclusion is that it cannot be that the *only* conclusive reason that we have for conforming to moral requirements is that it is in our self-interest to do so. In what follows, I shall accept this point: according to the account that I shall propose here, the overriding or conclusive reasons that we have for conforming to moral requirements must always include distinctively *moral* reasons, and need not include any reasons of self-interest at all. However, this point is obviously compatible with the thought that the correct account of the nature of reasons for action and of moral requirements could help us to understand why there are any moral reasons at all, and why these moral reasons include overriding or conclusive reasons for conforming to all moral requirements.

This is the sort of answer to our central question that I shall search for here. I shall start by giving a sketch of a general conception of reasons for action and of moral requirements; and then I shall argue that this conception can help us to understand why these two propositions are true.

2.2 A GENERAL CONCEPTION OF REASONS FOR ACTION

First, we need a conception of reasons for action. I shall start by making some comments on what it is for the reasons that you have to φ at *t* to count as "conclusive" or "overriding" reasons.[2]

Overriding or conclusive reasons seem to satisfy the following necessary condition. If an agent's reasons for φ-ing at a time *t* count as overriding or conclusive, then the agent *ought*, all things considered, to φ—where this occurrence of "ought" is what we could call the "objective practical 'ought'," indexed to the situation of this agent at this time *t*. The practical "ought," as I understand it, is neither moral nor non-moral, but reflects *all* reasons or considerations that bear on how the agent should act at *t*; it is, in other words, an *all-things-considered* "ought."[3] The "objective" practical "ought" differs from more "subjective" (or "information-relative") versions of the practical "ought" insofar as the facts that determine how the agent ought to act or choose at *t* are not limited to facts that the agent knows or is even in a position to know at *t*.

Indeed, it may be that the following more detailed connection holds between the objective practical "ought" and the notion of what we have overriding or conclusive reasons to do: it may be that the agent ought (in this sense) to φ at this time *t* if *and only if* φ-ing is at least *part* of the overall course of action that the agent has overriding or conclusive reasons to take at *t*. For our purposes, however, it does not matter whether this more detailed connection is exactly right. What we need is simply

[2] Some philosophers, such as Robert Audi (1993), define different senses for the two expressions "You *have* a reason to φ" and "*There is* a reason for you to φ." As I use these two expressions, however, they are synonymous and have exactly the same sense.

[3] For a more detailed account of the practical "ought," see Wedgwood (2007, chapter 4).

the necessary condition that I articulated above—that is, the thesis that if an agent has overriding reasons to φ at *t*, then the agent ought (in the objective all-things-considered practical sense) to φ at *t*.

This thesis makes it plausible that there is a more general connection between reasons and the practical "ought". At least so long as we restrict our attention to cases where it is possible for the relevant agent not to φ, it seems that if the agent ought to φ, there must be some explanation or reason why the agent ought to φ—and this explanation of why the agent ought to φ can surely be called a reason for the agent to φ.

In general, then, it seems plausible to follow the seminal suggestion of John Broome (2004), according to which a reason for an agent *x* to φ at a time *t* is a fact that plays what we could call the "*pro* φ-ing" role in an explanation of how *x* ought to act at *t*. Indeed, it seems plausible to me to suggest that all the facts about how an agent *ought* to act at *t* are explained by the following two factors:

a. Facts about what options are *available* to the agent at the time *t*.
b. The *reasons* for and against each of those options.

What is it exactly for a fact to play the "*pro* φ-ing" role in an explanation of how *x* ought to act at *t*? Again, I cannot attempt a full answer to this question, but I will try to clarify what is at issue by articulating some necessary conditions on reasons for action. As I mentioned at the outset, my goal here is to explore the implications of a distinctively *value-based* view of reasons for action. So I shall assume here that all reasons for action consist in facts about how the available options exemplify various *values*—i.e., facts about how these options are, in various ways, *good* or *bad*, or *better* or *worse* than each other.

More specifically, the conception of reasons for action that I shall assume here implies that a fact counts as a reason for you to φ if and only if that fact entails that φ-ing is in an appropriate way a *good thing for you to do*. In addition, I shall also assume here that goodness is reducible to betterness.[4] That is, for a course of action to be good is simply for it to be *better* than the appropriate benchmark of comparison. I shall not be able to say anything about this "benchmark of comparison" here, but presumably, if a course of action is better than this "benchmark," then

[4] For an argument in favour of this assumption, see Broome (1999).

that course of action must at least be better than *some* of the available alternatives.

There are, of course, many different ways in which a course of action can be as good as or better than some alternative. It is doubtful whether all of these different ways of being good can generate genuine reasons for action. If not, then only *some* of these ways of being good can ground any genuine reasons—in which case a full account of reasons for action would have to include a more precise account of which such ways of being good can do this. For present purposes, however, what has been said so far will suffice to give us a conception of reasons to work with in the rest of this discussion.

2.3 A GENERAL CONCEPTION OF MORAL REQUIREMENTS

What are "moral requirements"? It seems plausible to me that "required" here means neither more nor less than "needed," which seems to refer to what is *conditionally necessary*—in this case, what is necessary *for avoiding moral wrongness*. More precisely, then, we can give the following definition of what it is for an agent to be morally required to φ at a given time *t*:

> An agent *x* is *morally required* to φ at a time *t* if and only if, in every state of affairs available to *x* at *t* in which neither the *x*'s practical reasoning nor the *x*'s behaviour at *t* is *morally wrong*, *x* φ-s.[5]

If there is always a state of affairs at least in principle available to the agent at *t* in which neither the agent's practical reasoning nor the agent's behaviour at *t* is morally wrong, then this definition guarantees that the logic of what is morally required is so-called standard deontic logic (SDL)—in effect, the modal system known as KD.

SDL is controversial, but for our purposes the controversial features of SDL are not important. These controversial features flow from the right-to-left half of this definition (the thesis that *whenever* your φ-ing is necessary for you to avoid moral wrongness, you are morally required

[5] I have formulated this definition of moral requirements in this way to allow that moral wrongness can be exemplified in the agent's practical reasoning—that is, in the way in which the agent forms and revises her intentions or plans for action—as well as in behaviour.

to φ). What matters for our purposes is the left-to-right half of this definition—the thesis that if you are morally required to φ, then it will be impossible for you not to φ without either your behaviour or your reasoning being in some way morally wrong. This left-to-right half of the definition is much less controversial; and in the rest of this discussion, it is only this left-to-right half that I need to assume here.

Given what I said in the previous section about the all-things-considered practical "ought," it clearly need not always be the case that whenever one behaves otherwise than as one all-things-considered ought to, one's behaviour is morally wrong. All things considered, I ought to buy a new pair of shoes; but it surely need not be morally wrong of me not to buy any new shoes.

In what follows I shall assume a particular conception of moral wrongness—a conception that is inspired by some of what J. S. Mill says in *Utilitarianism* (1871, chapter 5). According to this conception, there is a fundamental connection between what is morally wrong and what is *blameworthy*. This connection does not imply that every case of moral wrongness is blameworthy. There are certainly cases where an agent's behaviour is morally wrong, but the agent is blameless because of having some appropriate *excuse* (such as blameless ignorance or the like). What an excuse of this sort does is to make it the case that even though the agent's behaviour is in some way morally wrong, the agent is not blameworthy for that behaviour.

However, it is only such excuses that can prevent wrongness from implying blameworthiness. If an agent's practical reasoning or behaviour at *t* is morally wrong, then, unless he or she has an excuse, the agent will be blameworthy for that reasoning or behaviour. So there is still a crucial connection between moral wrongness and blameworthiness.

In Section 2.1 I interpreted our central question as concerning two propositions, which I numbered (i) and (ii). The second of these two propositions is (ii)—the proposition that whenever we are subject to a non-trivial moral requirement, we have compelling or overriding reasons for conforming to it. If this second proposition is true, then presumably it is also true that whenever we are subject to a moral requirement, at least some of the reasons that we have for conforming to the requirement *ground* or *explain* our being subject to this requirement. These reasons could be called *moral reasons*. (If we use the term in this way, then even if you are morally required to φ, there could still be *non-moral* reasons that count in favour of your φ-ing—it is just that such non-moral reasons do not ground or explain your being subject to this moral requirement.)

If such moral reasons exist, it is plausible that there are also *other* moral reasons that do not ground any moral requirements. First, moral reasons that are themselves *overridden* by countervailing reasons presumably do not ground any moral requirement: for example, I might have a moral reason to keep my promise to meet you for coffee in the Common Room, but this reason will not ground a moral requirement if it is overridden by the reason that I have to help a badly injured person get to hospital without delay.

Secondly, some moral reasons are merely *supererogatory* considerations: for example, I have a moral reason to give up my expensive habit of buying opera tickets and to donate all the money to charity instead; but (I hope) this moral reason does not ground any full-blown moral requirement. In cases where a supererogatory option is available, the moral reason in favour of the supererogatory course of action does not "override" the non-moral reasons against that course of action. As we saw in the previous section, to say that your reasons for φ-ing are "overriding" or "compelling" is to say that φ-ing is not merely a good thing to do, but something that you—in the "all-things-considered" sense—*ought* to do. But even though the supererogatory course of action is not something that one ought *not* to do, it is surely also not something that one *ought* to do (in this "all-things-considered" sense). So the moral reason in favour of the supererogatory course of action does not override the reasons in favour of the alternative (even though the reasons in favour of the alternative presumably do not override the reasons in favour of the supererogatory course of action either). It seems clear, then, that these supererogatory moral reasons also do not ground any moral requirements.

These points are important for understanding the second of the two propositions with which we are concerned here, (ii). This proposition implies that the reasons supporting conforming to moral *requirements* are always overriding or conclusive; it does not imply the much stronger and less plausible proposition that absolutely *all* moral reasons are overriding.

2.4 THE MORAL/NON-MORAL DISTINCTION

It seems that the question that we are concerned with here would be much less interesting and important if absolutely *all* reasons for action were moral reasons.

Suppose that all reasons for action were moral reasons. Admittedly, we would still need to explain (i), the first of the two propositions with which we are concerned; that is, we would still need to explain why any moral reasons had the necessary qualities to ground a full-blown moral *requirement* (given that if you are morally required to φ, then φ-ing is necessary for avoiding wrongness—that is, for avoiding the kind of reasoning or behaviour that, at least in the absence of any excuse, would be blameworthy). But if one is morally required to φ, then it is not surprising that the *moral* reasons in favour of φ-ing outweigh or override all *moral* reasons for acting otherwise. So, if all reasons are moral reasons, it is also not surprising that the reasons in favour of conforming to a moral requirement override absolutely *all* countervailing reasons whatsoever, and so count as overriding or conclusive reasons; that is, it is not surprising that (ii), the second of the two propositions that we are concerned with, is true.

In this way, if there are no non-moral reasons pulling against the moral reasons, our central question would lose much of its apparent importance and urgency. So we need to make it plausible that there really is a distinction to be drawn between moral and non-moral reasons.

Within the framework of the conception of reasons that was articulated in Section 2.2, the distinction between moral and non-moral reasons corresponds to the distinction between moral and non-moral values. You have a moral reason to take a course of action if and only if that course of action is better than the appropriate benchmark with respect to some moral value—while there is a non-moral reason in favour of a course of action if and only if that course of action is better than the appropriate benchmark with respect to some non-moral value. But what exactly is this distinction between moral and non-moral value?

I shall not commit myself to any specific account of this distinction between moral and non-moral values here. Presumably, a full account of this distinction would have to say something about the connection between moral values and the reactive attitudes, like blame, which were mentioned in the previous section. However, defending any such account would require a thorough investigation, which would take us too far away from our main theme. Here, I shall just canvas a couple of interpretations of this distinction in order to fix ideas.

The first interpretation of the distinction that I shall canvas here is a view that I shall call "Sidgwickian consequentialism."[6] This view assumes a *consequentialist* conception of the value of action. That is, this view assumes that the only relevant kind of value that an action can have is a value that is derivative from, and wholly determined by, the value of the action's *total consequence*—where the "total consequence" of an action is simply the conjunction of all states of affairs that (i) *would* obtain if the action were to be performed, and (ii) *might not* obtain if the action were not performed.

However, Sidgwickian consequentialism differs from many forms of consequentialism in that it does not only focus on the *agent-neutral* value of consequences. Instead, Sidgwickian consequentialism implies that while *some* reasons for action are grounded in agent-neutral values, there are also *other* reasons for action that are grounded in *agent-relative* values.

Specifically, according to Sidgwickian consequentialism, there are two ways in which reasons for action are grounded in values. First, if the consequences of a course of action that is available to you are good in an *agent-neutral* way, this fact will ground a reason for you to take that course of action. Secondly, if the consequences of a course of action are *good for you*, that fact can also ground a reason for you to take that course of action. According to this Sidgwickian view, it is natural to identify the reasons that are grounded in the agent-neutral values with the *moral* reasons, and to identify the reasons that are grounded in agent-relative values of this sort with *reasons of self-interest*—where reasons of self-interest are plausibly categorized as non-moral reasons. On this Sidgwickian view, then, it is just a fundamental fact that the values that generate reasons for action include both agent-relative and agent-neutral values, and that in consequence we have both moral and non-moral reasons for action.

The second interpretation of the moral/non-moral distinction that I shall canvas here differs from Sidgwickian consequentialism in two fundamental respects. First, this second interpretation rejects the consequentialist conception of the value of actions.[7] Instead of implying

[6] This view is broadly inspired by Sidgwick (1907).

[7] Although this second interpretation is emphatically *non-consequentialist*, it is not *deontological* in the strict sense of advocating principles of rightness that have nothing to do with any notion of moral value or goodness. The value-based conception of reasons for action that I am assuming here makes it hard to reconcile any strictly deontological conception of morality with the belief in the existence of moral reasons.

that the value of an action is derivative from and determined by the value of its total consequences, this view permits actions to have an intrinsic value or disvalue of their own—so that it is possible, at least in principle, for there to be two actions that differ in value from each other even though the overall value of the total consequences of the two actions is exactly the same. (For example, if one of these actions *actively causes* a bad consequence while the other action merely *allows* a bad consequence to occur, then the first action may be worse than the second, even if the overall value of their total consequences is the same.) Secondly, according to this interpretation, all reasons for action are grounded in the intrinsic agent-neutral values that are instantiated by the acts themselves—there is no need to appeal to agent-relative values here.

More specifically, according to this interpretation of the moral/non-moral distinction, moral values are all fundamentally instantiated by *interpersonal relations*. For this reason, I shall call this interpretation the "relational view."[8] You have a moral reason in favour of a certain course of action if and only if that course of action would put you into a morally good relation to some person or persons; you have a moral reason against a course of action if and only if that course of action would put you into a morally bad relation to some person or persons.[9] For example, an action that involves saving a person's life will normally put you into a morally good relation to that person, while an action that involves killing a person will normally put you into a morally bad relation to that person.

In general, it seems plausible that the moral value or disvalue of a relation between persons is sensitive to the effect of that relation on the

[8] This idea is akin to the central idea of Scanlon (1998, 162), that moral reasons are grounded in the "value and appeal" of a certain "relation" that we might stand in to others; it is also akin to the conception of "second-personal" reasons that is advocated by Darwall (2006). However, this view is not committed to any sort of "contractualism" that is built around the idea of *principles that could not be reasonably rejected* or the like.

[9] There seem to be moral reasons that do *not* arise from the relations between the agent and any person who counts as either a *victim* or a *beneficiary* of the agent. For example, there seem to be moral reasons for not despoiling the natural environment, or for saving species or cultural traditions from becoming extinct; and acting against these moral reasons need not involve wronging any individual victims at all. Still, if one acts against these reasons, then one has arguably impaired one's relation to the whole "moral community," and it may be this that explains why there is a moral reason (as well as a non-moral reason) not to act in these ways.

values that ground the *other* reasons that the persons in question have. Morally good relations between persons often *help* at least some of those persons to exemplify or promote those values, while morally bad relations between persons typically *worsen* the position of at least some of those persons to exemplify or promote those values. For example, you have a reason to want to live a long flourishing life—a reason that is grounded in some value, such as the value of well-being or the like. So, if I save your life, the relation between us has helped you to exemplify this value to a greater degree, whereas if I allow you to die prematurely, this relation has, at least by comparison, put you in a worse position to exemplify this value. Proponents of the relational view could say that this difference is part of what explains why I stand in a morally better relation to you if I save your life than if I allow you to die prematurely—and also why I have a moral reason to save your life rather than to allow you to die prematurely.

As I said, in this section I have simply canvassed a couple of interpretations of the moral/non-moral distinction, for the purpose of fixing ideas about what is at stake in this debate.

In general, on any version of the value-based conception of reasons for action, there will be no great difficulty in explaining why you have moral reasons for action. As long as there are moral values, and some of the courses of action available to you instantiate these moral values to different degrees, this is enough to explain why you have moral reasons for action. Within this framework, there is a moral reason in favour of a course of action if and only if that course of action is *better*, with respect to one of these moral values, than the relevant benchmark.

So far, however, this framework can only explain why we have *some* moral reasons for action. So long as there are non-moral reasons as well as moral reasons, it could still be the case, for all that has been said so far, that moral reasons are extraordinarily *weak* reasons—reasons that sometimes conflict with non-moral reasons for action, and lose out to the non-moral reasons whenever any such conflict arises.

Intuitively, it seems that if moral reasons were such extraordinarily weak reasons, they would be *trivial*; and similarly, any moral requirements grounded in these moral reasons would be equally trivial. I shall assume that to explain why we are subject to "non-trivial moral requirements" (as I put it in (i), the first of the two propositions that I enumerated at the end of Section 2.1), we must explain why some moral reasons—specifically, the reasons that ground moral requirements—are not trivial in this way. On the contrary, there are cases in which we can have moral reasons to do something even though we have reasons for being tempted not to do it, and in many of these cases these reasons are not overridden or outweighed by those countervailing reasons.

To solve our problem, then, we must do much more than simply explain why there are moral reasons. We need to explain (i) why there are *non-trivial* moral reasons of this sort, which ground correspondingly non-trivial moral requirements, and (ii) why it is the case that whenever we are subject to such a moral requirement, we have an *overriding* or *conclusive* reason to act accordingly. We still do not yet understand how that can be the case.

2.5 THE REQUIRED/SUPEREROGATORY DISTINCTION

At this point it is illuminating to invoke a point that has perhaps been most clearly formulated by Sarah Stroud (1998).[10] As we have seen, there is a crucial difference between (a) moral reasons in general and (b) moral requirements. Some moral reasons are overridden or defeated by other reasons; and these defeated reasons do not ground any moral requirements. Moreover, moral reasons (even undefeated moral reasons) include *supererogatory* considerations, which also fall short of grounding full-blown moral *requirements*. So, what distinguishes between those (undefeated) moral reasons that ground full-blown moral requirements, and those that are mere supererogatory considerations instead? How should this distinction be drawn?

I propose the following account of the distinction: the difference between those actions that are *morally required* and those that are merely *supererogatory* consists precisely in the fact that the moral reasons in favour of the *former* actions are *overriding* reasons, while the moral reasons in favour of the *latter* actions are not overriding reasons, but only *sufficient* reasons for action.

This proposal is supported by the fact that whether we regard the reason for a certain action as grounding a moral requirement or not typically depends on the *strength* of the reasons (including non-moral reasons) *against* that action. In other words, when we judge whether a

[10] Closely related points have also been made by other philosophers, such as Shiffrin (1999).

moral reason grounds a moral requirement or is merely supererogatory, we always take account of the *other* reasons that the agent has. For example, if the cost in terms of the agent's other reasons for action is exorbitant, we are more likely to classify the moral reason as supererogatory rather than as grounding a moral requirement; whereas if the cost in terms of the agent's other reasons for action is relatively trivial, then even quite a weak moral reason may ground a full-blown moral requirement. The account of the required/supererogatory distinction that I have just articulated could give a straightforward explanation of this.

A further consideration supporting this account is the fact that it seems intuitively doubtful whether it could be appropriate to *blame* people for acting in a way in which they had a sufficient reason for acting. If this intuition is correct, then you can only be blameworthy for an action if you had an overriding reason not to perform that action. Given that the notion of moral requirements and moral wrongness has an intimate connection to blameworthiness, this intuition also supports the proposal that I am making in this section.

Clearly, this proposal explains the *second* of the two propositions with which we began—that is, it explains why it is the case that whenever you are subject to a moral requirement, you have overriding or conclusive reasons for conforming to it.

However, even if this proposal is correct, it only entails that *if* we are subject to any moral requirements, we have overriding reasons to comply with those requirements. This proposal does not entail that we *are* subject to any non-trivial moral requirements. It entails that a moral reason for a course of action cannot ground a moral requirement (as opposed to counting as a mere supererogatory consideration) unless it *overrides* all the reasons against that course of action. But why are there *any* non-trivial moral reasons that have the power to override all countervailing reasons in this way?

In effect, then, the second of the two propositions with which I started—the proposition that whenever we are subject to a moral requirement, we have overriding reasons to conform to it—turns out not to be the most challenging proposition for us to explain. It is rather the first proposition—the proposition that we are subject to non-trivial moral requirements at all—that is the most challenging problem for us to solve. To solve the problem, we will have to explain why there are any moral reasons that have the power to override all countervailing non-moral reasons for action.

2.6 ALTERNATIVES TO THE WEIGHING MODEL?

Some philosophers might be tempted to attack this problem by claiming that non-moral reasons never really conflict with moral requirements at all. The claim that non-moral reasons never conflict with moral requirements could be defended in many ways. For example, some versions of *eudaemonism* seem committed to a claim of this sort.[11] These versions of eudaemonism accept the following two theses: first, there is a reason for you to take a course of action if and only if that course of action contributes to your *well-being* to a sufficient degree; secondly, no course of action that involves failing to conform to a moral requirement can contribute to your well-being to a sufficient degree.[12]

An alternative version of this approach might concede that in some cases, courses of action that involve contravening a moral requirement may indeed make a significant contribution to the agent's well-being. Nonetheless, this alternative version of the approach might insist that in these cases there is no *reason* for the agent to take any such course of action. In this way, this version of this approach could hold on to the central idea that no genuine non-moral reasons ever conflict with any moral requirement. In some way, the fact about the course of action in question that would normally ground a reason for action is somehow "silenced" by the presence of the moral requirement; in other words, the moral requirement "disables" that fact from generating any genuine reason for action.[13]

For our purposes, these differences between the various ways of denying that there are ever any non-moral reasons against conforming to a moral requirement do not matter. All that matters is that these approaches are committed to the thesis that none of the apparent benefits provided by the agent's failing to conform to a moral requirement can ever ground a genuine reason against conforming to that requirement.

Given some extremely plausible assumptions, this thesis implies that there is a radical *discontinuity* among the apparent benefits in question. These apparent benefits *normally* ground a reason for acting; but however great these benefits may be, as soon as there is moral requirement

[11] For some discussion of the sort of eudaemonism that seems to have been assumed by Plato, see Wedgwood (2009b)

[12] Compare Irwin's (1994) interpretation of Plato as claiming that justice is a "dominant component of happiness."

[13] Compare McDowell (1979) and Dancy (2005, 41).

not to act in the way that secures those apparent benefits, they suddenly cease to ground any reason for action at all.

Such discontinuity claims have a grandiose sound to them, but they typically reveal themselves on closer inspection to face severe problems. The main reason for this is that *everything that matters in life comes in degrees*. The factors on which both reasons and values—including both moral and non-moral reasons and values—are grounded seem to be capable of varying in arbitrarily small increments from one case to another. For example, an act can be morally bad because it causes harm, but causing harm is a factor that can vary by tiny increments. For example, even if one harm is greater than a second, the degree to which the first harm is greater than the second may be extremely small. Indeed, it is not clear that there is any limit to how small the difference between a greater and a lesser harm can be; harms seem to be capable of varying by arbitrarily small increments.

We can explore the problems with these discontinuity claims by imagining the following spectrum of cases, C_1, … C_n. In each case C_i, there are two options available to you: A_i and B_i. There is at least one case C_k in which there is indisputably both a moral reason in favour of A_k and a non-moral reason in favour of B_k; and throughout this spectrum, the factors on which these reasons depend in C_k vary only very slightly between each case and its successor. But intuitively in the first case C_1 the moral reasons in favour of A_1 seem overriding, while in the last case C_n, the non-moral reason in favour of B_n seems clearly *not* to be overridden by any countervailing moral reason.[14]

To give a more concrete illustration of such a spectrum of cases, assume that in the case C_k where there is clearly both a moral reason and a non-moral reason, the moral reason in favour of A_k is grounded in the degree to which A_k's total consequences is *good for the world*, while the non-moral reason in favour of B_k is grounded in the degree to which B_k's total consequences are *good for the agent*. For example, suppose that in the first case C_1, option A_1 involves your saving millions of others from an agonizing death but not saving yourself from a paper cut, while option B_1 involves your saving yourself from a paper cut but not saving millions of others from an agonizing death. In each subsequent case C_i,

[14] This sort of spectrum of cases is inspired by those that are deployed by Williamson (2000, chapter 4).

the harm that *others* will suffer if you do B_i is slightly *less* than in the preceding case, while the cost that *you* will suffer if you do A_i is slightly *greater* than in the preceding case. Finally, in the last case C_n, option A_n involves sacrificing your own life to save the life of a distant stranger, while option B_n involves saving yourself but not saving the life of that distant stranger.

The following claims about these cases seem plausible:

a. In case C_1 you are morally required to choose option A_1.
b. In case C_n you are *not* morally required to choose option A_n.
c. There is a case C_j on this spectrum such that $1 \leq j < n$, where C_j is the last case in which you are morally required to choose the *A*-option.
d. There is at least one case C_k such that $j < k \leq n$, where option A_k is supererogatory but not morally required.

The views that we are considering in this section must all posit a radical discontinuity on this spectrum of cases immediately after case C_j—in spite of the apparent continuity of this spectrum of cases. But if there really is a radical discontinuity on this spectrum, then in the crucial respects this is not really a continuous spectrum of cases at all. C_j is a special point on the spectrum, since immediately after C_j you cease to be morally required to help others rather than yourself; and after C_j, the fact that one of the available courses of action will spare you physical harm suddenly jumps up from not grounding any sort of reason for action at all, to grounding a very significant reason. It seems impossible to explain why these factors would behave in such a radically different way in the two cases on either side of this special point, given how extremely similar to each other these two cases seem to be. So, views of this sort—which as we have seen include both the eudaemonist view and the silencing view—seem implausible.

So, it seems, there must always have been a non-moral reason in favour of the *B*-option in *all* of the cases on the spectrum, even in cases in which you were morally required to choose the *A*-option. The non-moral reason in favour of the *B*-option was really present in all these cases, conflicting with the moral reason in favour of the *A*-option; it is simply that in those cases this non-moral reason was not strong enough, and the moral reason was strong enough, to ensure that the non-moral reason in favour of the *B*-option was overridden by the moral reason in favour of the *A*-option.

It seems, then, that the best way to explain these phenomena is by invoking a simple *weighing* model. According to this model, in each case C_i, the totality of reasons in favour of A_i and the totality of reasons in favour of B_i both have a *weight*—where these weights can be at least partially ordered, and out of a given set of alternative acts {*A*, *B* . . .}, there is overriding reason in favour of *A* if and only if the totality of reasons in favour of *A* is *weightier* than the totality in favour of each alternative.

The weighing model can easily explain everything that it is plausible to say about this spectrum of cases, given the assumption that—at least in these cases—the weight of the reasons in favour of each act is an *increasing function* of the *harm* that that act prevents. In C_1, the reasons in favour of A_1 are clearly weightier than the reasons in favour of B_1. Then, as we go along the spectrum of cases, the reasons in favour of the *A*-option become gradually less weighty, while the reasons in favour of the *B*-option become gradually weightier, until we pass case C_j, after which point the reasons in favour of the *A*-option *cease* to be weightier than the reasons in favour of the *B*-option (which is not to say that after that point, the reasons in favour of the *B*-option become weightier than the reasons in favour of the *A*-option—indeed, in case C_k, where option A_k is supererogatory, neither the totality of reasons in favour of A_k nor the totality of reasons in favour of B_k is weightier than the other).

Some philosophers may be tempted to rescue the "silencing" approach by appealing to Joseph Raz's notion of an "exclusionary reason."[15] In the cases in which one is morally required to help others rather than oneself, perhaps there is—in addition to a moral reason to help the others—an exclusionary reason against even *considering* any countervailing reasons. However, it is doubtful whether there is always such an exclusionary reason in all these cases. If the case is very close to the line between the cases where helping others is morally required and the cases where helping others is morally supererogatory, then there may not be a reason against your considering the reason to help yourself rather than the others.

Moreover, even if there always is such an exclusionary reason, it will often be such a weak reason that it is itself overridden by countervailing

[15] See Raz (1999b, 35–48).

reasons or other factors. So this appeal to exclusionary reasons seems to presuppose that the exclusionary reason is itself a compelling or overriding reason. But why should this exclusionary reason count as compelling or overriding in this way? Clearly this is a question of exactly the same kind as the fundamental problem that we are seeking to address. So, even if there is always an exclusionary reason against even considering reasons for courses of action that involve contravening a moral requirement, it seems that the appeal to exclusionary reasons does not solve the fundamental problem that concerns us.

It seems, then, that we cannot solve our problem by denying that non-moral reasons ever conflict with moral requirements. This approach fails to recognize that both moral reasons and non-moral reasons come in degrees, and that some moral reasons are stronger or weightier than others. There are, at least sometimes, genuine non-moral reasons in favour of courses of action that involve failing to conform to a moral requirement.

Instead, then, I propose, when moral reasons override non-moral reasons, that it is simply because they *outweigh* them. But why do the moral reasons outweigh the countervailing non-moral reasons in cases of this kind?

2.7 MORAL VALUES AS BIG VALUES

Giving a complete theory of what determines the weight of reasons would be far too large a task for me to attempt here. So instead, in this final section, I shall enumerate some of the factors that seem to be involved in determining how much weight a given reason has. Once these factors are in view, they may help to make it seem less surprising and puzzling that moral reasons often outweigh countervailing non-moral reasons.

The kind of "weight" that I have just been talking about attaches to the *totality* of the reasons in favour of each of the available acts; it does not attach to any of the *individual* reasons belonging to this totality. Nonetheless, it may be that in at least some cases this weight can itself be regarded simply as the *sum* of the weights or strengths of all the various individual reasons in favour of the act in question.

Some philosophers will be tempted to think that this "additive" version of the weighing model is hopelessly crude; these philosophers may insist that the weight of the totality of reasons in favour of an act may in some cases be determined by *complex interactions* between the various individual reasons in favour of the act and in favour of its alternatives, and so cannot be identified with the sum of the weights of all the individual reasons. However, we could always say that the effect of these "complex interactions" is simply to generate *new* reasons in favour of (or against) some of the available acts, in which case the presence of these new reasons will obviously affect the sum of the weights of all the reasons. So it is not clear that it is such a crude mistake to assume an additive version of the weighing model.

According to this additive version of the weighing model, the crucial question will be what determines the weight of the *individual* reasons in favour of each of the available acts. As I explained in Section 2.2, I am working here within the framework of a value-based conception of reasons for action. According to this value-based conception, there is a reason in favour of an act if and only if that act is *better* (with respect to one of the relevant values) than the relevant benchmark—which implies that this act is better than at least some available alternatives.

Now, there is one factor that seems normally to make a crucial difference to the weight of a reason: at least other things equal, the *greater* the *degree* to which an action is better (with respect to the relevant value) than the available alternatives, the *weightier* the reason for that action. If there is only a *small* difference between the available acts with respect to a suitable value, then (other things equal) this value will ground only a relatively *weak* reason in favour of the acts that are better than the relevant benchmark with respect to that value. If there is a *huge* difference between these acts with respect to the value, then (other things equal) this value will ground a much *stronger* reason in favour of the acts that are better than the benchmark with respect to that value.

As I described this factor, however, it affects only the relative weights of different reasons that are grounded in the *same* value. If the reason in favour of an act *A* is grounded in the fact that *A* is better than the relevant benchmark with respect to a certain value, then other things equal, this reason would have been stronger if the degree to which *A* is better than the benchmark with respect to that value had been greater than it actually is.

One might attempt to extend this point to reasons that are grounded in *different* values. Even if a reason in favour of *A* and a reason in favour of *B* are grounded in two *different values* V_A and V_B, could it sometimes still be the case that the degree to which *A* is better than the benchmark in terms of V_A is greater than the degree to which *B* is better than the benchmark in terms of V_B? If so, then perhaps it might be true here too that other things equal, the reason in favour of *A* grounded in V_A is weightier than the reason in favour of *B* that is grounded in V_B.

However, there are grounds for doubting whether this is a promising approach.[16] It seems that it does not always make sense to compare the difference between two items with respect to one value with the difference between a pair of items with respect to another value. Is the difference in musical value between Beethoven's *Missa Solemnis* and Ravel's *Bolero* greater or lesser than the difference in hedonic value between the typical migraine that was suffered by Virginia Woolf and the agony of those who were crucified to death by the ancient Romans? It is not clear that such comparisons make any sense at all; even when all of the items being compared are acts, it is still not obvious that such comparisons can really be made. (Of course, we could make such comparisons if by saying that the difference between *A* and the benchmark in terms of one value V_A is "greater than" the difference between *B* and the benchmark in terms of a second value V_B, we meant no more than that the reason in favour of *A* grounded in V_A is weightier than the reason in favour of *B* that is grounded in V_B; but in that case, this comparison obviously could not *explain why* the first reason is weightier than the second.)

Still, there may be another factor that applies to reasons that are grounded in different values, which explains why reasons that are grounded in one value V_1 are often weightier than reasons grounded in a second value V_2. This factor is present most clearly in cases where the fact about the first value V_1 that grounds the first reason in a sense "subsumes" the fact about the second value V_2 that grounds the second reason.

We can illustrate this point by considering the Sidgwickian form of utilitarianism. According to this form of utilitarianism, a *reason of self-interest* in favour of an act *A* is ultimately grounded in the degrees to which the state of affairs of the agent's doing *A* is *better for the agent* than each of the relevant alternatives. By contrast, a *moral reason* in favour

[16] I am indebted to an anonymous referee for convincing me that this approach is not really promising.

of an act *B* is grounded in the way to which the state of affairs of the agent's doing *B* compares to each of the relevant alternative acts in terms of *sum* of the degrees to which the act is better or worse than each of the relevant alternatives for every individual person who will exist if the act is performed. In effect, as we might put it, the evaluative fact that grounds the moral reason is *partially constituted* by the evaluative fact that grounds the reason of self-interest.

For example, suppose that I set off a bomb, killing myself and hundreds of other people. According to the Sidgwickian form of utilitarianism, the reason of self-interest against this act is grounded in a fact that is constituted by the effects that this act has on my well-being, while the moral reason against this act is grounded in a fact that is constituted by the effects that this act will have on my well-being *and* on the well-being of everyone else as well.

The similar point will hold for many other consequentialist views besides utilitarianism. The degree to which the consequences of the available acts differ in how good they are for the individual agent is constituted solely by the effects of those consequences on the life of the agent alone, while the degree to which they differ in the relevant agent-neutral intrinsic value is constituted by their effects on the world as a whole (*including* the life of the agent). The point seems to lie behind an observation that Aristotle makes in the *Nicomachean Ethics* (1094b5–6):

> For even though the human good is the same for an individual and for a community, the good of the community is manifestly a greater and more perfect good both to attain and to preserve. To secure the good even for one person is worthwhile, but it is a finer and more divine thing to secure the good for a nation or for communities.

A similar point may well be true on the relational view as well. The relational view rejects the consequentialist conception of the value of courses of action, but it still interprets the moral value of an act as arising from the effect of the relationship that that act creates between two or more agents on the values that ground the reasons that those agents have. In effect, the moral value of this relationship between agents is partially constituted by the facts about values that ground those other reasons that those agents have.

On the relational view, then, the facts about moral value that ground the moral reasons that each individual has in this sense "subsume" the facts about non-moral values that ground the reasons that are possessed by several individuals. For example, according to this view, the moral reason that I have for saving your life if I can do so at negligible cost to myself is grounded in the fact that I put myself into a morally better relation to you if I save your life than if I fail to do so; and that fact is partially constituted by whatever fact about values (such as the value of your well-being) grounds the reason that you have for wanting to go on living.

If it is true that moral values "subsume" other values in this way, then, as we might put it, moral values are "*big values*"—values that put together the facts about other values to make a bigger and more encompassing evaluative fact.[17] According to the tentative proposal that I wish to make here, the fact that moral values are "big values" in this way is at least part of what explains why moral reasons are often so weighty that they outweigh all countervailing non-moral reasons.

At the same time, it seems that there must be other factors affecting the strength of a reason besides how big the difference is between acts (in terms of a single value), and how "big" the relevant value is. As we have seen in cases of supererogation, moral values do not always outweigh countervailing non-moral reasons. My reason to promote my own well-being seems much weightier than the reason that I have to take equally burdensome steps to promote the well-being of a stranger who is wholly unknown to me, and the reason for this seems to have to do with factors of a completely different kind. For example, this factor might be something like the *motivational centrality* of the value of one's own well-being in the mental life of normal human beings; or perhaps it has to do with the *proximity* of the state of affairs that fundamentally exemplifies this value to the practical situation of the agent in question.

At all events, the central proposal that I wish to make here is that the weight of moral reasons is explained, at least in part, by the fact that moral values are big values in this way. However, this proposal cannot quite be the whole of the story. For this fact about the moral values to give reasons for *us*, we need to have the right sort of *opportunities*: an appropriate array of alternative acts must actually be available to us, in the many situations in which the moral reasons emerge as overriding or compelling reasons for action. This is because it is only when the acts that are *available* to us differ sufficiently greatly in their moral value

[17] For the contrast between "big values" and "small values," see Raz (1999b, 30).

that we have the particularly weighty moral reasons that I have been describing here.

What explains why the opportunities available to us are so often of this sort? This is ultimately an empirical question, about the explanation of a contingent fact about human beings. Still, the following speculation seems plausible: the explanation has to do with the *radically social* nature of human beings—that is, with the fact that the evaluatively significant aspects of our lives consist to such a great extent in our interactions with other persons. If there were intelligent agents who were less deeply social than we are, they might rarely be in situations in which the available acts differed greatly in moral value, while they might more often be in situations in which the available acts differ greatly with respect to non-moral values; and it seems that these non-social agents would be subject to less demanding moral requirements than we are.

At all events, an approach along these lines seems to be the most promising way for a value-based conception of reasons for action to explain why it is that we human beings are subject to non-trivial moral requirements, which correspond to overriding reasons for action. The acts available to us often differ significantly in their moral values; and moral values are, as I put it, particularly big values. This may be what explains why we so often have moral reasons for action that are weighty enough to outweigh all the reasons that we have for acting otherwise.[18]

REFERENCES

Audi, Robert (1993). *The Structure of Justification* (Oxford: Clarendon Press).

Broome, John (1999). "Goodness is reducible to betterness: the evil of death is the value of life," in John Broome, *Ethics Out of Economics* (Cambridge: Cambridge University Press): 162–73.

—— (2004). "Reasons," in *Reason and Value: Essays on the Moral Philosophy of Joseph Raz*, ed. R. J. Wallace, Michael Smith, Samuel Scheffler, and Philip Pettit (Oxford: Oxford University Press).

Dancy, Jonathan (2005). *Ethics Without Principles* (Oxford: Clarendon Press).

Darwall, Stephen (2006). *The Second-Person Standpoint* (Cambridge, MA: Harvard University Press).

Irwin, T. H. (1994). *Plato's Ethics* (Oxford: Clarendon Press).

[18] An earlier version of this essay was presented at the University of Arizona's Normative Ethics Workshop, and at the Central Division Meeting of the APA in Chicago. I am grateful to those audiences, and also to Jonathan Dancy, Michael Otsuka, Joseph Raz, and an anonymous referee for this volume, for helpful comments.

McDowell, John (1979). “Virtue and reason,” *The Monist* 62: 331–50.

Mill, J. S. (1871). *Utilitarianism*, 4th edition (London: Longmans, Green, Reader, and Dyer).

Price, Richard (1787). *Review of the Principal Questions of Morals*, 3rd revised edition (London: T. Cadell).

Prichard, H. A. (1912). “Does moral philosophy rest on a mistake?,” *Mind* 21 (81): 21–37.

Raz, Joseph (1999a). *Engaging Reason* (Oxford: Clarendon Press).

—— (1999b). *Practical Reason and Norms*, 3rd edition (Oxford: Clarendon Press).

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

Shiffrin, Seana Valentine (1999). “Moral overridingness and moral subjectivism,” *Ethics* 109: 772–94.

Sidgwick, Henry (1907). *The Methods of Ethics*, 7th edition (London: Macmillan).

Stroud, Sarah (1998). “Moral overridingness and moral theory,” *Pacific Philosophical Quarterly* 79: 170–89.

Wedgwood, Ralph (2007). *The Nature of Normativity* (Oxford: Clarendon Press).

—— (2009a). “Intrinsic reasons and reasons for action,” *Philosophical Issues* 19: 342–63.

—— (2009b). “Diotima’s eudaemonism: Intrinsic value and rational motivation in Plato’s *Symposium*,” *Phronesis* 54: 297–325.

Williamson, Timothy (2000). *Knowledge and Its Limits* (Oxford: Clarendon Press).

3

Scanlon’s Promising Proposal and the Right Kind of Reasons to Believe[1]

MARK VAN ROOJEN

> . . . Here is a proposed analysis. When I say “I promise to be there at ten o’clock to help you,” the effect is the same as if I had said “I will be there at ten o’clock to help you. Trust me.” In either of these utterances I do several things. I claim to have a certain intention. I make this claim with the clear aim of getting you to believe I have this intention, and I do this in circumstances in which it is clear that if you do believe it then the truth of this belief will matter to you . . . Finally, I indicate to you that I believe and take seriously the fact that, once I have declared this intention under the circumstances, and have reason to believe that you are convinced by it, it would be wrong of me not to show up (in the absence of some truly compelling reason for failing to appear).
>
> T. M. Scanlon, *Promises and Practices*, p. 211

In a nutshell, the above paragraph gives Scanlon’s analysis of promising. His account of the obligation to keep promises contrasts with more traditional accounts that rely on promising’s status as a kind of social practice along with some sort of duty to support such practices. He suggests that the duty is better explained using a principle that governs actions broader than promising, Principle F:

[1] Thanks to Ken Baynes, Matt Bedke, Aaron Bronfman, Brad Cokelet, Roger Crisp, Janice Dowell, Jennifer Haley, Reina Hayaki, Mark Heller, Harry Ide, Robert Johnson, Clare LaFrance, Chris Malony, Sarah McGrath, Sean McKeever, Jennifer McKitrick, Joe Mendola, Howard Nye, Hille Paakkunainen, Andrew Reisner, Holly Smith, David Sobel, Mark Timmons, Robert Van Gulick, and Stephen White for helpful discussion. Thanks also to members of the audience at the Tucson Normative Ethics Workshop in January 2012 and at a presentation at Syracuse University in February 2012 for helpful questions and comments. I owe thanks as well for helpful comments to two reviewers for Oxford University Press, one of whom outed himself as Drew Schroeder and the other of whom remains anonymous to me. Special thanks are due to Sarah Stroud and Ralph Wedgwood for helping me to see how to put my point in the appropriately strong way, and to Neal Tognazzini for the discussion that led me to this essay.

> If (1) A voluntarily and intentionally leads B to expect that A will do X (unless B consents to A's not doing so); (2) A knows that B wants to be assured of this; (3) A acts with the aim of providing this assurance, and has good reason to believe that she has done so; (4) B knows that A has the beliefs and intentions just described; (5) A intends for B to know this, and knows that B does know it; and (6) B knows that A has this knowledge and intent, then, in the absence of special justification, A must do X unless B consents to X's not being done.[2]

Principle F can apply even if one has not made a promise. Promising, however, relies on Principle F in an especially tight way—it exploits Principle F to provide the assurance that triggers its application. In other words, mutual recognition by promisor and promisee of the force and relevance of principle F is necessary to the ability of promises to provide assurance to the promisee of the promisor's future compliance. One succeeds in promising (according to this account) if one gets another to feel so assured in virtue of believing that one is motivated not to violate the very obligation to follow through that Principle F codifies.

The resulting account gives us two sorts of explanation. First, it gives us an explanation of why, as we believe, we generally have an obligation to keep our promises unless released by the promisee. Second, it gives us a vindicating account of how the institution/practice of promising could arise without self-deception or irrationality. This second level of explanation relies on the first. It is because the actions involved in promising do in fact trigger the relevant obligation that it makes sense for us to make and abide by promises. Of course the second level depends on other claims as well, such as the assumption that people often enough comply with their obligations.

The two explanations together allow Scanlon to avoid relying on a more standard sort of justification for the keeping of promises—one which invokes an obligation to comply with, support, or uphold going beneficial social practices.[3] Such theories may be pure or hybrid; each sort requires there to be an existing social practice of promising to ground our obligations to keep promises. Pure social practice theorists ground our obligation to keep promises in the fact that there is a social

[2] Scanlon (1990), p. 208. Principle F also makes an appearance in Scanlon (1998a) at 304, and in (1998b) at p. 245.

[3] Such as the accounts of promising in Rawls (1971) and Hume (1888).

practice of promising that enables us to obtain various benefits, including the benefits of compliance with promises, but also the benefits of assurance that something will be done. And they typically couple this with some normative principle requiring that one comply with just or useful social practices. Hybrid theorists, notably Kolodny and Wallace,[4] suggest that social practices have some such grounding role while also suggesting that our current obligations to keep promises derive their normative force from something like Principle F in addition to such practice-based justifications.

In support of their view Kolodny and Wallace (following Pratt, Anscombe, Warnock, and Hume)[5] have pressed a circularity objection against Scanlon's account. A person succeeds in promising only if she creates the expectation in clause (1) by persuading the promisee to believe she will do what is expected because Principle F applies. But Principle F will apply only if the promisee has the required expectation. It thus looks like the person forming the expectation must already have that very expectation for her to be justified in coming to expect the promisor will do as promised because not doing so would violate Principle F. In the absence of that expectation a lack of follow-through would not violate Principle F.

This line of argument can be used in support of pure social practice accounts, but Kolodny and Wallace employ it in support of a hybrid account, which grafts Scanlon's Principle F-based explanation of our obligation to keep promises onto a prior practice-based justification for keeping promises. Hybrid strategies, as I think of them, employ some other principle to underwrite the promisee's rational expectation that the promisor will follow through and then to bolster that expectation with reasoning employing Principle F. Kolodny and Wallace's particular suggestion is to use a practice-based principle to underwrite the initial expectation on the part of the promisee that the promisor will follow through. It does so by deploying a general obligation to do one's part to uphold just beneficial practices. However, once the practice-based expectation is in place, Scanlonian reasoning can both reinforce the obligation to comply and the expectation

[4] Kolodny and Wallace (2003). Kolodny and Wallace themselves draw on Pratt (2003).

[5] Anscombe (1978); Pratt (2003); Warnock (1971), 99–101; Hume (1888), Bk III, Part 2, ch. 5.

that promisors will follow through. Since this suggestion relies on the status of promising as a social practice to obviate the objection it would not allow Scanlon to escape reliance on practices at a fundamental level.

Somewhat ironically, the quotation at the start of this essay sets up Scanlon's own answer to objections of this ilk. The particular version is one pressed by Elizabeth Anscombe. Even so, many philosophers have found his reply wanting. Not only did Kolodny and Wallace write their critical paper after having read the reply, but others have followed them in thinking Scanlon's account problematic for this very reason. This seems to be the dominant view, though these things are hard to judge.[6]

Dominant or not, I think the conclusion is mistaken. In this essay I will argue that any vicious circularity in Scanlon's account can be avoided without recourse to promising's status as a social practice or to any duty to comply with social practices from which one benefits. In fact, I will argue that the quoted paragraph contains nearly everything needed to provide a satisfactory answer to the circularity objection. In service of that overall thesis I will argue that the "circularity" objection really boils down to a right-kind-of-reasons objection[7] employing assumptions about the right kind of reasons for belief. And I will argue that so construed the objection can be met because other perfectly respectable lines of reasoning would also run afoul of the assumptions leading to the objection. Along the way I will note that there is also a circularity worry for the practice view—and one more troublesome than that leveled at Scanlon.

My essay has three main parts. In the first I note some advantages which Scanlon's proposal has over its practice-based competitor. In the second I discuss the Kolodny/Wallace objection and explain why it boils down to a right-kind-of-reasons objection. And then, in the third part I provide some reasons for thinking that so construed, the objection can be overcome.

[6] See, for example, Tognazzini (2007), pp. 209–13.

[7] Readers may be familiar with such objections from the literature on the Kavka's Toxin puzzle or Pascal's wager. For those who are not familiar I give an explication in Section 3.2.

3.1 Some Advantages of Scanlon's Promising Proposal Here over Practice Accounts

The basic idea of social practice accounts is that we have some kind of duty to comport ourselves with just and/or mutually advantageous social practices. Coupling this with the uncontroversial assumption that the institution of the social practice of promising is generally beneficial, yields a duty to keep our promises. Rawls, for example, invokes a Principle of Fairness, which he believes obligates the voluntary beneficiaries of just social practices to do their fair share to support them by complying with the obligations that such practices place on them.[8] It would not be hard to come up with other variants of such vindicating social-practice theories of promising.[9] Common to such theories will be the reliance on a certain kind of social fact—that there is a social practice of promising from which most/all benefit—to ground the duty to obey one's promises. This social fact then interacts with a general obligation to do one's part in supporting beneficial social institutions to generate the specific duty to keep the promise in question. For in promising, the promisor obtains a benefit from the social practice—being able to provide assurance to the promisee—thereby incurring an obligation to do his or her part to uphold that practice. If that is the justification for keeping promises, we would expect that all of us who are part of the practice and who do our bit to uphold it are equally in a position to complain when a promise is broken. The unfaithful promisor would be taking advantage of us to obtain a benefit without reciprocating. Relatedly, the expectation of apology for infractions would be shared generally, as would the right to forgive those who transgress.

As Scanlon (1990, p. 221) notes, this does not fit very well with our actual attitudes towards promising and towards transgressions of the duty to keep promises. The obligation to keep a promise seems directed toward the person to whom we make that promise. Although others can fairly criticize infractions, the promisee has special status. It is to her that

[8] Rawls, (1971) pp. 344–50.

[9] In addition to Rawls, Scanlon cites Hume's *Treatise of Human Nature*, bk. III, Pt. 2, chap. 5 as using an ideal observer theory along with the thought that ideal observers would disapprove of defection from beneficial social practices to suggest a similar verdict with respect to breaking promises.

performance is owed. And if the promisor is to be forgiven for breaking the promise, it is the promisee whose forgiveness matters, to whom apologies are owed, and to whom reparations must be made. The rest of us may judge the transgressor harshly, remonstrate with him about following through, and blame him, but we cannot take over that role from the promisee.

Scanlon's account is better placed to explain this feature of the obligation to keep promises. Principle F itself is a duty we would think of naturally as directed towards those whose expectations we raise. In the particular transaction that creates assurance there are two parties whose concerns and interests are of focal concern: the person who for whatever reason wants to provide assurance, and the person who wants and seeks it.[10] When the assurance provider fails to do as expected, that will presumably be because his interests are better served by not following through. So he has nothing to complain about. But the interests of the recipient will not normally have been taken into account—for if they were she would presumably have been willing to release the assurance provider of the obligation to follow through.[11] Thus it is no surprise that Scanlon's Principle F-based rationale fits well with the directed nature of promising.

This provides his theory with an explanatory advantage over purely practice-based accounts of promising. And that is one reason why Kolodny and Wallace find their hybrid proposal attractive; insofar as hybrid accounts include Principle F in the basis for complete promissory obligations, they can explain the special status of promisees.

But there is another advantage that the Principle F-based account has over its practice-based rivals, and this is one that hybridization will not mitigate. Practice-based accounts are subject to a particular circularity or regress worry of their own—one that hybridization will not defuse. They require that there in fact be an existing social practice that one fails to support when one breaks one's promises. While different social practice theories vary as to the exact nature of the normative requirement, they all rely on a previously existing social practice to get off the ground. Rawls

[10] I am speaking here of assurer and recipients rather than promisors and promisees because, as noted earlier, Principle F can generate obligations even when no promise is made.

[11] Complications involving the dead and absent are fully noted; it might be worth adding some clause about when we know they would release us to Principle F.

invoked something like a duty of fairness not to free-ride on the efforts of others in creating a generally beneficial social practice.[12] Hume seems to invoke general benevolence to underwrite impartial disapproval of activities that might undermine useful social practices.[13] But these diverse normative principles all require an actual social practice already in place to generate an obligation in a particular case. You are not free-riding on the efforts of others if there is no socially created practice of promising on which the success in providing assurance depends. And you are not undermining a useful social practice in the absence of an actual practice of the sort we are trying to explain and rationalize. As far as I can see, this cannot be fixed by tweaking the relevant obligations to apply to possible but not actual social practices. You will not be free-riding if you do not take what would be the first of many steps towards creating a useful practice, at least in the absence of an agreement to create such a practice. And it is not obvious why disinterested benevolence would motivate us to disapprove of someone who does not take steps that would be part of a useful practice if only there only was such a practice.

So far as I can tell, that leaves the advocates the option of first giving a non-vindicating account of the origins of the practice of promising and then arguing that once it is in place we have good reason to comply, because now that the practice is up and running it creates an obligation that did not previously bind those who made promises. Luckily for us, enough previous promisors mistakenly thought themselves bound that we can now say there is an existing social practice of providing assurance by making promises. And now, given that fact, we for the first time have an obligation to keep the promises we make. This is not an incoherent story, but it does give us reason to be curious about how the previous participants got themselves confused enough to keep their promises in the absence of any obligation to do so.[14]

[12] Rawls (1971), pp. 344ff.

[13] Hume (1888), pp. 516–22.

[14] The worry is particularly pressing for Kolodny and Wallace, who write: "In promising, one signals one's recognition of a moral obligation not to undermine or exploit not merely some social practice of agreement making, but specifically the social practice of promising: the practice that consists in participants' signaling that they adhere to a policy of fidelity because they recognize moral obligations to adhere to the policy" (123). This rules out explaining the genesis of promising by piggybacking it on a distinct but similar practice. They go on to consider a charge of bootstrapping related to my worry. But their answer really only speaks to reasons why we might continue with such a practice once it is in place—something I concede they can provide.

It is a virtue of the Scanlonian account of promising that it requires no such error-theoretic story. And that is because his account puts no requirements on the past. Reasons for accepting and keeping a promise are grounded entirely contemporaneous with and subsequent to the promising itself. As a result, it does not require promising already to exist before one can succeed in making a binding promise.[15]

Could a Kolodny and Wallace style hybrid theory employing Principle F obviate this sort of circularity worry? Not without giving up their solution to the other circularity objection—the one which they press against Scanlon and which motivates their own account in the first place. Kolodny and Wallace accept the hybrid proposal precisely because they think there must a Principle F-independent ground for the expectation that promises will be kept before principle F can be invoked. And they employ the standard social practice account to provide that prior expectation. If that account is itself unable to provide a reason for the expectation in the absence of a pre-existing social practice, it will not do the work of filling the alleged need to provide a Principle F-independent reason for the expectation. And if hybridizing the social practice account with principle F enables that account to provide a vindicating story of the genesis of the social practices on which the social practice-based expectation relies, the Principle F-based story should have been sufficient on its own.

3.2 The Circularity Objection is a Right Kind of Reasons for Belief Objection

As I already noted, whatever circularity there may be in Scanlon's account, it is not the kind of circularity that generates an infinite regress in time. Rather, the circularity, if there is any, is logical. One's reasons to keep a promise depend on another's reasons to expect one to keep the promise, and yet that other person's reasons in turn depend on your having reason to keep the very promise in question. Kolodny and Wallace put the objection thus:

[15] Ulrike Heuer (2012) attributes a circularity objection to David Owens in a forthcoming book on promising. Owens' objection as outlined by Heuer seems to involve the same worries I am explicating here.

(a) In order for B to be assured—by appeal to F, in the way in which Scanlon describes—that A will do X, A must first be obligated by F to do X.

(b) In order for A to be obligated by F to do X, condition (1) of F must first be satisfied.

(c) In order for condition (1) of F to be satisfied, B must first be assured that A will do X.

(d) In order for B to be assured—by appeal to F, in the way in which Scanlon describes—that A will do X, B must first be assured—by some other means—that A will do X. (Kolodny amd Wallace, pp. 131–2)

It is easy to see how this sets up Kolodny and Wallace's hybrid solution to the problem; practice-based reasons can provide assurance by another means.

Several of the premises in this argument and the conclusion make claims about what must "first" obtain for a person to succeed in providing assurance by employing Principle F. But nothing in Principle F requires any kind of temporal priority so long as all its clauses are simultaneously satisfied. Nor is there anything built into using Principle F to generate assurance that requires temporal priority as opposed to contemporaneous mutual recognition that the promisor has met the conditions of Principle F and thereby obligated herself. The alleged regress must involve a different worry.

The real worry is that being assured is a kind of cognitive attitude—one that has to be based on evidence to be well-grounded. It will only be well-grounded if there is some basis for thinking the expectation—that is, the belief that the promisor will follow through—is justified. Promising is supposed to ground that belief in the thought that the promisor would be running afoul of Principle F if she does not keep the promise. But for that to be so, the promisee needs to have the expectation—that is, the belief that the promisor will follow through. What we have, then, seems to be a belief that is justified only if it is accepted. And the worry is that rational belief requires awareness of the justification to be rational, a justification that will not be assured until the promisee believes the proposition in need of justification.

The point here is one about justification. As Kolodny and Wallace nicely put it, the challenge is: "Can an account of the obligation of fidelity spell out clearly the conditions that trigger that obligation, and would the obligation survive under explicit awareness on the part of both promisors and promisees of the conditions that provide its basis?" (p. 134). Understood in this way, it will not be an answer to point out that so long as the promisee fools the promisor into thinking that she believes the promisor will follow through, she will have a reason to expect follow-through. Nor would it be an answer to point out that if the promisee mistakenly and irrationally comes to believe that the promisor will follow through, Principle F will trigger an obligation. So we need to think about whether the promisee's expectation can be justified without someone making an error. And that requires asking whether the promisee can have the right kind of reason to believe that the promisor will follow through where that reason invokes Principle F.

3.3 DOES THE PROMISEE HAVE THE RIGHT KIND OF REASON TO BELIEVE?

To put the objection most sharply, someone who doubts that the promisee has the right kind of reason to believe might liken the belief in question to some sort of wishful thinking. It cannot, the objectors might say, be a reason to believe that the promisor will follow through that it would serve my purposes—the purposes of securing a binding promise—that I believe the promisor will. Our desires should not be able to make rational a belief that would not be rational in their absence. As Nishi Shah among others has noted, reasons for belief must be such as to make sense of our taking these reasons as relevant to determining whether the belief in question is true.[16] Believing something just because we would like it to be true is not that kind of reason.

3.3.1 Is this kind of reasoning compatible with believing the conclusion non-accidently true?

We should note, however, that the situation is unusual in one important respect. The promisee is deciding whether to believe that the promisor will do as suggested on the basis of her coming to have that very belief. Absent worries about the promisor's good intentions, the belief will be

[16] Shah (2003).

true if she comes to have it (at least if she is right about the other conditions she is in). And the promisee and promisor are in a position to know this. A principle of belief adoption which allowed one to believe such self-validating contents would thus be reliable in the sense that it would not lead one to adopt false beliefs. That by itself does not tell us whether the promisee has the right kind of reason to believe the proposition in question, but it does reduce the urgency of the worry at least to this extent. Forming beliefs in this way will not lead us away from the truth.

In the remainder of this essay I will in effect be arguing that the reasoning in question is not really a form of wishful thinking or even relevantly like it. The promisee is not forming the belief that the promisor will follow through *just* because she wants it to be true, even though the fact that she wants it to be true enters into her reasoning in a certain way. I will pursue the issue by looking at other sorts of reasoning that have a similar structure to that involved in generating the objection here.

3.3.2 Reasons for optimism

3.3.2.1 A real-life example

I finished graduate school and took my present job before cellular phones were widely available at a reasonable price. Jennifer and I packed all of our belongings into the largest U-Haul truck available, filling it to the point that it sagged on its suspension. We drove west out of New Jersey to Nebraska, with a stop in Illinois where my parents lived. In Illinois we picked up an old Renault sedan which my parents no longer wanted, since Jenny and I did not own a car. I drove the van, and Jenny followed in the Renault. Since I knew the route, Jenny was planning just to follow the large and hard-to-miss truck. Many miles from my parent's house and probably long before I realized it, Jenny discovered that we had become separated. At some point I also came to that realization.

We each stopped to ponder our predicaments, and each eventually figured out that it would make the most sense to call my parents at their home which we had left earlier in the day. We called them about 5 minutes apart, and we were then able to pass messages along to one another. Using this method we were able to coordinate a meeting and to continue on our way as planned.

I believe Jenny and I were correct to think it made sense to call my parents as a strategy for getting reconnected. But its making sense depended on further beliefs that we each had—among them the belief that each of us had that the other was going to call my parents because that is what it made sense to do. It would have been of no use calling my folks if they were not going to have a way of communicating with Jenny. And they were not going to have that unless she called them. I believed she would call them because it would make sense for her to call if she believed that I was going to do so as well. So it appears that each of us formed the belief that the other would call, based on an inference from a thought about what the other could reasonably believe. And the belief that we each thought it reasonable for the other to have itself depended on a belief which would be reasonable only if that very belief were true. That is just the feature that is somewhat troubling about the promisee's belief that the promisor will follow through in Scanlon's account of promising.

3.3.2.2 Game-theoretic parallels

The driving story just told is an instance of a kind of problem/situation that has received a lot of attention in the decision-theory and game-theory literature. The situation Jennifer and I were in is structurally similar to that of players in a coordination game invented by Thomas Schelling:

> You are to meet in New York City. You have not been instructed where to meet; you have not prior understanding with the person on where to meet; and you cannot communicate with each other. You are simply told that you will have to guess where to meet and that he is being told the same thing and that you will just have to make your guesses coincide. (Schelling, 1960, p. 56)

A majority of test subjects who played this game in 1960 chose Grand Central Station as the place to meet. If they were asked also to converge on a time, noon was the dominant choice (Schelling, p. 55, fn. 1). Schelling called the salient coordination points "focal points," and argued that for many such coordination games and in their real-life analogs sufficiently creative individuals could reasonably form hypotheses about focal points that would attract attention from a cooperating partner.

What is important for my purposes is that these choices seem rational and that the rationality of the choices depend on the rationality of cognitive states which the players might adopt. It is reasonable for a player to expect other players to choose a certain strategy on the assumption that it is reasonable for that player him or herself to choose a complementary strategy. But at the same time, it may be reasonable for a given player to choose that complementary strategy only if she correctly predicts that the other player will choose its complement. Here we seem to have a prediction about another party's response to a prediction about that very judgment, where the rationality of each state of mind reflexively depends on the rationality of what the other party believes and does. And that is the same feature that raised issues for Scanlon's account. If the jointly cooperative coordination point can be reasonably predicted by a player in this game, there might in principle be nothing wrong with the kind of reflexivity involved in Scanlon's account of promising.

Someone might resist my use of these examples by pointing out that the rationality of choosing such a focal point does not really depend on a full belief that one's partner will converge on the same point. All you really need to think is that the point in question is the most likely, or among the most likely, places to meet. Or perhaps, the objection might go, you do not even need a belief that it is more likely so long as one's own credence that they will choose the relevant focal point is higher than one's credence in their choosing any other alternative. I do not think that this changes the moral of the story. The rationality of the choice in these cases still depends on what cognitive states it is rational to form in the circumstances in question. I only have reason to raise my credence in my partner's going to Grand Central Station if I can rationally expect her to raise her credence in my coming to raise my credence in that very hypothesis. Credences, beliefs about probability, and beliefs about relative probability are all cognitive states. If a putative reason is the wrong kind to support belief—including beliefs about probability—it should also be of the wrong kind to support raising one's credence.

3.3.3 Lessons from cooperation and focal points

At this level of abstraction, such cooperative games give us grounds for optimism—but we would like more than that. It would be helpful to be able to note some features which focal-point reasoning and promising have in common. As it happens, there are several important similarities. Firstly, both sorts of reasoning rely mostly on being able to reason about how the other party will reason, and not so much on empirical information to the effect that the other party has reasoned in that way.

Secondly, the parties have a common goal, meeting in one case, assuring the promisee in the other. Thirdly, the parties can each reason to a series of actions and intentions which will yield an optimal result for both of them. Fourthly, each course of reasoning involves a self-reinforcing feedback loop such that increased confidence in one's expectation that the other will take a jointly beneficial course of action bolsters one's rational confidence in one's grounds for having that level of confidence.

In the paradigm, cooperative games parties are not in a position to communicate with one another. This means, I think, that the actual credence of one party in some claim about what the other party will do does not enter directly into making the other parties expectations and actions rationally permissible. What matters for the justification of one party's expectations and beliefs is what they can rationally think the other party will do and believe given what is common knowledge between the parties.[17] So in these cases, it is not my coming to believe that my partner will go to Grand Central Station that makes it rational for her to go there; it is her believing that it is rational for me to believe that, and her thought that I will probably do and believe what it is rational for me to do and believe. This may to some extent defuse the worry that the parties' reasoning involves illegitimate bootstrapping; the rationality of my belief does not depend on my believing it. Rather, it depends on the features of the situation that make it rational for my partner to believe that it is rational for me to believe it. So it is also for the promisor and promisee once a promise has been made. Let me name the promisor Olga and the promisee Emma for ease of explication of this and subsequent points. Emma's reason to expect follow-through from Olga is based in part on Emma's own reasons to think that Olga has reasons to think that Emma herself will rationally expect the promisor to follow through. Emma need not have additional information to the effect that Olga has made the relevant inferences.

The second and third similarities are straightforward. Emma and Olga each would like to do what is needed to see that Emma has assurance of Olga's future follow-through, and they can both see what it would take from each of them to provide that assurance. The fourth commonality is also present in the promising case—once Emma has some reason to

[17] Robert Sugden, "A Theory of Focal Points." *Economic Journal* 105 (1995), 533–50.

think that Olga will follow through that reason feeds back to provide more reason for Olga to follow through and more reason for Emma to expect that Olga will.

These similarities provide us with clues as to how the relevantly similar beliefs can be justified where promising is at issue. Given the positive feedback loop of support we need only a relatively weak reason to form an expectation to get the process of reasoning started. Just as in 1960 the salience of Grand Central Station provided an initially weak reason to think one's partner would try to meet there, we need only an initially weak reason to think the promisor might do as promised. Both parties, if they think things through, can see what is needed to achieve their common goals. The promisee wishes to be assured that the promisor will do the action or actions that are the subject of the promise. The promisor wants to assure the promisee of just that as well. They can also both see that if they can jointly generate a rationally grounded expectation of the promisee that the promisor will do as promised, Principle F will generate grounds to expect just that. Thus they can see generating the conditions that trigger Principle F as an optimal joint strategy for providing the assurance they both want.

As it happens, Olga is best placed to provide Emma with the needed weak reason to expect Olga to do what they both wish to assure Emma that Olga will do. Olga can just decide to do whatever it is that she intends to promise to do and to communicate the intention to do it to Emma. Assuming that background norms of veracity are in play, that communication will generate some reason to expect Olga to do as she intends. And the expectation can be strengthened if Olga further communicates that the intention in question is not the sort that she is likely to reconsider.

That will, in turn, plausibly put Olga in the domain governed by Principle F, at least if the expectations Principle F invokes include creedal states short of full belief. And, consistent with its spirit it should cover such states of mind. The reasons it would be wrong to lead someone to fully expect one to do something and then not follow through are of a piece with the reasons it would be wrong to lead them to a similar merely probable expectation and then not follow through. Perhaps the strength of these reasons varies with the strength of the reasons given to form the underlying expectation. But the reasons need not be conclusive to make Principle F relevant. Once it is relevant, that relevance will give a promisor like Olga yet more reason to do what she intended and communicated to the promisee. And that itself will be yet more reason for Emma to expect her to follow through. Thus, once the promisee has some reasonable confidence that the promisor will follow through, the promisor has a Principle F-based reason to do what she said she would, and we have our needed positive feedback.

I do not see anything wrong with this story as one way to get the promise off the ground.[18] But it is worth noting how this is similar and yet different from Kolodny and Wallace's hybrid strategy.[19] Like their account, the present suggestion starts with a reason on the part of the promisor to follow through and do the action of the sort that he promises he will do which is not itself rooted in Principle F. Where they use a social practice-based expectation I invoke an expectation grounded in a norm of veracity. This reason then creates a reason to expect the promisor to do what the promisee is hoping he will do, and that fact then creates the Principle F-based reason to do as promised. But, unlike the practice-based reason that triggers Principle F according to Kolodny and Wallace, the reason here requires no Principle F-independent obligation to do what the promisor promises. Norms of veracity generate obligations to say what one believes to be true, but not an obligation to make true what one says.[20]

There is one more thing to learn from the game-theoretic parallel. There need be no independent decision to do as one will promise to do prior to making the promise itself. In coordination games it does not seem that we really need to generate the different steps which ground

[18] Some readers might worry that deciding to do what the promisee wants in order to trigger the promissory obligation is the wrong kind of reason to intend to do an action, perhaps because they buy into Parfit's state-based/object-based distinction as drawing the line for the right kinds of reason to intend (see, for example, Parfit, 2001). Or perhaps they just think the upshot of the toxin puzzle (Kavka, 1983) is that reasons to intend must always be reasons to do the action in question. I am not persuaded, for two reasons. There are good reasons to doubt this as an account of the relevant distinction (see Schroeder, 2012). And, in any case, the promisor probably has plenty of good reason to do the action she wishes to promise to do. One such reason is that the promisee wishes the promisor to do that action. There is more to say here, but I think these materials suggest a successful strategy to avoid this line of objection. I thank Howard Nye for raising it.

[19] Matt Bedke helped me notice this.

[20] Kolodny and Wallace (pp. 146ff.) themselves argue this in an attempt to show that such a norm cannot do the work that Scanlon would need it to do.

the expectation that we and our partner will wind up at a focal point in any particular temporal order. In other words, we do not have first to form some small expectation that our partner will head to Grand Central Station. Note that this gives us some reason to go to Grand Central ourselves, go on from there to note that this in turn gives our partner more reason to go there, raise our expectation, and so on. We can all at once, as it were, grasp that the optimal joint strategy is to go to Grand Central Station, and this line of reasoning terminates with an expectation that one's partner will go there and an intention to go there oneself. There seems to be nothing irrational in that kind of immediate response to the structure of the situation.

If you can already see that a process of good reasoning will lead to a certain conclusion, that itself is a reason to endorse that conclusion. Weak and therefore plausible reflection principles on individual reasoning endorse this claim.[21] Things do not seem to be much different where two people who trust each other to answer to reasons are concerned. That might suggest something like the following for joint deliberation:

> If (1) two parties are going through a mutually recognized process of practical reasoning that depends on each recognizing and responding to the other's reasons at various stages, and if (2) they both see that rational error free continuation of that reasoning will lead to certain beliefs and intentions on each party's part, then they can each rationally move directly to those beliefs and intentions.

In the focal-point reasoning highlighted in Shelling's examples, this seems like the right thing to think, and it has some attraction as a more general claim. Similarly with our promising example. If two people A and B can see that A by communicating an intention to Φ will give B reason to raise her expectation that A will Φ, and that this very expectation will make Principle F generate a further reason for A to Φ and B to expect A to Φ, and A does in fact communicate the intention to Φ to B, then B has reason to conclude that A will Φ, and A has reason to expect B to conclude just that. Both parties can reasonably come to the relevant

[21] For discussion of some variants, see Briggs (2009).

judgments and intentions all at once, without need for any temporally prior step. A promise need involve no more than its reflexively recognized expression to generate the relevant reasons to believe and to act.

3.4 A WORRY AND A REPLY

You might be worried that the argument shows too much. If a statement of intention can be ratcheted up into a promise by the applicability of Principle F, what is to stop every similar statement about one's future conduct to obligate one to follow through? I leave the house, telling you I am going to the hardware store, only to realize half way there that the store does not carry the parts I need. Instead I will have to get them at the auto parts shop. But I have raised your expectation that I will go to the hardware store, so now Principle F requires me to follow through. I have inadvertently promised that I am going to the hardware store, so now I must go. That is crazy![22]

It is crazy, and not a consequence either of Scanlon's original explanation nor of anything I have claimed in the course of arguing that it eludes the circularity objection. Principle F already includes clauses to distinguish promising from such a situation. You probably were not hoping to be assured that I was going to the hardware store, and if you were I probably did not know it. And even if you did want that and I did know it, I normally would not have said what I did *in order to* provide assurance that I would go there. Principle F is thus normally not applicable.

But suppose I did say it in order to provide you with such assurance. Principle F may then give me a reason to go to the store. But that will not yet make my statement of intention into a promise to go there. Promises are distinguished from other Principle F-triggering grounds by their exploitation of principle F in providing that assurance. So if I figured only that you would be assured in virtue of my veracity and the likelihood that I knew where I was going, and not in virtue of thinking that Principle F would constrain me to follow through, I have not yet made a promise. If I was aiming to provide you with desired assurance I might still have incurred an obligation even though that will not be

[22] Thanks to Stephen White for raising this objection.

enough to have incurred it by promising. But that result is not all that crazy. Ordinary statements of intention will not generate similar obligations, so we need not worry about inadvertently promising or creating Principle F-based obligations.

3.5 CONCLUSION

The main point of this essay has been to assess the robustness of Scanlon's account of promising in the face of a circularity objection pushed most forcefully against him by Kolodny and Wallace. I have argued that the right way of construing the objection is as an objection to the reasoning that a promisee might go through in coming to expect a promisor to follow through with what is promised. In particular, I have argued that we should construe it as a complaint that the promisee's reasoning is of the wrong sort to support belief or similar cognitive states. From there I have shown that promising as conceived by Scanlon shares many features with other perfectly acceptable forms of practical reasoning also involving the expectation that a cooperating partner will act in some particular way. And I have argued that the shared features vindicate the reasoning involved as being of the right sort.

REFERENCES

Anscombe, G. E. M. "Rights, Rules and Promises," *Midwest Studies in Philosophy, Volume* III*: Ethical Theory* (1978), 318–23.

Briggs, Rachael. "Distorted Reflection," *Philosophical Review* 118 (1) (2009), 59–85.

Heuer, Ulrich. "Promising: Part 1," forthcoming in *Philosophy Compass* 7 (2012), 832–41.

Hume, David *Treatise of Human Nature*, ed. L. A. Selby-Bigge (Oxford: Clarendon Press, 1888; reprinted, 1958).

Kavka, Gregory. "The Toxin Puzzle," *Analysis* 43 (1) (1983), 33–6.

Kolodny, Niko, and Wallace, Jay. "Promises and Practices Revisited," *Philosophy and Public Affairs* 31 (2) (2003), 119–54.

Owens, David. *Shaping the Normative Landscape* (Oxford: Oxford University Press, 2012).

Parfit, Derek. "Rationality and Reasons," in D. Egonsson, J. Josefsson, B. Petterson, and T. Rønnow-Rasmussen(eds.), *Exploring Practical Philosophy* (Burlington, VT: Ashgate, 2001) 17–39.

Pratt, Michael. "Promises and Perlocutions," in *Scanlon and Contractualism*, ed. Matt Matravers (London: Frank Cass & Co., 2003).

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

Scanlon, T. M. "Promises and Practices," *Philosophy and Public Affairs* 19 (3) (1990), 199–226.

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

—— "Promises and Contracts," in *The Difficulty of Tolerance* (Cambridge, MA: Harvard University Press, 1998), 234–69.

Schelling, T. C. *The Strategy of Conflict* (Cambridge; Harvard University Press, 1960).

Schroeder, Mark. "The Ubiquity of State-Given Reasons," *Ethics* 122 (2012), 457–88 .

Shah, Nishi. "How Truth Governs Belief," *The Philosophical Review* 112 (4) (2003), 447–82.

Robert Sugden. "A Theory of Focal Points." *Economic Journal* 105 (1995), 533–50.

Tognazzini, Neal. "The Hybrid Nature of Promissory Obligation," *Philosophy and Public Affairs* 35 (2007), 203–32.

Warnock, G. J. *The Object of Morality* (London: Methuen, 1971).

4

Evil Achievements and the Principle of Recursion

GWEN BRADFORD

Achievements, many people think, play a central role in the best kind of human life. Indeed, many people think that achievements are of such importance that they are worth pursuing at the expense of serious sacrifices. But just what makes achievements valuable? What makes running a marathon, writing a book, climbing Mount Everest, or decoding the human genome worth doing? There is a great deal to be investigated about these questions. In this essay I look at one particular aspect of the value of achievements by examining the implications of a highly plausible axiological principle: the principle of *Recursion*. Doing so not only sheds light on the value of achievements, but also on how best to construe Recursion.

Let us suppose that an *achievement* has (among other features) the following structure: *a process that culminates in a product*. Writing a novel, for example, consists of a process of writing, which culminates in a product, the novel.

One question that we might have about the value of achievements is this: which of these two components of achievements, the process or the product, is the *source* of the value for achievements? A natural thought is that achievements are valuable primarily because of the value of their *products*: developing the cure for cancer, for example, seems valuable because it results in saving many lives.

However, having a valuable product is clearly not necessary for an achievement to have value. Many paradigmatic achievements have products that are of zero value: the product of winning a chess match, for example, is having the pieces configured in a certain way, and this state of affairs alone hardly seems to make winning a chess match valuable. Similarly, being in a location 26.2 miles away does not have any particular value—certainly not enough to account for the value of running a marathon.

It is plausible that achievements are all valuable (at least in part) in virtue of features that are characteristic of achievements as such—that is, that are common to all achievements. So if some valuable achievements have products of zero value, it is not very promising to suppose that the product is what grounds the value of achievements. Instead, this observation supports the view that the process is the source of value of achievements. Since it is plausible that achievements are all valuable in virtue of some features that are characteristic of them as such, this supports the view that the value of *all* achievements is grounded in the process.

Call this the *Process Thesis*: *the process is a source of the intrinsic value of an achievement*. Of course, there may be other features that shape the value of any particular achievement (no doubt having a valuable product contributes to the overall value of an achievement), but we at least hold that at least *some* of the value of all achievements is sourced in the process.

In this essay I will assume that the Process Thesis is true,[1] and turn instead to consider achievements that have products of positive or negative value. In particular, my focus will be products that have some negative value. How might the value of the product contribute to the achievement's value?

One highly plausible axiological principle that might be of service here is the principle of *Recursion*. Recursion as a value-theoretic principle captures the thought that, most broadly, *it is good to love the good*. In other words, given some thing of intrinsic value, the love of this valuable thing is itself of intrinsic value. Recursion also makes the corresponding claims that loving the bad is bad, and hating the good is bad, and hating the bad is good. Loving, here, is broadly understood as encompassing a range of pro-attitudes, including *desiring, wishing for, taking pleasure in*, and, most relevant to the issues at hand, *pursuing*. Hating, likewise, encompasses a range of con attitudes: *avoiding, destroying, being pained by*. (Since pursuit is what is relevant to the issues here, it will be largely the focus in what follows.)

The general idea of Recursion has been rigorously formulated by Thomas Hurka:[2]

[1] I defend the Process Thesis more extensively in "The Value of Achievements," *Pacific Philosophical Quarterly*, 94 (2013): 204-24.

[2] *Virtue, Vice, and Value* (Oxford: Oxford University Press, 2001), henceforth abbreviated as VVV. Hurka makes use of Recursion insofar as his view is that *virtue* is loving the good and

(1) For some base intrinsic good, G, the loving of G is itself of positive intrinsic value; the hating of G is of negative intrinsic value.
(2) For some base intrinsic evil, B, the hating of B is of positive intrinsic value; the loving of B is of negative intrinsic value.[3]

Thus, much like recursion in logic, value-theoretic Recursion concerns an operation performed on a base. We begin with a base intrinsic good, or a base bad, as the case may be. The operation is a particular attitude toward the base. Recursion applies to "attitudes" more broadly: particular pro and con attitudes, such as loving and hating. The particular aspect of Recursion that is relevant here is *pursuit* of a base good or bad.[4] Taking pursuit as our form of loving, the pursuit of an intrinsic good is *itself* of intrinsic positive value. Similarly, the opposite of pursuit, avoidance, of intrinsic bad is of positive value. Mismatches of pursuit of values—pursuit of bad or avoidance of good—are, accordingly, of *negative* value.

Hurka's formulation of Recursion is helpful because it crystallizes this value theoretic principle whose subscribers, in one form or another, appear to have included Aristotle, Brentano, Moore, Nozick and others.[5] It is arguable that a proto-recursive view is held by Aristotle, who holds that at least some good things—activities—have a fitting or "proper" pleasure.[6] And, famously, that the pleasure "completes" a good activity, presumably thereby augmenting its value, as "the bloom on youths."[7]

hating the bad, and vice is thus its opposite. We need not agree with this analysis of virtue, however, to make use of Hurka's account of the principle of Recursion, which has independent plausibility as such.

[3] This is a modified version of the principle of Recursion as presented by Hurka (VVV).

[4] I will follow Hurka and take "pursuit" here as a *form* of *loving* (VVV, pp. 11ff).

[5] Hurka attributes versions of Recursion to these philosophers, as well as Hastings Rashdall and W. D. Ross (VVV, pp. 23–8).

[6] From *Nicomachean Ethics*: 'since activities differ in degrees of decency and badness, and some are choiceworthy, some to be avoided, some neither, the same is true of pleasures; for each activity has its proper pleasure. Hence the pleasure proper to an excellent activity is decent, and the one proper to a base activity is vicious; for, similarly, appetites for fine things are praiseworthy, and appetites for shameful things are blameworthy' (1175b25–30). It is unclear whether or not Aristotle's view is that the pleasures in the respectively good and bad activities are *intrinsically* good or bad. Hurka takes this to be an expression of Recursion (VVV, p. 23), but also notes that there are differences between Aristotle's views in *Nicomachean Ethics* and fully-formed Recursion. Moreover, it seems that Aristotle's view is that the pleasure that "completes" the activity is *caused* by the activity, when the activity is done the right way: he says it arises "as a consequent end" (1174b30), which seems slightly different from the relevant pleasure for Recursion. In any case, there is something of the spirit of the principle in Aristotle.

[7] For example, see *Nicomachean Ethics*, 1174b20–25, 1174b30.

Chisholm attributes a version of Recursion to Brentano, which is certainly in line with Brentano's definition of intrinsic value as that which loving is *correct*. Brentano, on Chisholm's reading, also holds that correct love of the good is also good.[8]

A recursive view is also often associated with G. E. Moore, since a great deal of his discussion of organic unities involves the positive value of wholes involving loving the good or beautiful and hating the bad or ugly.[9] Moore's view, on one reading, is that the value accrued in situations involving a pro-attitude toward a good is accrued by the *whole* composed of the attitude and base good. He explicitly says that the mere existence of a beautiful object on its own is of very little value, and the pro-attitude, on its own, would have hardly any value.[10] In this respect, Moore's discussion departs from Recursion as we understand it here, where the base indeed has significant positive intrinsic value.

There has been some debate over whether this is the best reading of organic unities: is the value accrued to a *whole*, or is it rather the case that the elements themselves *change* in value when put together?[11] The latter approach is the *conditional* approach to organic unities. This view ultimately denies that there is indeed a *bona fide* organic unity, holding instead that the value of the parts changes, rather than being a whole which has different value. Hurka's account of Recursion is, in a sense, a relative of this view—the loving itself accrues value, rather than the attitude-plus-object whole.

As we will see, the view I ultimately endorse here challenges Recursion in this respect. After considering what value theoretic implications we would hope to see for evil achievements, it turns out that the best construal of Recursion is indeed as *bona fide* organic unity.

[8] Chisholm, *Brentano and Intrinsic Value* (Cambridge: Cambridge University Press, 1986), p. 63.

[9] Moore, *Principia Ethica* (Cambridge: Cambridge University Press, 1971 [1903]), pp. 191ff.

[10] *Principia Ethica*, p. 190.

[11] For some of this debate, see Ben Bradley, "Is Intrinsic Value Conditional?" *Philosophical Studies* 107 (2002): 23–44; Johan Brännmark, "Three Kinds of Organic Unity," in Wlodek Rabinowicz and Toni Rønnow-Rasmussen (eds.), *Patterns of Value: Essays on Formal Axiology and Value Analysis*, vol. 2 (Lund: Lund Philosophy Reports, 2004), pp. 80–94; Thomas Hurka, "Two Kinds of Organic Unity," *The Journal of Ethics* 2 (1998): 299–320.

Achievements, according to my earlier sketch, because of their process-product structure, involve the *pursuit* of some product. In cases where the product is of non-zero value (that is, some positive or negative value), then, Recursion entails that *something happens to the value of the process,* in virtue of its being the pursuit of some good or bad. If the product of the achievement is of positive value, Recursion entails that the pursuit of this product is good. If the product of the achievement is of negative value, Recursion entails that the pursuit of this product is *bad.*

The latter case is puzzling: according to the Process Thesis, the *positive* value of achievements is in part grounded in the *process*. But if the product is of negative value, and if Recursion is true, then *what happens to this positive value*?

I am going to consider three possible ways of understanding how Recursion works, and the implications of these construals for the value of achievements—evil achievements in particular. Doing so will reveal not only a good way to account for the value of achievements, but also the best way to understand how exactly the principle of Recursion works.

Here are the three construals: (1) the value of the process is *polarized* by the value of the product; (2) the value of the process is augmented or diminished *summatively*; (3) there is an augment or diminishment of value which is not strictly summative, but rather an instance of genuine *organic unity*. I will examine each of these and ultimately conclude that the final option, though not without its drawbacks, is the best.

According to the polarization construal, Recursion simply governs the valence of value of the pursuit of a good or bad. When the product is of positive value, the process of an achievement is unaltered in value by Recursion. When the product has a *negative* value, such as an elaborately planned murder, the value of the process is *negative*. It thus undergoes a reverse polarization of value.[12]

[12] On this construal, then, Recursion has no effect on the *degree* to which the value of the process is of positive or negative value (this is governed by the features of the process itself), but Recursion determines the valence. A more sophisticated version of the polarization construal could also shape degree, i.e., the worse the product, the worse the negative value of the process. But this version would fail to help the polarization construal from the objection I subsequently consider.

However, polarization fails in two ways. First, a large part of the motivation for accepting Recursion as a general value theoretic principle is the thought that the pursuit of a good is *all the more good* in virtue of its being a pursuit of a good. So a pursuit that is otherwise valuable should accrue *more* value in virtue of its being the pursuit of some good. But the polarization construal of Recursion does not have the resources to account for such a bonus. Recursion governs only the *valence* of value on this view. The process is *already* good, and so Recursion has no further influence over its value.

Second, the polarization construal gives an overly restrictive account of achievements that have products of *low-level* evil, such as practical jokes, or even moderate-level negative value, such as art heists. These not-too-evil evil achievements, such as practical jokes, seem to have (at least some) positive intrinsic value. If the positive value of the process undergoes reverse polarization when the product is of negative value, as this approach to Recursion will have it do, then these achievements actually have *no* positive value whatsoever. This is schoolmarmish and overly restrictive. So we should reject the polarization construal of Recursion in favour of one that allows us to retain some positive value to petty evil achievements.

The second construal I now turn to consider does just this. The *summative* construal of Recursion holds that Recursion generates a value, V_R, that is summed with the value of the process—to be precise, V_R is summed with the value that the process would have, independently of the value of the product in which it culminates. V_R is positive when the value of the product (V_{prod}) is positive, negative when it is negative, and, we can further elaborate, proportionate to the value of the product in magnitude. Thus, for example, a very evil product results in a large negative value for V_R, a product of small positive value results in a small positive value for V_R, and so on.

Given the value that the process would have, independently of its product having any value, positive or negative (call this the *independent value* of the process, $V_{proc\ indep}$), V_R is added to $V_{proc\ indep}$, giving us the actual value of the process, V_{proc}. So according to the summative construal of Recursion:

$$V_{proc} = V_{proc\ indep} + V_R$$

Assuming the value of an achievement is the value of the process plus the value of the product:

$V_{ach} = V_{proc} + V_{prod}$
$V_{ach} = (V_{proc\ indep} + V_R) + V_{prod}$

The summative view thus gives the following desirable results. Achievements that have products of positive value accrue additional value. In a petty evil achievement, the value of V_R and V_{prod} will be a small negative value—not enough to outweigh the positive value of $V_{proc\ indep}$. Thus petty achievements come out as of *positive* value on this construal, which is just the result we were hoping for. In a significantly bad achievement, where the value of the product is very bad, the achievement is bad *overall*, since the negative value of the product outweighs the positive value of the process.

However, the summative construal fails to give a satisfactory answer in certain kinds of cases. In these cases, the achiever *believes* that the product he is pursuing has positive value, but turns out to have *no* positive value. For example:

> *Cure-All.* Dr Cure-All is a highly competent scientist working on the vaccine for a debilitating disease. Even though there was every indication as he was working on this medicine that it would indeed inoculate against the disease, years later it is discovered that the vaccine actually induces a much more horrible fatal disease in all who take it.

Dr Cure-All's valiant pursuit of the vaccine is earnestly effortful and competent, and thus presumably valuable in the same way as the process of any achievement. The product of his efforts—namely, the deaths of many people—is of *negative value*.

What does the summative construal of Recursion say about the value of Dr Cure-All's achievement? Since the product is negative, V_R is negative. Thus $V_{proc\ indep}$ will be diminished. In this case, V_R is greater (*ex hypothesi*) in negative value than $V_{proc\ indep}$ is positive. Thus the summative construal tells us that the value of the process is *negative*. This means that all the hard work and dedication of Dr Cure-All toward curing a debilitating disease actually has negative value—the summative construal entails that the process has no positive value at all.

This does not seem right. It seems there should be at least some positive value in Dr Cure-All's pursuit of what he believed, albeit falsely, was a cure for a debilitating disease. The implausibility is even more evident in contrast to the efforts of Dr Kill-All:

Kill-All. Dr Kill-All labors intensely to design a terrible toxin, which he intentionally administers and thereby kills many people.

Surely Dr Kill-All's efforts are worse than Dr Cure-All's efforts. But the summative view as stated has no way of distinguishing the value in these two cases: in virtue of their equally evil products, the summative construal of Recursion says that Dr Kill-All's efforts are just as bad as Dr Cure-All's.

A large part of the motivation of accepting the principle of Recursion is to capture the thought that there is indeed *positive* value in efforts such as those of Dr Cure-All—that is, that the *pursuit* of the good is itself good, even if the good toward which it aims fails to come to fruition. The summative construal is not able to distinguish the value between the efforts of Dr Cure-All and those of Dr Kill-All.

We can revise the summative construal. The best way to understand Recursion is looking to the value of the *intentional object* of pursuit, rather than the actual, resultant product.[13] By "intentional object" I mean that toward which the pursuit is aiming. The intentional object of Dr Cure-All's pursuit is what Dr Cure-All is aiming for: namely, the vaccine.

Not only does this version of the view best capture the distinction between Dr Cure-All and Dr Kill-All, but it also gives better results in other, non-achievement cases where the principle of Recursion is at play. For example, taking pleasure in non-existent goods, products that never come to fruition, or wishing for goods; i.e., in contrast to taking pleasure in non-existent bads, pursuing bads that never come to fruition, or wishing for bads. If we construe the principle of Recursion as taking the *intentional object* of the pursuit as the element that figures in the value generated by Recursion, this gives a much better account than if we look to the value of the product of the achievement.

So, on the revised summative construal, the value that the process accrues from Recursion is generated from the value of the intentional object of the process, V_{obj}, rather than the value of the actual resultant product. To be precise, intentional objects themselves do not have any actual value, so V_R is generated from the value that the object of the

[13] I believe that this is very close to what Hurka has in mind.

intentional object would have, were it to come about. It is this value, V_R, which is the amount by which $V_{proc\ indep}$ is augmented or diminished.

The overall value of the achievement on the revised summative construal will be a matter of the value of the actual product and the value of the process. The value of the process, as I have just described, is the $V_{proc\ indep}$ plus V_R, where V_R is positive just in case the intentional object of the process would be positive, and negative just in case the intentional object would be negative. To the value of the process we then add the value of the actual product, which gives the value of the achievement overall. This allows the summative construal to avoid counterexamples such as Cure-All.

But even with this refinement, the summative construal runs aground. It entails that we can have a case like the following.

Zero Case. Villain is striving with great ingenuity toward a heinous end, which he succeeds in achieving. Accordingly, $V_{proc\ indep}$ has a positive value (according to the Process Thesis), V_{prod} has a negative value, and V_R is thus of negative value. In this case, it just so happens that $(V_R + V_{prod})$ is the same amount in *negative value* as the $V_{proc\ indep}$ is of *positive value.* Thus the value of the achievement overall is *zero*.

To be precise, in the Zero Case:

$$(V_R + V_{prod}) = -(V_{proc\ indep})$$
$$\text{Thus } V_{ach} = 0$$

In Zero Case, V_{prod} and V_R are negative to a degree such that they *exactly negate* the positive value of the process. As a result, the overall value of the achievement is zero. It is entirely *neutral in value*. This means that an ingenious, heinous, evil murderous scheme that succeeds in bringing about a terrible result has no value whatsoever, positive or negative. This seems inaccurate. Surely there is something of negative value here.

Now, one might reply that the case is not so bad. After all, there really is something of negative value: namely, the product. It just so happens that its negative value is exactly counterbalanced by the positive value of the process. So even though there is indeed an achievement here that has no value at all, we are reacting to the value of its components.

I do not find this particularly satisfying, since we are evaluating the value of the *achievement*, which is zero. To really get to the heart of the matter we would have to take a stand on the status of values of the individual components when they are summed within a single thing or state. (Do these values still "count" in some sense, even though they are outweighed?) Nonetheless, resolving these issues is unnecessary, since there is a more troublesome version of the zero case for the summative account.

*Zero Case**. Villain is striving with great ingenuity toward yet another heinous end, which, by luck, does not come about. Thus, V_{prod} is zero, and V_R is negative. But it just so happens that V_R is the same amount in *negative value* as the $V_{proc\ indep}$ is of *positive value.* In other words, V_R is as bad as $V_{proc\ indep}$ is good. So the overall value is zero.

To be precise, in Zero Case*:

$$V_{proc\ indep} = -(V_R)$$

$$\text{Thus } V_{proc} = 0$$

$$\text{And } V_{prod} = 0$$

As a result:

$$V_{ach} = (V_{proc\ indep} + V_R) + V_{prod}$$

$$V_{ach} = 0 + 0$$

$$V_{ach} = 0$$

In Zero Case*, the negative value of the intentional object of the achievement generates an amount to be deducted from the independent value of the process that is *the same in amount as the independent value of the process*. As a result, the overall value of the process of the achievement is zero. It is entirely *neutral in value*.

This means that the attempt of an ingenious, heinous, evil murderous scheme is equivalent in value to a sneeze: zero. This seems incorrect. And here is an argument to support this position: we can take what we consider to be appropriate reactions to things as evidence of their value. If a pro attitude is appropriate, this is an indication that something is of some positive value; if a con attitude is appropriate, this is some evidence that something is of negative value. It seems clear that some kind of con attitude is appropriate toward the evil pursuit—even if unsuccessful. In any case, it certainly seems clear that *neither* a pro nor con attitude is appropriate toward the sneeze: this is evidence that it is of neutral value. But it is clear that *some* attitude is appropriate toward the murderous scheme. This is evidence that there is some sort of value to the murderous scheme. But the summative view fails to give us this result. Rather, it tells us that the value of the process of an evil attempt is zero. This seems incorrect. Moreover, the intuition that pursuit of evil is evil is the very sort of intuition that motivates the principle of Recursion in the first place, and the summative construal fails to capture this.

One might be inclined to say that the explanation for why it seems that some attitude is appropriate in Zero Case* is because of the *instrumental value* that the pursuit of an evil process would have, typically. Even though there is no *actual* bad (or good) here, had the process come to fruition, there would have been some bad in the world. Thus the process has a kind of instrumental value that is negative, and this is the source of our reaction.

This is a possible way that we could go if we really want to support the summative construal, and were there no other appealing construals of Recursion I might be inclined to find this response adequate. But I am far more inclined to say that it is not an adequate response: indeed, there *really is something in the world that is bad* in Zero Case*—the attempt of an evil—and something that is good—the difficult and competent pursuit. These things are not merely instrumentally good and bad, and are not merely possibly good and bad: their value is actual.

Moreover, a good part of the theoretical motivation behind the principle of Recursion is to capture the negative value of the pursuit of evil. And the summative construal fails to capture this. Indeed, it says that there is *nothing* that is evil in Zero Case*.

To pinpoint the source of the problem with the summative construal, note that according to the summative construal the value that is accrued or lost via Recursion is the actual value of the process. That is, Recursion governs the *value of the process*. This, it seems, is the source of the problem with zero cases: Recursion can result in the process having very little or even *no* value whatsoever, and thus, in some cases, the achievement has *zero* value.

I propose, then, that we adopt the third construal of Recursion, which allows the process to retain all its value. On this construal, Recursion is such that the value that is accrued or lost is not gained or lost by the process—that is, the value located in the process is *not* what is affected by Recursion. Instead, Recursion governs the value of the *whole*—process *and* product together—as a unit. In addition to the process, and the product, there is an entity that is composed of the process and product: the achievement itself. This whole is the location of the value accrued from Recursion. While the previous construals of Recursion located the recursive value *in* the process part, this construal does not do this. The process, on this construal, *retains* its independent value: the *actual* value of the process is its independent value. The recursive value is located in the whole, not in the process. The parts of the whole thus retain their independent intrinsic value. As a result, the value of the whole, process and product, may *not be equal* to the sum of value of the process and the value of the product.

In value theory, of course, we have a name for such an occurrence: *organic unity*. Organic unity occurs when the value of the whole differs from the sum of the values of the parts.[14] On this approach, then, Recursion is an instance of true organic unity.

To elaborate, an achievement is a whole with a process and product, in which the process, according to the Process Thesis, is always a source of positive value. On the organic unity construal, Recursion entails that an achievement in which the product is of positive value will have positive value as a whole, and an achievement in which the product has negative value will have negative value as a whole.

Although this is the result we were hoping for in the case of achievements that have significantly evil products, it is not the result we wanted for petty evil products: the intuition here is that petty evil achievements such as practical jokes have some positive value. Indeed, the organic unity construal entails that, *as a whole*, the value of an achievement with a negative product will be negative.

[14] This rough gloss should not be taken as an authoritative definition. There is debate about precisely what characterizes organic unities. Moore appears to characterize organic wholes such that the value of the whole “bears no regular proportion to the sum of the values of its parts” (*Principia Ethica*, 27), but Fred Feldman shows that this is problematic (*Utilitarianism, Hedonism and Desert*, Cambridge: Cambridge University Press, 1997, pp. 112–24); Chisholm offers a different analysis (*Brentano and Intrinsic Value*, p. 75), with which Lemos takes issue (*Intrinsic Value: Concept and Warrant*, Cambridge: Cambridge University Press, 1994, pp. 196–200; and also “Organic Unities” *The Journal of Ethics* 2 (1998): 323–4).

But there is more to be said. We can avail ourselves of Moore's distinction between value *as a whole* and value *on the whole*.[15] The value *on the whole* is the sum total of the value of the whole as a whole, plus the values of its parts. Tallying the value of all the entities—process, product, whole—gives us a sum total of the value of the whole *on the whole*. The parts, after all, in an organic unity, retain the value that they would have independently of the whole. Thus we can add up the value of these parts, and then add this sum to the value of the whole itself as a whole. This gives what we might call the net value of all the entities—that is, as it is called by Moore, the value of the whole *on the whole*.

In a petty evil case, then, we have a product with a small negative value, a process of high positive value, and a whole, of small negative value. Added together, two small negatives and one large positive, we thus have a *positive* value on the whole. The positive value of the process is positive enough to outweigh the negative value of the product and the whole. Petty evil achievements thus have a positive value. This is indeed the result that we were hoping for.

In a significantly evil achievement, we would be hoping for the view to tell us that the achievement is of *negative* value on the whole. And it does. In a significantly evil case we have an achievement with a product of very large negative value, and a process of positive value, and a whole of negative value. Here, if the negative value of the product and whole are negative *enough*, as they would be in a case of a very evil achievement, they outweigh the positive value of the process. Thus, the value of the very evil achievement is on the whole negative.

On this construal, however, unlike the others, there is still some *positive* value that is retained by the process in the evil achievement. Thus there is still some respect in which the evil achievement is good. I am inclined to say that this is a plausible implication: after all, given the process thesis, the evil achievement shares the very same good-making features that make non-evil achievements valuable. An incredibly ingenious and clever yet diabolical plan to commit the perfect crime is nonetheless ingenious and clever, and, assuming it is an achievement, it shares the other features common to all achievements, including those features that make them valuable. Yet, the organic unities construal of

[15] *Principia Ethica*, pp. 214–6.

Recursion tells us that the value of the very evil achievement on the whole is *negative*. As a result, things overall would be better off *without* the evil achievement. So it is overall bad, even though there is a respect in which it is good. This captures quite nicely what we might be inclined to say about very evil achievements. Thus the organic unities construal succeeds in giving us yet another nice result.

The organic unity construal similarly gives us the implications we want to see in achievements with products of positive value: there is an additional bonus of value—the value of the whole as a whole is positive, thus adding a bonus of positive value to the whole as a whole. Thus, overall the organic unities view captures the implications about achievements that we wanted, and so fares better than the other construals considered so far.

But surely the organic unities construal is subject to a version of Zero Case objection as well. Could not it be possible, after all, that the value on the whole could be zero? In such a case the positive value of the process exactly counterbalances the negative value of the product and the negative value of the whole as a whole. Thus the value of the whole on the whole amounts to zero.

Indeed, this is possible. However, what is problematic in the objection to the summative construal is no longer problematic. The objection to the summative construal is that in Zero Case* there is *nothing* of intrinsic value whatsoever. The elaborate pursuit of an evil end is equal in value to a mere sneeze.

But given the organic unities construal, even if the value of the whole on the whole sums to zero, the parts *retain their original intrinsic value*. On this construal, in every case of achievements the process *retains* its value. And the whole as a whole, in an instance of an evil achievement, retains its negative value as such. So even if the overall state of the value of the whole on the whole adds to zero, there are still components of the achievement that retain their original value, positive and negative. This was the original complaint with the Zero Case on the summative construal: because the summative construal is such that the actual value of the *process* is determined by Recursion, it is susceptible to cases where the process has *no* value (and also where the whole achievement had no value). But the organic unities construal is not susceptible to this problem.

One might raise a further worry, however, that the organic unities construal is unable to account for the disvalue of *unsuccessful attempts*. The organic unity construal of Recursion governs the value of wholes, which are, in achievements, made up of two parts: the process and product. But if no product is produced, Recursion appears to have nothing to say about the value of the whole, which consists of only one part: namely, the process.

Consider a different case with Dr Kill-All:

Kill-All 2. Dr Kill-All wants to kill lots of people by making a terrible toxin. He labors intensely designing his toxin and the way in which he will administer it, but the plan is foiled, and no one gets killed.

Here we have a process, which, according to the Process Thesis, is of some positive value. The product would have negative value, but does not, since it does not exist. The whole, then, appears to consist only of the process—there is no evil product that is a part of this whole. Now, this is indeed an unfortunate implication, since of course a great deal of the motivation for accepting Recursion is to capture the negative value of processes such as this one here. How can the organic unity construal of Recursion account for the value of evil achievements in cases where there is no evil product to shape the value of the whole?

The issue concerns organic wholes more generally. In some organic wholes, the intentional object of a pursuit or, more broadly, attitude, does not *exist* in what we might call a metaphysically robust sense, yet we are inclined to think that the intentional object shapes the value of the whole nonetheless. For example, take a case of cruelty, where A cruelly enjoys B's pain. Proponents of organic unities accounts are inclined to classify such instances of cruel pleasure as an organic whole that is disvaluable as a whole; moreover, cruel pleasure is a disvaluable whole *regardless* of whether or not the intentional object of the pleasure—B's pain—is real, or is imagined. Moore, for example, holds the view that cruelty is intrinsically bad regardless of whether or not the pain that is the object of enjoyment is imaginary pain or real pain: in other words, cruelty is characterized by enjoyment of another's pain, and is disvaluable as such; the enjoyment of the pain is *equally bad* regardless of whether or not it is *actual* pain that is being enjoyed.[16] Suppose that B's pain is indeed merely imagined by A. Here the organic whole consists of two putative parts: the pleasure, and its intentional object, the imagined

[16] Moore, *Principia Ethica*, p. 210.

pain. But the intentional object of the pleasure here is a pain that does not exist. How can the non-existent pain be said to be a *part* of the whole, and thus shape its value as an organic whole? To put the same puzzle in our current context: how can the non-existent product of an evil achievement be said to be a part, and thus shape the value of the achievement as a whole?

This question, many think, can be answered only by appealing to the going definition of *parts*. But it turns out that it is a difficult matter just what it is for something to be a "part" of a whole, and there is contention over how this is to be defined.[17]

However, it is my view that looking toward an account of "part" in order to resolve this question is unnecessary. We do not need to know the correct analysis of "part" in order to resolve this issue. Rather, we have a very strong grasp on the position that intentional objects *can and do* shape the value of wholes even in cases where they do not exist in a robust sense. Cruel pleasure in another's pain, as we just considered, is bad *regardless* of whether the pain is real or imagined. The pain, which is the object of malicious pleasure, shapes the value regardless of whether or not it is *actual* pain being experienced by someone, or if it is just *imagined* the person who is enjoying it.

Whatever it is that we have in mind by "part"—where we mean the components of an organic whole that shape its value—the intentional object (robustly existing or not) counts as one of these. Intentional objects *are* parts in the relevant sense insofar as they clearly shape the value of the whole. Our impression of the negative value of malicious pleasure or of the negative value accrued to the pursuit of an evil end that fails to come to fruition are *evidence* that intentional objects are "parts" in the sense that is relevant for organic unities. That is to say: we have a stronger grasp on the truth of this premise—namely, that intentional objects shape the value of organic wholes, than we do on the truth of any analysis of part-hood. Thus, we should take capturing the former

[17] Much of the debate appeals to the Brentano–Chisholm definition of "proper parts" (Chisholm, *Brentano and Intrinsic Value*, p. 73) but this view entails the intentional object of an attitude is not a part of a whole. Hurka says that this is a peculiar understanding of "part," and prefers to see an attitude-plus-intentional-object as a single thing with *no* parts ("Two Kinds of Organic Unity"), but this means that, for examplee, malicious pleasure is *both* intrinsically good and intrinsically bad, and I agree with Lemos that the same thing cannot be good and bad all over at the same time (Lemos, "Organic Unities," p. 326).

as a desideratum of the latter, and accept that intentional objects are the relevant components of organic wholes.

To be clear, in this essay I am arguing for the relatively narrow conclusion that the organic unities construal is the best understanding of the principle of Recursion: that is to say, instances of loving the good and hating the bad. Whether or not all instances of organic unities are best understood on a similar construal (as opposed to a conditional account, for example, according to which the values of the parts change within the context of the whole) is another matter. But I do take it that my points here about part-hood and its status in organic unities apply to organic unities more broadly.

As a result, there is no concern that the organic unities construal of Recursion will have difficulty accounting for cases of failed attempts at evil achievements. Thus we safely can take the following as how the organic unities construal will account for failed attempts at evil products. Given a difficult attempt of a very evil goal that is not attained, the pursuit, according to the Process Thesis, has positive value. The intentional object, were it to come to fruition, would be of very negative value, but is of zero actual value. The organic unities construal of Recursion holds that the pursuit of an evil end is of negative value as a whole. Since the product has zero actual value, the components that we add together to tally our value *on* the whole will be the positive value of the process and the negative value of the whole as a whole. Assuming the positive value of the process is not greater than the negative value of the whole as a whole, on the whole it is of negative value. Of course, *were* the product to be attained, the negative value of the product would further be added to our calculation of the value of the whole on the whole: thus a *successful* attempt at an evil end would be *worse* than an unsuccessful one. This seems correct. So the organic unities construal captures the desired results.

However, we might now worry that if the pursuit of the evil end is significantly elaborate (and so has to a great degree those features that make achievements valuable), then on the whole a *very* evil achievement could be of positive value. Could the positive value of the process outweigh the very negative value of the product and the whole as a whole?

The organic unities construal, like any construal of Recursion, reports on proportionality: the pursuit of a very evil end is *worse* as a whole than the pursuit of a less evil end, and so forth. In cases where a very evil end is being pursued, Recursion can say that as a whole it will be much worse than in a case where there is a pursuit of a merely minor evil.

Yet it seems possible nonetheless that there could be a very valuable pursuit of a significantly disvaluable end, such that the positive value of the pursuit outweighs the negative value of the product and the whole as a whole, meaning that an incredibly elaborate attempt of a very evil end could be on the whole good.

However, Recursion captures a second kind of proportionality as well: it is worse to pursue an evil *more intently* than it is to pursue the same evil less intently. Similarly, it is better to pursue a good more intently than it is to pursue it less intently. Hurka's full account of Recursion includes such a consideration to capture proportionality here, as well as in the earlier sense, so I will not go into the ensuing details.[18] But the point here is that any construal of Recursion indeed has the resources to accommodate the potential counterexamples we are currently entertaining.

Yet even after adjusting for proportionality, the possibility appears to remain open that on the organic unity construal, an evil achievement could ultimately be of *positive value* on the whole. These would be cases where the pursuit is so very impressive and elaborate that its value in virtue of these features would be so great as to outweigh the negative value of the product and the whole as a whole, even after the value is adjusted in response to proportionality.

But whether or not such a case would indeed be of positive value on the whole is a delicate balance: as the effort of the pursuit increases, so does its positive value, but so does also the *negative* value of the whole as a whole, as a result of the second kind of proportionality. Whether or not the positive value could be greater than the negative will depend on the details of the numbers. But even if it *is* of positive value on the whole, given the organic unities construal, we can reassure ourselves that it is nonetheless of great negative value *as* a whole.[19]

[18] VVV, pp. 58ff.

[19] The degree to which this implication is unpalatable will vary according to intuitions about, say, ingenious art heists. I am inclined to say that indeed evil schemes of that sort could be of positive value on the whole, and so I see this implication as appealing rather than as a bullet to bite. For those who are less inclined to agree, I might point out that we typically think that it is perfectly appropriate to enjoy and admire elaborate art heists in fictional depiction, and so it seems that there is at least that to be said to motivate the intuition for their having some positive value. Could there be a case with a very, very evil product—such as a heinous

Be that as it may, the organic unities construal does very well at capturing our intuitions about the value of evil achievements, and avoids the implausible implications of the alternative construals.

The view is not without peculiarities, however. The organic unities construal is considerably more complex than the alternative construals of Recursion. Adding all these extra features—whole as a whole, whole on the whole, and so forth—may seem akin to Ptolemaic epicycles, bordering on Byzantine, and should be avoided if there is a more theoretically streamlined approach available. But the more theoretically streamlined approaches that are available—the polarizing and summative views—have been *rejected* for their counterintuitive implications. I have argued that the organic unities view gives better results for the cases that we have considered. I do not really find the epicycle objection a compelling one[20] —if the world is complex, our theories should reflect that.

murder? Again, this might be possible, but the negative value of the whole as a whole would presumably be so great as to outweigh the positive value of the process in most cases. If indeed the balance were just so, and even the pursuit of a heinous murder could be of positive value on the whole, then so be it. Once again we can take refuge in the fact that there is nonetheless a great deal of negative value—the whole as a whole. Moreover, if the pursuit were indeed so unimaginably clever and ingenious, then perhaps we would not find it so counterintuitive to say that it could be of positive value. If we are still inclined to think that the real thing ought to be deemed of negative value on the whole, bear in mind as well that our intuitions may be pulled in that direction, since to be sure one could make a case that evil schemes are of negative value *instrumentally*—on some account or other of instrumental value—in that *typically* such a scheme would cause something of negative value.

[20] Motion is, after all, relative.

5

Self-Ownership and the Conflation Problem[1]

DAVID SOBEL

This essay will explore problems and potential solutions for a moral theory which claims that our most basic and powerful deontological rights stem from our self-ownership. Call this the Self-Ownership Thesis. Such views have attracted those yearning for an explanation and vindication of the thought that we enjoy powerful protections from interference when we are minding our own business even if more social good would result if we were interfered with.[2] After all, you may not take my kidney without my consent merely because it could do more good elsewhere. Self-ownership is attractive because

[1] I am very grateful for helpful discussion with Jason Brennan, Aaron Bronfman, Eric Mack, Joe Mendola, Justin Moss, Michael Otsuka, Dan Russell, David Schmidtz, Drew Schroeder, John Thrasher, Mark van Roojen, the students in my graduate seminar at Nebraska on these issues, commentators on two Pea-Soup blog posts, the audience at the 2012 Arizona Workshop on Normative Ethics and at the 2012 Noise conference, an anonymous referee for this volume, and, especially, Janice Dowell, Connie Rosati, Peter Vallentyne, and Steve Wall. This essay is a companion paper with my "Backing Away from Libertarian Self-Ownership," *Ethics* 123 (October 2012), 32–60. Each considers different replies to the same problematic feature of self-ownership views. Thus there is overlap between them, and some paragraphs in common. While the arguments presented in each of them are meant to be independent of each other, I believe they supplement each other.

[2] See, among others, John Locke, *Second Treatise of Government* (Hackett Publishing, 1980), p. 19. Locke writes, "every man has a property in his own person: this no body has any right to but himself." See also Nozick, *Anarchy, State, and Utopia* (Basic Books, 1974), pp. 172, 281–3, and 286; Murray Rothbard, *The Ethics of Liberty* (New York University Press, 2002), p. 113, goes further and claims that the only human rights are property rights; John Hospers, "The Libertarian Manifesto," in *Morality in Practice*, ed. James Sterba (Wadsworth, 1997), p. 21, writes, in his defense of libertarianism, that "libertarianism . . . is the doctrine that every person is the owner of his own life, and that no one is the owner of anyone else's life. . ."; Eric Mack, "Self-Ownership, Marxism, and Egalitarianism," parts 1 and 2, *Politics, Philosophy, and Economics*, vol. 1, no. 1 (February 2002), and vol. 1, no. 2 (June 2002), explicitly embraces the Self-Ownership Thesis in these and other works. Left-Libertarian writings that champion the Self-Ownership Thesis include, among others, Vallentyne, Steiner, and Otsuka, "Why Left-Libertarianism Isn't Incoherent, Indeterminate, or Irrelevant: A Reply to Fried," *Philosophy and*

it appears to offer a satisfyingly direct and not very hostage to empirical fortune justification for such protections. That something is mine—that I own it—provides an obvious and much relied upon rationale for my authority over what may happen to a thing even when others can create more good with it. Further, it is deeply plausible that one has a non-conventional claim to decide what may be done with one's body and to not having it messed with without one's consent. Self-ownership, far from a cobbled-together rationalization for protecting the privileges of the privileged, is an intuitive and tempting foundation for a non-consequentialist morality.[3] Small wonder, then, that leftist non-consequentialist egalitarians are now busily exploring the prospects of vindicating their view from within a self-ownership framework.

And of course, self-ownership does not only protect our kidneys. It is also thought to grant us broad protections from coercion when we are engaged in self-regarding actions with other competent self-owners. Let us call the attractive set of rights that libertarian self-ownership views are thought to vindicate, the Millian liberties. Consequentialism, it is often thought, makes the protection of our Millian liberties too hostage to empirical fortune, for we can easily imagine cases in which interfering with a person's Millian liberties creates more value. Self-ownership views, it is thought, provide a more secure, and therefore more attractive, justification for such liberties.

The Self-Ownership Thesis is traditionally taken to maintain that only rarely, at best, may we infringe upon the property rights of the one for the sake of the good of the many.[4] Thus it is, at best, rare that

Public Affairs 33 (2005): 201–15; Michael Otsuka, *Libertarianism Without Inequality* (Oxford University Press, 2003); Peter Vallentyne, "Left-Libertarianism and Liberty," in *Debates in Political Philosophy*, ed. Christiano and Christman (Blackman Publishers, 2009), pp. 137–51; Hillel Steiner, "Original Rights and Just Redistribution," in Vallentyne and Steiner (eds.) *Left-Libertarianism and Its Critics*, Palgrave, 2000. See also J. J. Thomson, *The Realm of Rights* (Harvard University Press, 2000). She concludes: "people own their own bodies." See also Michael Huemer, "America's Unjust Drug War," in James and Stuart Rachels (ed.) *The Right Thing to Do* (McGraw Hill, 2010).

[3] G. A. Cohen highlighted such advantages of self-ownership views in his "Self-Ownership, World-Ownership, and Equality," in his *Self-Ownership, World-Ownership, and Equality* (Cambridge University Press, 1995).

[4] I here speak of "property rights" rather than "property rights in oneself" because the former is less cumbersome and more general. I mean to be talking about property rights that are purported to have the moral force that the Self-Ownership Thesis attributes to our rights over our own bodies. It is, of course, an interesting question (and one that divides left and right-libertarians) how widely beyond the self we have property rights so conceived. Broadly, left-libertarians maintain it is at best rare to have such powerful property rights in things other than oneself while right-libertarians maintain that it is common.

we may tax the rich merely to benefit the poor or take a person's spare kidney or blood just because it is badly needed elsewhere. More generally, there are classes of actions that infringe upon a person's property rights, and we enjoy powerful, if not absolute, protections from all such actions. On this conception, my property rights, whether they protect something important to me or not, provide powerful protections. This is what allows the above simple and powerful argument against a range of activity that would infringe upon what a person owns without requiring an investigation into the significance of the infringement.

I will argue that the most plausible understanding of the rights of self-ownership is significantly different from the traditional understanding. I will argue that maintaining that we are entitled to powerful protections against even trivial infringements on our property has consequences no one is willing to accept. Rather, the most plausible understanding of the rights of property owners must allow that small infringements for significant gains are quite generally permissible.[5] Two implications of what I claim is the genuine upshot of the Self-Ownership Thesis are particularly at odds with the traditional understanding. First, I claim that, best understood, our rights of self-ownership do not provide powerful protections against all redistributive takings of our property. The strength of the claim that we have against redistributive efforts that involve infringing property rights for the sake of promoting the social good is significantly more variable, and thus sometimes much weaker, than the tradition supposes. Thus it will not infrequently be permissible to take some property from a person who will little miss it and give it to those who truly need it. Secondly, I claim the Self-Ownership Thesis's vindication of our Millian liberties remains importantly hostage to empirical fortune. If I am right, proponents and opponents of self-ownership libertarianism have significantly misunderstood what the moral upshot would be of our being self-owners. Further, while the upshot I claim flows from the Self-Ownership Thesis is more congenial to my consequentialism-polluted intuitions than the traditional view,

[5] There is a different but related question of how regularly it in fact is the case that infringements would make the world significantly better. Some libertarians are doubtful that things like redistributive taxation would in fact make for more social good. If they were right about that, the practical upshot of the view I propose here would be less different from the traditional view than I go on to suggest. Nonetheless the significant change in why such redistribution is impermissible would remain.

I avoid reliance on such intuitions in shaping the rights of self-ownership. I aim to argue for my non-traditional conception of the rights that flow from self-ownership using only intuitions that libertarians and non-libertarians share.

The plan for the essay is to start with a problem for the Self-Ownership Thesis. The problem is that it is implausible that we enjoy the same degree of protection against all actions that infringe upon our self-ownership. If we enjoyed such powerful protection against trivial infringements too much would be impermissible. Your trivial pollution, for example, that eventually falls to the earth and causes some small risk of minor skin irritation, would seem to infringe upon my property rights over my skin.[6] If so, the Self-Ownership Thesis threatens to make all fires impermissible unless they are unanimously consented to by everyone that might be affected by them. Nozick saw this worry and significantly adjusted his view to try to solve this problem. Peter Railton emphasized this problem in arguing that such views are fundamentally problematic.[7] After explicating this worry further, I offer what I take to be the most natural and plausible fix; namely, that we allow that different property infringements are differentially important and we are owed different levels of protection against them. We will also consider at length and reject a few attempts to solve the problem without this move.

Then I turn to the issue of how we might understand what makes one property right weightier than another. Broadly, one might use an objective or subjective measure of the significance of different infringements. I argue that the objective measure, divorced as it is from the agent's own assessment of the significance of the infringement, abandons the self-ownership framework for our rights. Alternatively, one might tie the strength of the right to the strength of the agent's contingent concern about the infringement. The latter, I claim, fails to vindicate the thought that there are classes of actions, such as self-regarding actions or acts that infringe upon our freedom of religion, that we all enjoy powerful protections from. Rather, the upshot of such a view would be

[6] See Samantha Brennan, "Thresholds for Rights," *Southern Journal of Philosophy* 33 (1995), 143–168, and Michael Smith and Frank Jackson, "Absolutist Moral Theories and Uncertainty," *The Journal of Philosophy* 103 (2006), pp. 267–83.

[7] See Nozick, *Anarchy, State, and Utopia*, chapter 4, and Railton, "Locke, Stock, and Peril: Natural Property Rights, Pollution, and Risk," in his *Facts, Values, and Norms* (Cambridge University Press, 2003).

that we are only owed powerful protections against infringements that we quite mind. On such a view, some will enjoy only quite weak protections against some property infringements. In sum, distinguishing the size of the protections we are owed against various infringements, as seems the most plausible way to develop the framework, forces self-ownership views into unfamiliar waters in which our Millian liberties are once again hostage to empirical fortune. While it seems awkward to count it as a cost of the emerging view that it would generally allow as permissible redistributive taxation of money (or hair) from those who little feel the loss to those who are seriously in need, this would be a most significant change in the view.

I should also emphasize that I am claiming that a self-ownership view that does not allow that differentially significant infringements merit different levels of protection has grave difficulties, and I have not found a way to make them plausible. If such views are not plausible, then self-ownership views are forced to take what may well feel to be a half-step towards consequentialism. The moral disvalue of infringements would be more broadly fungible for social good than such views have traditionally supposed. However, it is important to see that the resulting view would remain deontological and rights-based rather than consequentialist. The resulting view could say that you may not impose infringement harm of size N on me just to avoid 5N of harm befalling others. But the view does abandon the odd but traditional thought that the size of the infringement harm to me of your action has no impact on the amount of social good needed to make such an infringement permissible.[8]

5.1 THE FAILURE TO DISTINGUISH THE SERIOUSNESS OF PROPERTY INFRINGEMENTS

Some actions are morally worse than others. It is not merely that the morally permissible acts are better than the impermissible. Some bad types of action are worse than others. I have a claim that you not take my tennis racket without asking but it is worse to so take my kidney. And in a range of cases, how morally bad an action is affects how much

[8] "Social Good" stands in here for a variety of possible views and need not be simply aggregate welfare.

social good can make such actions permissible. If the only way to save Joe's life is to use my racket without my permission, then surely you may do so. But if Joe will die unless you take my kidney it still seems you may not take it. If Joe will be run over by a bus unless you push his body out of the way without being able to get his consent beforehand, you may do so. But you may not do so just for a laugh.[9]

Some imaginable deontological theories would have difficulty explaining this. The class of deontological theories that seem to be embarrassed by this difficulty I will call "Broad, All or Nothing" deontological theories. Such views aim to illuminate a broad swath of the moral terrain with a single principle. Further, this principle most naturally suggests that an action either fully has the morally problematic feature or fully fails to have it. Intuitively the problem is that such views conflate cases on the trivial end of the spectrum and on the serious end and treat them as if they were equally morally important. Call this the Conflation Problem. The case that springs to mind here is Kant—especially his Universal Law formulations of the categorical imperative. Contradictions, whether in the will or in conception, seem to be all or nothing rather than coming in degrees. Further, Kantians interpret the Universal Law formulations such that it is meant to settle a very wide swath of morality.[10] Thus I suspect that at least Kant's Universal Law formulations will prove vulnerable to the kind of concerns urged here.[11]

[9] See Steven Wall's excellent discussion of problems that self-ownership views have handling cases of soft-paternalism, "Self-Ownership and Paternalism," *Journal of Political Philosophy* 17 (2009), 399–417. Presumably, self-ownership views must rely significantly on social conventions and tacit consent to make permissible such things as slapping a good friend on the back after she gets tenure without getting consent beforehand. Recall that Nozick strongly rejects tacit consent. He writes: "tacit consent is not worth the paper it is not written on." (287).

[10] Tom Hill, in "Making Exceptions Without Abandoning the Principle," in his *Dignity and Practical Reasoning* (Cornell University Press, 1992), considers the sort of challenge to such Kantian principles I have in mind. Hill writes that there is a "tendency to append a 'catastrophe clause' to familiar principles whenever the consequences of adhering to the principles are so repugnant that it seems morally perverse to refuse the exception... [But] if the balance of consequences determines what to do in the extreme cases, why not in the case slightly less extreme, and so on?" (199–200). Hill's suggestion for attempting to solve this problem "is that the dignity principle applies first to decisions about the basic system of public laws and only then to individual decisions remaining underdetermined by those laws" (208). Essentially, the suggestion is that we take Kant to be talking, like Rawls, primarily about the basic structure of a just society rather than individual morality.

[11] It may be that whether or not one can universalize a maxim in part depends on the amount of social good at stake. If so, Kantian views have an obvious method of avoiding the Conflation Problem. If not, then they seem vulnerable to the problem. Nozick's use of Kant's idea that we must not use people as a means seems directly vulnerable to the problem.

This essay will be focused only on how this worry plays out for the Self-Ownership Thesis. But the conclusions I draw here apply at least against any view that claims we have uniformly strong property rights forceful enough to vindicate the stringent traditional conclusions against nearly all redistribution and paternalism, whether such a scheme is justified via self-ownership or not.

Some terminological issues must briefly detain us. I will refer to unconsented crossings upon another person's property, with Nozick, as a "boundary" or "border" crossing. One would have naturally expected that a view that takes property rights very seriously would maintain that at least all harmful border crossings without consent are rights infringements. However, according to some views, as we will see, a boundary crossing would only become a rights infringement if, for example, adequate compensation is not paid or if the harm from the crossing exceeds a certain threshold.[12] An infringement involves doing something that someone's rights protect her against. Nonetheless, a rights infringement could potentially be permissible, at least according to non-absolutist variants of the view, if it would, for example, avoid a catastrophe.[13] In such cases let us say the infringement is justified. Infringements that are not so justified are impermissible and violate a person's rights.[14]

Traditional self-ownership views have tended to have the two features that generate the Conflation Problem. Such views aspire to illuminate a wide swath of the moral terrain, perhaps all of enforceable morality. And such views have tended to suggest that an action either fully infringes a property right to one's own body or it fully fails to do so. Intuitively, a theory that has the Conflation Problem will either treat serious matters too lightly or treat trivial matters too seriously. In the self-ownership tradition, infringements have always been taken to be a very big deal. Thus the way the Conflation Problem will manifest itself in such a view will be by treating relatively trivial infringements as if they were more significant than they are. The friend of the

[12] I think we should say that on a non-absolutist version of the view, enough social good can make it permissible to infringe a right, rather than say that in such cases the right disappears. Nozick says: "Overridden rights do not disappear; they leave a trace," p. 180.

[13] Nozick, p. 30, expresses agnosticism between an absolutist version of the view and one that permits rights infringements to avoid a "catastrophic moral horror."

[14] I follow Vallentyne's helpful terminological conventions here. See his "Enforcement Rights against Non-Culpable, Non-Just Intrusions," *Ratio* 24 (2011), 422–42.

Self-Ownership Thesis might try to persuade us that it is appropriate to treat such radically different infringements with equal care. This would be to argue that while the view does conflate seriously different sorts of infringements, it is not a problem that it does so. I think when we see the upshot of this strategy we will find this implausible. Alternatively, the view could be modified to avoid the most dramatic sorts of Conflation Problem.

The four most obvious ways by which the traditional view could be modified to avoid the Conflation Problem would be either to 1) make it harder for an action to count as an infringement by maintaining that a border crossing only counts as a rights infringement if the harm of the border crossing exceeds a certain threshold. As this view rejects the idea that there is no lower limit to the border-crossing harm that results in an infringement, call this to reject No Lower Limit; 2) maintain that differentially important infringements can be made permissible by different amounts of social good (call this rejecting All Infringements are Equal); or 3) claim that our owning something does not give us a claim against just any harmful boundary crossing even if it is above a threshold, but rather only against certain types of such crossings such as those in which the border-crossing harm is intended, foreseen, or in which the owner is used as a means (call this rejecting Property Rights Protect against Border-Crossing Harm); or 4) reject the view that one's property boundaries create powerful side-constraints and instead maintain that one may freely cross the boundary of a person's property without her consent so long as you adequately compensate her for this. Elsewhere I try to show that this view, which I dub "cross and compensate," is Nozick's view. Much of the rest of this essay will explore the first three of these options and argue that only the second option holds real promise of adequately responding to the Conflation Problem. I explore the virtues and vices of cross and compensate elsewhere, and conclude that it is inadequate to our problem and has independent issues.[15] A fifth option of seriously downgrading the significance of all infringements would obviously have the Conflation Problem in the opposite way—that is, it would treat as too trivial very serious matters.

[15] "Backing Away from Self-Ownership."

5.2 NOZICK

Nozick briefly but illuminatingly considered the view that border crossings must involve at least a certain threshold of harm to count as an infringement, and rejected such views. In doing so he highlighted important issues concerning risk. He wrote:

> Actions that risk crossing another's boundary pose serious problems for natural-rights positions . . . Imposing how slight a probability of a harm that violates someone's rights also violates his rights. Instead of one cutoff probability for all harms, perhaps the cutoff probability is lower the more severe the harm. Here one might have the picture of a specified value, the same for all acts, to mark the boundary of rights violations; an action violates someone's rights if it is expected harm to him . . . is greater than the specified value. [74]

In other words, to count as an infringement an action that risks crossing a border would have to produce above a threshold level of expected harm. But against that tempting modification, Nozick continued:

> This construal of the problem cannot be utilized by a tradition which holds that stealing a penny or a pin or anything from someone violates his rights. That tradition does not select a threshold measure of harm as a lower limit, in the case of harms certain to occur. It is difficult to imagine a principled way in which the natural-rights tradition can draw the line to fix which probabilities impose unacceptably great risks upon others. [75]

Nozick's reasoning here seems to me persuasive, if in need of some fleshing out. Nozick is charitably interpreted as thinking not only that historically self-ownership views have held that there is not a lower limit of border-crossing harm below which the action does not count as an infringement (No Lower Limit), but that the tradition has good reasons for that commitment. What might those reasons be? Nozick does not speak much of this question. Vallentyne, Steiner, and Otsuka (VSO) offer several possible responses to the Conflation Problem, including the possibility of rejecting No Lower Limit. We will shortly consider difficulties in rejecting No Lower Limit in the context of assessing VSO's view.

For now I will just say that Nozick's argument seems sound that if there is no lower limit to the amount of boundary crossing harm that constitutes an infringement in cases of harms certain to occur, then there is no principled ground for maintaining that there is such a limit in cases of risk or expected harm. In other words, if we accept No Lower Limit we must accept No Lower Expected Limit.[16]

When we turn our attention away from the question of whether or not an act constituted an infringement, and to the question of how much social good it would take to make such an infringement permissible, new options open up. We want to say that I have some claim against your stealing my penny or my pin, but we also want to say that I enjoy less protection against your borrowing my tennis racket than stealing my kidney. And we could say all this by saying that as the significance of the infringement diminishes, the size of the social good needed to make such actions permissible becomes less. So, we could seemingly vindicate our commonsense intuitions here if we said that if you need to borrow my racket without my permission to ward off an attacker, doing so does infringe upon my property rights, but if things will be very bad if you do not so infringe, it may well nonetheless be permissible for you to do so.

It is not clear whether Nozick considered the possibility of selling different infringements at different prices.[17] In any case, recall that Nozick expresses agnosticism between an absolutist view and one that allows that an infringement is permissible to avoid a "catastrophic moral horror."[18] Both of these views suggest that the size of social good needed to make permissible any infringement, regardless of its significance, remains constant; that is, "All Infringements are Equal." The view need not be that all infringements are in all ways normatively alike, but rather

[16] Two quite recent and highly recommended works that deal with the often neglected chapter 4 of *Anarchy, State, and Utopia* are Eric Mack's "Nozickian Arguments for the More-Than-Minimal State" and Barbara Fried's ""Does Nozick Have a Theory of Property Rights," both in *The Cambridge Companion to Nozick's Anarchy, State, and Utopia*, ed. Ralf Bader and John Meadowcroft (Cambridge, 2012). Kasper Lippert-Rasmussen, "Against Self-Ownership," *Philosophy and Public Affairs* 36 (2008), no 1, is also highly recommended.

[17] Nozick's considered view, I believe, allows that different border-crossings take different amounts of compensation to make such crossings not count as infringements. This is different from different infringements getting different weight.

[18] Nozick, p. 30. It is commonly maintained that deontological theories, even those that have no natural home for considerations of the overall good, can nonetheless reasonably maintain that the proposed side-constraint is overridden when a great amount of good would be lost if we heeded the constraint. That is, it is claimed that such deontological theories are not rendered problematically *ad ho*c by helping themselves to such an addition no matter how inorganic to the rationale on offer for the side-constraints. But consider the apparent gruesomeness of a consequentialist view that says that we should maximize welfare, unless doing so would create a rights-violation catastrophe. After noting some suspicion that consequentialist

only that in terms of the amount of social good needed to make such an infringement permissible, they are equal.

Otsuka makes clear that the strict rights of self-ownership he accepts cover one's hair, and so he is committed to strict rights even against trivial infringements. Thus one might be encouraged to think that he is committed to All Infringements are Equal. In passing, however, Otsuka seems favourably disposed to a luxury tax on that which is self-owned but not necessary for a decent life if such a tax could prevent people from freezing to death.[19] Such thoughts do not receive further discussion, but the view just described may look as though it requires the thought that different infringements are being sold at different prices (infringements not very harmful to the owner being sold more cheaply than infringements that are—hence the tax only on what is a luxury). Still, it seems fair to me to say that Otsuka affirms a very stringent, if not absolute, right over one's hair even if the harm to one of it being taken is trivial. Perhaps we may say that while Otsuka is not clearly committed to All Infringements are Equal, he is clearly committed to All Infringements are a Very Big Deal.

Vallentyne calls himself, in conversation, a reluctant absolutist. I take it to be clear that all absolutists must accept All Infringements are Equal. Vallentyne claims that even in cases in which slightly injuring a person would save millions of lives, the self-ownership of the one who would be injured makes it unjust to impose such an injury.[20]

As I discuss elsewhere, Nozick tries to respond to cases where even trivial pollution might be thought to infringe on so many rights as to be impermissible not by distinguishing between more and less harmful infringements but by appeal to what I call "cross and compensate." Cross and compensate permits us to harmfully cross other people's boundaries provided we provide enough compensation to make the person whose

views are being held to a higher standard here, let us share the common assumption that there is nothing problematically *ad hoc* about such deontological views with escape clauses for welfare catastrophes.

[19] Otsuka, pp. 17 and 19.

[20] Peter Vallentyne, "Left-Libertarianism and Liberty," p. 7. Confusingly, Vallentyne writes: "It may simply be that it is reasonable to behave unjustly in such extreme circumstances." One would wish for some unpacking of the notion of "reasonable" in the above sentence. Vallentyne does not consider the case of risk, but I am supposing that No Lower Limit implies No Lower Expected Limit, and so any sized risk of imposing any infringement harm is unjust on this view.

boundary is crossed at least as well off by her own lights as she would have been had the crossing not taken place. On this view, a border crossing would become an infringement only if such compensation were not provided. Thus our pollution case could end up infringing on no rights. But if the compensation were not provided, then even trivial pollution would still amount to a full infringement and be permissible only to avoid a catastrophe. Nozick, characteristically for champions of views in the neighborhood of self-ownership, tries to find ways to fix problems that intuitively have to do with differentially significant infringements by fiddling with what counts as the one-size fits all infringement.[21]

All Infringements are Equal, by my lights unattractive on its own, becomes intolerable when combined with No Lower Limit. A view with both of these features seems to force us to conclude that flying an aircraft over a person's head with a one-in-a-trillion chance of the aircraft breaking down and crashing into that person must be counted as a rights infringement. And, since on this view rights infringements are normatively a big deal, we seem to reach the conclusion that flying aircraft, emitting pollution, and so on, are going to be generally impermissible, as such actions will involve infringing on many people's rights. If any unconsented to use of or risk to a person's property is an infringement even if the use is nearly harmless, and if our rights against any infringement are very strong, the predictable consequence is that, for example, even innocuous pollution that results from badly needed activity such as building a fire is a violation unless it is universally consented to. This combination threatens to make impermissible a range of activity needed to enjoy an acceptable level of liberty. Presumably, even throwing a stick for my dog would be impermissible given the very small chance that it may hit an unseen person and so infringe her rights.[22] Accepting both No Lower Limit and All Infringements are Equal yields an unacceptably severe version of the Conflation Problem.

We might try saying that risks to a person's property that do not eventuate in harms are not infringements. But this will not help against trivial actual border-crossings, such as some possible pollution cases, and it will

[21] See my "Backing Away from Libertarian Self-Ownership."

[22] Nozick notices that not being able to impose any risk of a property-rights infringement on another would make a criminal justice system impermissible. He writes: "For any system we can devise which sometimes does actually punish someone will involve some appreciable risk of punishing an innocent person, and it almost certainly will do so…" [96].

make surprisingly problematic our claim that others not play Russian roulette with us against our will.[23]

I will treat views that make impermissible any sized risk of even trivial infringements as implausible. Certainly such views no longer seem attractive from the point of view of liberty. Although such issues do not receive enough direct attention, I believe that traditional versions of the self-ownership view tends to combine No Lower Limit and All Infringements are Equal and then struggle with the consequences I am highlighting here. This combination is unpromising. If the friend of self-ownership had to give up one of the two, All Infringements are Equal seems the more tempting option to lose as it is less intuitively plausible, and, as I will shortly argue, rejecting No Lower Limit is problematic. We will therefore shortly be focusing on how the view might go without this component.

But there may still appear to be another way out of this jam without abandoning All Infringements are Equal. One might try saying that we only infringe a right of someone when we do so intentionally or when we use them as a means in the pursuit of our plans. Michael Otsuka takes such a path, suggesting that our rights of self-ownership protect us only from being intentionally used as a means or from taking from us income earned with using only our own bodies.[24] But as I have argued at more length elsewhere, this would leave our bodies unprotected from other people blowing up an area near our property to clear some land, foreseeing that this would kill us.[25] Doing so is not intentionally using us as a means in the relevant sense.[26] Such a move would, most charitably interpreted, radically sacrifice the breadth aspiration of self-ownership views, leaving such views significantly incomplete even concerning enforceable

[23] Railton nicely stresses this point. Nozick discusses such issues without taking a stand on how to resolve them around pp. 75–6.

[24] Otsuka, p. 15. His conception of the rights of self-ownership centrally include "A very stringent right of control over and use of one's mind and body that bars others from intentionally using one as a means..."

[25] I consider and find insufficient such moves in more detail in my "Backing Away from Self-Ownership."

[26] In "Backing Away" I distinguish two different senses of using someone as a means. There is the sense at play in Kant's categorical imperative that is relatively broad. And there is a narrower notion of literally making use of a person in the aid of one's plan, such as pushing the fat man in front of the trolley. Otsuka clearly has in mind this latter, narrower understanding of what it is to use someone in mind.

morality, and require significant supplementation even concerning what claims we have that others not mess with our bodies. Further, I take it to be clear that owning something gives one claims well beyond others not intentionally using it as a means without one's consent. Thus I think we can safely set aside such views.

I believe there are several reasons why it is no accident that traditional forms of the Self-Ownership Thesis have been developed in ways that generate such severe Conflation Problems.[27] One of these reasons is the strength of the Cohen-inspired understanding of the rights of ownership which immediately generates very severe Conflation Problems. I discuss this below. But here I will speculate that part of the problem stems from its advocates focusing on offering a corrective to the perceived unacceptable willingness of consequentialism to violate the claims of the one for the sake of the group. Nozick generalized from cases in which the one is seriously sacrificed for others towards principles that make it a Very Big Deal if the one suffers even incredibly minor infringements for the sake of the group. Recall that Nozick only contemplated tolerating any infringement if it was needed to avoid a catastrophe. When we remind ourselves how minor infringements can be, granted No Lower Limit, and we keep in mind Nozick's point about what this requires that we say of actions that impose very tiny risks, we see that treating every bit of infringement-harm as a Very Big Deal is implausible. Nozick wrote that the sort of side-constraints he championed reflect the fact that no moral balancing act can take place among us; there is no moral outweighing of one of our lives by others so as to lead to a greater overall *social* good. There is no justified sacrifice of some of us for others. This root idea—namely, that there are different individuals with separate lives and

[27] Obviously, even if our property rights are much less powerful than the Self-Ownership Thesis maintains they are, we may well deserve to be thought of as the owner of something or my own body. I am keeping in place, as I believe G. A. Cohen did in his influential understanding of the rights of self-ownership, the idea that a key desideratum of such a conception of rights is that they reasonably be thought capable of serving as a premise in establishing our libertarian's very strong conclusions against nearly all paternalism and redistribution. Some influential contemporary libertarians, such as David Schmidtz, appear to treat the stringency of these conclusions as much more negotiable. See his "Property and Justice," *Social Philosophy and Policy* 27 (2010), 79–100.

so no one may be sacrificed for others—underlies the existence of moral side-constraints.[28]

This statement, which develops Nozick's Kantian rationale for side-constraints, taken literally, rules out the most trivial infringement even if it would produce amazingly important benefits. It can seem that it is Nozick's Kantian rationale that is the problem we have found here. The commitment to not sacrifice the one, no matter how trivially, for the sake of the many, immediately threatens us with Conflation Problems. Our example of pollution shows just how difficult it is to accept literally the view Nozick expresses above. Such thoughts amount to an over-reaction to the perceived excessive willingness of consequentialism to sacrifice the one for the sake of the group. We are finding that not only do the cases of pollution and risk show difficulties with absolutist variants of the view but also that it puts great pressure on the idea that All Infringements are Equal and on All Infringements are a Very Big Deal.

5.3 VALLENTYNE, STEINER, AND OTSUKA: TOWARDS THE OBVIOUS SOLUTION

By now it may well seem that part of the problem is stemming from indeterminacy in exactly what rights I have in virtue of being a self-owner. G. A. Cohen responded to the persistent worry that self-ownership is problematically indeterminate by offering a principled way of sharpening the proposal. He maintained that "the stipulation that self-ownership confers the fullest right a person (logically) can have over herself provided that each other person also has just such a right generates a procedure for determining the content of self-ownership."[29] Vallentyne, Steiner, and Otsuka (VSO) follow Cohen in insisting that their proposal is made more determinate by appealing to the notion of full self-ownership. Full self-ownership "is simply (roughly) the logically strongest set of ownership rights over a thing that a person can have compatibly with others having such rights over everything else." They claim such an understanding of "full-ownership

[28] Nozick, p. 33. Recall that Nozick did not insist on an absolutist version of the view. Given this, apparently he should not be seen to insist that we take this passage literally.

[29] Cohen, "Self-Ownership: Delineating the Concept," p. 213. See also Hospers, p. 22, who writes: "Each human being has a right to live his life as he chooses, compatibly with the equal right of all other human beings to live their lives as they choose."

has a relatively determinate content."[30] They then note that this notion of full self-ownership—what they eventually call strict self-ownership—"has some rather radical implications." These include the claim that my self-ownership is violated "if, in the process of putting out a dangerous fire, you inadvertently send a small bit of stone one hundred yards away, where it lightly flicks my hand. Most people with strong libertarian inclinations will want to reject these implications and thus reject full self-ownership in the strict sense."[31]

In response, VSO tell us that their view is compatible with four possible exceptions, in any combination, from full self-ownership. The exceptions are actions where it is the case that:

1. There is only a very small probability that it will result in an incursion against oneself.
2. If there is an incursion, the harm to oneself will be trivial.
3. The harm was not reasonably foreseeable.
4. The benefits to others of performing the action are enormous (e.g., avoidance of social catastrophe).

The first thing that should strike us about this proposal is that it is *ad hoc*. There is no effort to make a case that the overall proposal has any more unity than a determinate conception of self-ownership together with four independent exceptions to handle counter-intuitive cases.

Further, although the Cohen-inspired understanding of the content of the rights of self-ownership is theoretically attractive and principled, it seems inevitably to force the view into the unacceptable corners I have been discussing. The exceptions VSO allow here are just the sort that both seem most obviously needed to make the view plausible and yet to inevitably and by design fail to provide "the logically strongest set of ownership rights over a thing that a person can have compatibly with others having such rights over everything else."[32] The Cohen-inspired conception, despite its significant virtues, will inevitably lead the view towards exactly the implausible results that VSO's modifications are

[30] Vallentyne, Steiner, and Otsuka, pp. 204–5. See also the papers by the left-libertarians referred to in note 2.

[31] Vallentyne, Steiner, and Otsuka, pp. 206 and 207. Steiner and Vallentyne, when writing alone, do not clearly reject such an implication.

[32] Vallentyne, Steiner, and Otsuak, p. 204.

designed to avoid. There seems to be a significant tension between this principled way to understand what the rights of self-ownership are and the more plausible understandings of our rights. As I will argue below, rejecting All Infringements Are Equal seems the least *ad hoc* way of addressing the problem even though it will continue to conflict with the Cohen-inspired understanding of the rights of self-ownership.

VSO's exceptions do not handle the counter-intuitive cases satisfactorily, I will argue. Obviously I cannot here show this for each of the sixteen variants of VSO's proposal. What I will hope to do here is show that the general shape of each of the exceptions they offer do not individually look to be adequate. Recall that we have already considered the prospects for non-absolutist versions of the view that permits rights infringements to avoid catastrophes (VSO's fourth option), and found that this move on its own failed to distinguish adequately between more and less serious rights violations and so was unable to make permissible intuitively acceptable risks and pollution.

VSO's third option that there is no rights violation if the harm was not reasonably foreseeable seems unhelpful. Saying that a border crossing was not "reasonably foreseeable" might be taken to mean that such a crossing was very unlikely. If we take it this way, it collapses into the first VSO response that I will consider below. But instead we might take it to suggest that the focus should be on the reasonable subjective probabilities available to the actor, not on the objective chance that an action would cross a border. So understood, the distinction is of little help with our problem. The cases that are causing the problem for supposing that our property rights are uniformly very strong are cases of trivial harm or risk. The problem is just as acute in cases where the actor is aware of the small risk.

I speculate that the reason Russian roulette with a great many chambers seems obviously impermissible and flying a normally safe plane seems obviously permissible, despite both activities imposing the same chance of border-crossing harm, has nothing to do with the foreseeability of the harm and everything to do with the significantly different amounts of social good likely to be created by such actions.[33]

Let us now consider the two last modifications VSO offer to full-ownership which essentially amount to rejecting No Lower Limit. Consider

[33] Nozick, in a special context, says just this; see p. 74.

first the idea that if the risk of a boundary crossing is small enough, this amounts to no infringement at all. First, as Nozick showed above, this proposal is unmotivated if we accept No Lower Limit in non-risky cases. Second, there will implausibly be cases where risks just below the threshold are no problem as far as rights are concerned but just over it is a full rights infringement. Such views will be forced to maintain that arbitrarily small additional impositions of risks make a very great moral difference—a much greater difference than a larger amount of risk that took us near to the threshold. Third, this would mean that a state lottery in which there is a small chance that the rich have their assets redistributively taxed, would not violate the rights of the rich. Nor would a similar lottery where my organs would be used for others. Fourth, this would mean that we did not have a right that others not play Russian roulette with our head so long as the chance of killing us is small enough. Fifth, uncoordinated acts each of which is below the threshold could add up to an arbitrarily high chance of an infringement, yet no one infringes upon my rights.[34] Sixth, some acts that bring no one's risk above a threshold could raise a lot of people's risk a little. Presumably the standards should be higher for imposing a small risk on billions of people than it should be for imposing such a risk on one person. Releasing a carcinogen that has a one in a billion chance of killing those who inhale it is very different if it is imposed on only one person than if it is imposed on a billion people. Seventh, nothing has yet been said about the value of the risk-imposing act to the person who imposes the risk. Surely if the act promises only trivial or no value for the actor, or only an infinitesimal prospect of a value, then such actions should be less permissible. For example, if the act promises an N chance of a benefit to the actor and it imposes a 2N chance of an infringement-loss of that same-sized benefit to the person affected by the act, then presumably the act should not be permitted even when 2N is still a quite small risk.

Let us now consider VSO's final amendment: the view that if an act would harm someone only a small amount, below some low threshold, it would not amount to a property infringement. This is just to reject

[34] Such a complaint forces us to distinguish between the view that each person may permissibly impose up to N amount of risk and the view that each person may permissibly have up to N amount of risk imposed upon her. The latter view will maintain that whether an act of mine violates your rights depends on what others have done. Nozick discusses such issues on p. 74 and surrounding.

No Lower Limit. Against this view, note first that the second, fifth, sixth, and seventh concerns above about imposing small risks can obviously be modified to pose problems in the case of small harms. Second, as Nozick pointed out, what would happen to ownership rights over a pen or a pin if relatively trivial harms did not count as rights violations? Third, we might see property rights as fundamental and pre-social natural rights or we might see the institution of property as itself justified by considerations of social value. Allowing that our property rights are not violated in cases in which we are harmed only slightly suggests the latter picture.[35] Such a picture would make perfect sense if the point of the institution of property were to serve as a very socially useful convention.

The good thing about VSO's exceptions from full self-ownership is that they allow us to mark important moral distinctions between actions that impose tiny risks and other kinds of actions that impose greater risks, actions that impose great harms and actions that impose less than great harms, and actions that would infringe rights but produce tremendous benefits and actions that would not produce such great benefits. All of this is helpful, and moves in the right direction. Unfortunately, however, these moves remain uncoordinated and *ad hoc*, and because they are developed only to allow exceptions when harm or risk is very small or benefit is very large, they miss a great many other distinctions that should matter to us. For example, we presumably also want to distinguish between and treat differently the imposition of mid-sized risk and very high risk. It is hard to see a principled rationale for allowing that some such differences matter a lot and others matter not at all.

There is an obvious way to remedy these problems with VSO's proposal. We could let the badness of the rights infringement vary continuously with the size of the risk and the harm. And we could sell different-sized infringements for different amounts of social welfare. What seems plausible, and what VSO's modifications from the

[35] Many have championed a broadly consequentialist rationale for the institution of property. Surely it is implausible on its face that things would go better without stable expectations to enjoy and plan around the availability of certain goods and the incentive structure provided by such stable expectation. But such a rationale for property need provide no reason to think that, for example, progressive taxation rates that largely leave such attractive features of property in place should be thought to violate our legitimate expectations to our property. The arguments presented here are in no way hostile to the institution of property, but only tell against treating property rights that are independent of such considerations as uniformly morally powerful and fundamental.

implications of full self-ownership begin to capture, is that the lower the risk of an infringement an act causes, and the less harm it threatens, the cheaper it should be in terms of social utility to make permissible. The most obvious view in this neighborhood would say that as the welfare costs of an infringement diminish, and as the risk of a boundary crossing diminish, so does the amount of social welfare needed to make permissible such an action. On any reasonable conception of welfare, some infringements of our property rights threaten us with only trivial harms and so should presumably be more cheaply bought by social welfare. Indeed, as we have seen this must be so if just about any pollution or tiny risks of infringement is to be permissible.

On the resulting conception, the fact that something is one's property provides protection in proportion to how important it is for the owner that the thing not be infringed upon. So perhaps the fact that an infringement causes N amount of the relevant sort of infringement-harm requires that the act produce at least 20N of social good to be permissible.[36] This allows the less serious infringements to be bought for less social good, it retains a deontological, rights-based approach, it vindicates the thought that because something is mine I have say over what may be done with it well beyond the extent to which I can create the most good with it, it vindicates the thought that the fact that we own a penny or a pin gives us a claim that others not take such things from us, yet it can explain in a principled way why flying normally safe aircraft overhead is permissible. It vindicates the intuition that you may borrow my tennis racket without my permission if you need to do so to save a life yet you may not take my kidney to save a life. It avoids in a principled way the Conflation Problems.[37] Such a view looks much less *ad hoc* than VSO's proposal, it is more determinate than their proposal, and it solves the concerns just mentioned above about their view. Finally, the emerging view can offer an explanation for why someone might have

[36] Obviously I am just following out the simplest variant of such a picture. The framework need not be welfarist and could include prioritarian thoughts as well. Below, however, I will argue that the self-ownership view fits best with a subjectivist conception of the relevant value.

[37] Richard Arneson, "Self-Ownership and World Ownership: Against Left-Libertarianism," *Social Philosophy and Policy* 27 (2) (Winter 2010), 192, also find that the most promising left-libertarianism must be modified so that "the level of bad consequences that suffices to trigger a moral permission or requirement to infringe Lockean moral rights is variable, depending on the moral importance of the rights at stake in the situation."

been mislead into supposing our property rights are so uniformly powerful. A person might make such a mistake by focusing on and inducting from cases like the kidney example above where our property rights really do offer very powerful protections.

A person might well complain that there will be no single multiplier that will allow us to capture both the intuition that we must not take the kidney and that we may turn the trolley. This seems a legitimate worry. There are several possible replies: 1) we could reject the entire framework; 2) we could live with some of our intuitions being bruised and settle for them being bruised less than on a traditional self-ownership view or a consequentialist view; or 3) we could add to the framework by, for example, claiming that the multiplier is greater for harms intended rather than merely foreseen.[38]

The most obvious way to respond to the general challenge highlighted by the pollution and risk cases would be to distinguish between important property rights and relatively trivial ones and be willing to sell violations of the less important property rights relatively cheaply for social good. That is, the view might provide a theory of value that explains why some property rights are more significant than others by showing that some such rights protect more valuable things and others protect only trivial things.[39]

5.4 PROBLEMS WITH THE OBVIOUS SOLUTION FOR THE TRADITIONAL CONCLUSIONS

The most obvious, compelling, and principled way we have found to respond to the Conflation Problem within a self-ownership framework is to distinguish between important property rights and relatively trivial ones and be willing to sell violations of the less important property rights relatively cheaply for social good. That is, the view would provide a theory that explains why some property rights are more significant

[38] The unadorned view would work something like the view Scheffler outlined in *The Rejection of Consequentialism* (Oxford University Press, 1982). But it would avoid many of the problems that beset that view.

[39] I think of this as pursuing a line that G. A. Cohen suggested but did not pursue. He wrote: "Self-Ownership: Assessing the Thesis," in his *Self-Ownership, Freedom, and Equality*, p. 231, a "limited dose of forced labour is massively different, normatively, from the life-long forced labour that characterizes a slave."

than others by showing that some such rights protect more valuable things and others protect only trivial things. The central problem we have found is not that the rights of self-ownership are conceived to be very strong, although they are, but rather that there is too little discrimination in what sorts of infringements trigger the full protection of such rights.

The problem for the emerging self-ownership view (call it the Value-Sensitive Self-Ownership View) is not that there are no credible ways to distinguish between more and less significant infringements. Rather, the problem is that there is no reason to be hopeful that there is a plausible theory of value that explains the differences in significance in property rights infringements in a way that simultaneously provides three things our traditional libertarian needs: 1) the distinction in significance of rights stems somehow from the thought that we are self-owners, or at least is not *ad hoc* or in tension with the Self-Ownership Thesis; 2) it vindicates traditional libertarian conclusions such as the broad impermissibility of paternalism and the near inviolability of our body when it comes to taking money, hair, or blood for others who badly need it; yet 3) it makes room for infringements on our property rights where we think we surely must permit them, such as in the case of pollution, acceptable risks such as flying aircraft over people's heads, and soft-paternalism cases such as pushing people out of the way of buses. Call these the three criteria of adequacy for a theory of value that vindicated the Self-Ownership Thesis as it has been traditionally urged. The challenge for traditional understandings of the upshot of the Self-Ownership Thesis is to come up with a theory of value that combines these three features.[40]

Let us consider the two broad types of theories of the variable significance of different infringements our libertarian might offer. This theory of value can either defer to the agent's own point of view in determining what makes an infringement more or less serious or not. Just to have labels, albeit imperfect ones, let us call the former subjectivist and the latter objectivist. There will be many different objective and subjective theories. All subjective theories will, in one way or another, defer to the

[40] For similar issues in a different context, see Samantha Brennan, "How is the Strength of a Right Determined? Assessing the Harm View," *American Philosophical Quarterly* 32 (1995), 383–92.

agent's will, choices, or preferences under certain conditions. To simplify matters I will focus on cases where the agent's will or choices are aligned with her preferences. That is, all the relevant subjectivist options point in the same direction.[41]

If the theory of value is objectivist, then although the agent who does not consent to infringement A or B but who expressly wills and prefers that A happen rather than B will, in some cases, nonetheless enjoy less protection from B than from A simply because infringement B is considered a less significant infringement on that agent's property. There will be cases where there is enough social good involved to make infringement B permissible but not infringement A. I want to say that whatever is generating this view about the relative moral significance of these actions, it is not stemming from the agent's self-ownership. To the extent that our ranking of the significance of different actions was stemming from the self-ownership of the agent the ranking should reflect in some way the agent's own view of the significance of the infringements upon her. If anything is bad about, for example, paternalism on a self-ownership view, it is that others who do not own something are making decisions about what will happen to that thing without gaining the consent of the person whose property it is. Naturally then, if we are looking for an account of what made one infringement worse than another on a self-ownership view we should think it is that one infringement is even less responsive to the will of the person whose property is being infringed.

Imagine what it would be for the opposite to be true. We would have to think that while Joe is a competent adult who fully owns something and unreservedly prefers that we infringe upon it in way A rather than B, nonetheless, *because it is his property* we should infringe in way B rather than A. The absurdity of that claim suggests that the objective account of what makes infringements differentially significant is a poor fit with the self-ownership view. That is, I am claiming this theory of value fails to capture well the first criteria of adequacy discussed above for a theory of value that fits with the Self-Ownership Thesis.

[41] Nozick and Otsuka both embrace subjective accounts. Nozick seems to defer to an agent's actual preferences, perhaps after they have experienced the boundary crossing. Otsuka appeals to more fully informed desires when discussing the sort of equality of opportunity for welfare view he adopts for goods that are not owner-occupied.

Alternatively, if our libertarian uses a subjectivist theory of value here, she will have a hard time vindicating the claim that there are powerful general considerations against forbidding someone from engaging in homosexual sex or forcing them to avoid saturated fats. People care about such prohibitions to different extents. Some may not much mind such infringements while not doing anything that counts as having consented to them. We do not consent to something merely by not minding it much. So the problem on the subjective side is that what different people value can differ so widely. As a result, the subjective theory of value will not be able to vindicate the thought that there are classes of actions, such as freedom of conscience, or freedom from interference with self-regarding actions, from which we all enjoy powerful protections. On the view under discussion there will be people who do not value or only slightly value this or that traditional libertarian sphere of protection, and then they will not enjoy powerful protections against infringements into that sphere. Such people will, on the subjectivist theory of value under consideration, enjoy only weak protections against some infringements. The traditional libertarian claim that we all enjoy powerful protections against action that infringes upon our property in such ways will not be vindicated. The threatened result would be that the status of our supposed libertarian protections of our Millian liberties on the value-responsive libertarian account will be quite subject to empirical fortune, not unlike consequentialism. Thus I am claiming such views score poorly in vindicating the second criteria of adequacy for a theory of value that can serve the needs of the Self-Ownership Thesis as it has traditionally been conceived.

I have so far tried to stress reasons that a self-ownership view seems to fit best with a subjective account of the significance of infringements. But in truth I would think any reasonable conception of the significance of infringements, whether subjectivist or objectivist, would have to allow that a range of infringements of the sort the traditional view rules out are not very significant and so we are entitled to only weak protections against them.

Thus, whether our libertarian embraces a conception of value that defers to the agent's own point of view or not, there appear to be serious costs for the Self-Ownership Thesis as traditionally understood. The above is sufficient to fuel the suspicion that there is nothing significantly more important about the aspect of self-ownership that is protected by the property rights that libertarians have traditionally stressed than the property rights that protect us from things such as pollution, risk, and soft-paternalism. That is, there is nothing about the value of what such rights protect that provides a principled basis for our libertarian's insistence that the former sort of rights are very stringent while allowing that the latter sort of rights are much more easily made permissibly infringeable for the sake of social goods. Thus, whatever reasonable theory of value we use to distinguish serious from trivial infringements such that flying aircraft and trivial pollution are permissible, will also justify a significant amount of redistribution and paternalism.

5.5 CONCLUSION

I have argued that the Conflation Problem poses a serious challenge to The Self-Ownership Thesis, and that the best way for such views to respond to the challenge is to reject All Infringements are Equal and allow that different kinds of infringements merit different levels of protection. We considered several other ways by which the view might be amended to handle the challenge, and found them less compelling.

Traditionally libertarian self-ownership views have claimed that the fact that something is our property gives us powerful protections against even trivial infringements on what we own. Thus, they have claimed, respecting property rights requires that we never or only rarely infringe on some people's property even for significant gains in social good. I have argued that this is not the most tempting or plausible interpretation of the rights of self-ownership. Maintaining that trivial property rights have such great strength would have consequences that no one is willing to accept. The most plausible understanding of the rights of property owners must allow that small infringements for significant gains are quite generally permissible, and thus that, for example, redistributive takings from those who will be little harmed by the loss and giving to the desperately needy will be broadly permissible on the most principled version of the Self-Ownership Thesis.[42]

[42] It is important to stress that I am only claiming that such redistribution is much less likely to infringe the best understanding of our natural property rights than the tradition has maintained. I have not tried to rule out that such redistribution might be morally forbidden or unwise for other reasons—such as that it is in fact counter-productive in the long run to social good. "Bleeding Heart Libertarians" maintain that traditional libertarian conclusions are warranted not because they flow from self-ownership but because they tend to make the world a better place. I have yet to begin to consider such views.

6

Consequentializing and Deontologizing: Clogging the Consequentialist Vacuum[1]

PAUL HURLEY

6.1 INTRODUCTION

A great deal of philosophical ink has been spilled of late arguing that all relevant values, including values that are traditionally endorsed by opponents of consequentialism, can be consequentialized. That many or even all plausible values can be consequentialized—incorporated into a ranking of states of affairs (typically evaluated as better or worse)—is often taken to lend support to the view that apparent alternatives *to* consequentialism are more plausibly understood as alternative forms *of* consequentialism, understood as the theory that the relevant deontic statuses of actions (such as their status as right or wrong) are determined through appeal to such a ranking of states of affairs.[2] All candidate moral theories worth considering are, on such a view, sucked up by a consequentialist vacuum.[3] Philosophers who

[1] I have benefitted from discussions of this essay with members of the Claremont Colleges Works in Progress Group—in particular, with Rivka Weinberg and Julie Tannenbaum. I am grateful to Mark Timmons for organizing the superb 2012 Arizona Normative Ethics Workshop, and to Sarah Stroud, Houston Smit, Thomas Hurka, Sean McCarron, Julia Driver, Jamie Dreier, and Doug Portmore for helpful discussions during the course of the Workshop. I have also benefitted from written comments on an earlier draft by Josh Glasgow and Alex Rajczi, and from extensive sets of comments on two different drafts from both Andrew Schroeder and Douglas Portmore. I regret that I cannot do full justice to their suggestions and insights in the space available.

[2] Douglas Portmore, "Consequentialism," in C. Miller (ed.), *The Continuum Companion to Ethics* (New York: Continuum Press, 2011).

[3] The "vacuum" metaphor was introduced by David McNaughton and Piers Rawling, "Agent-Relativity and the Doing-Happening Distinction," *Philosophical Studies* 63 (1991),

take themselves to be opposing consequentialism with alternatives are, on such an account, in the grips of a deep confusion. Properly understood, they are merely opposing one form of consequentialism with another.

Arguments that contain such appeals to consequentializing take two very different forms. The first, put forward most forcefully by Samuel Scheffler, Shelly Kagan, and David Cummiskey,[4] is concerned with the relationship between morally right action and states of affairs evaluated evaluator-neutrally. The second is concerned with the relationship between what agents ought and ought not to do and an evaluator-relative ranking of outcomes.[5] In what follows I will challenge many of the consequentializing arguments for both evaluator-neutral and evaluator-relative forms of consequentialism. I will argue in the next section that on the first, evaluator-neutral form of the consequentializing project, the plausibility of such impersonal consequential*izing* of values in itself has no implications for the acceptance of an evaluator-neutral consequential*ist* theory. It will become apparent, moreover, that the very plausibility of the evaluator-neutral consequentializing of certain values establishes the implausibility of an evaluator-neutral consequentialist account of such values.

In Section 6.3 I will take up the claim that an evaluator-relative form of the consequentializing vacuum succeeds where its evaluator-neutral counterpart fails. The problems I have identified with the

167–85. See also the invocation of the "consequentialist umbrella" by, for example, Jennie Louise, "Relativity of Value and the Consequentialist Umbrella," *The Philosophical Quarterly* 54 (2004), 518–36.

[4] Samuel Scheffler, *The Rejection of Consequentialism* (Oxford: Oxford University Press, 1982); Shelly Kagan, *The Limits of Morality* (Oxford: Clarendon Press, 1989); David Cummiskey, *Kantian Consequentialism* (Oxford: Oxford University Press, 1996).

[5] Variants of this evaluator-relative form have been developed by Amartya Sen, Jamie Dreier, Michael Smith, Jennie Louise, and Douglas Portmore, among others. See, for example, Amartya Sen, "Evaluator Relativity and Consequential Evaluation," *Philosophy and Public Affairs* 12 (1983), 113–32; Jamie Dreier, "Structures of Normative Theories," *The Monist* 76 (1993), 22–40; Michael Smith, "Neutral and Relative Value After Moore," *Ethics* 113 (2003), 576–98; and Jennie Louise, "Relativity of Value and the Consequentialist Umbrella." 54 (2004), 518–36. A variant upon arguments of this sort that appeals to the teleological nature of practical reasons is also offered by Douglas Portmore, for example, in "Consequentializing Moral Theories," *Pacific Philosophical Quarterly* 88 (2007), 39–73, and *Commonsense Consequentialism* (Oxford: Oxford University Press, 2011).

evaluator-neutral consequentializing argument do not, I will argue, beset its evaluator-relative counterpart. But I will demonstrate in Section 6.4 that the plausible candidate theories that can be consequentialized, located within a consequentialist evaluative framework, can also readily be "deontologized," located within an evaluative framework that is congenial to the articulation of deontological and other non-consequentialist theories. This suggests that it is other premises that are doing the heavy lifting in any evaluator-relative consequentializing argument for consequentialism. It is not consequentializing *per se*, but the possibility of such consequentializing in light of what are taken to be certain distinctive advantages of a consequentialist evaluative framework, that constitutes the consequentializing case for consequentialism. In particular, the consequentialist evaluative framework, which relates the deontic status of actions to the telic status of states of affairs, is taken by many to be uniquely suited to accommodate both the Compelling Idea that it is always permissible to do what is best,[6] and the Explanatory Thought that a satisfactory explanation of deontic evaluations of actions—of what it is right to do and what ought to be done—will appeal to goodness.[7] According to this Thought, a theory of the goodness of states of affairs will play a central role in any adequate rationale for deontic evaluation of actions.[8] If plausible alternatives must

[6] This Compelling Idea, as standardly formulated, relates deontic evaluation of actions to the evaluation of outcomes or states of affairs as better, worse, and best. For traditional characterizations of this Idea, see Samuel Scheffler, *The Rejection of Consequentialism*, p. 4, and Philippa Foot, "Utilitarianism and the Virtues," in Samuel Scheffler (ed.), *Consequentialism and Its Critics* (Oxford: Oxford University Press, 1988), p. 227. For more recent characterizations of the Idea as relating deontological evaluation of actions to the goodness of outcomes, see Mark Schroeder, "Teleology, Agent-Relative Value, and 'Good'," *Ethics* 117 (2007), 265–300, and Jamie Dreier, "In Defense of Consequentializing," Mark Timmons (ed.), *Oxford Studies in Normative Ethics, Vol.* 1 (New York: Oxford University Press, 2011), pp. 97–119. I will work with this standard characterization of the Compelling Idea going forward. Douglas Portmore has recently offered an alternative characterization of the Compelling Idea in his *Commonsense Consequentialism* (Oxford: Oxford University Press, 2011), for example, p. 5.

[7] For various characterizations of this Explanatory Thought see Schroeder, "Teleology, Agent-Relative Value, and 'Good'," Benjamin Sachs, "Consequentialism's Double-Edged Sword," *Utilitas* 22 (2010), 258–71, and the discussions of value-based rationales in Kagan, *The Limits of Morality*, and Hurley, "Agent-centered Restrictions: Clearing the Air of Paradox," *Ethics* 108, pp. 120–46.

[8] But see, for example, Dreier ("In Defense of Consequentializing"), who endorses the Compelling Idea, and takes it to count in favour of considering all theories in their consequentialized forms, but eschews the Explanatory Thought. For other consequentialists, however, the Explanatory Thought is considered central to consequentialism, and a consequentialized theory that eschews the Explanatory Thought is considered to be a non-consequentialist theory presented in consequentialized form.

accommodate this Compelling Idea and/or the Explanatory Thought, and only theories in their consequentialized form—formulated within a consequentialist evaluative framework—can plausibly accommodate this Idea and/or this Thought, then any theory that *can* be consequentialized *should* be consequentialized.

But I will demonstrate in Section 6.5 that the deontologizing alternative introduced in Section 6.4 suggests an alternative evaluative framework that can also accommodate what is attractive in both the Compelling Idea and the Explanatory Thought. This evaluative framework takes the deontic evaluation of actions as permissible and obligatory to be determined through appeal to the evaluation of actions and reasons for action as better or worse, and in some cases as best. On this framework, agents ought to perform the best action, the action decisively supported by good reasons. Thus, whereas the consequentialist evaluative framework holds that deontic statuses of actions are determined through appeal to the ranking of states of affairs, typically as better or worse, this alternative evaluative framework holds that deontic statuses of actions are determined through appeal to the evaluation of actions and reasons for action as good, better, and best.[9] It will become clear that such an alternative evaluative framework can accommodate what is compelling in consequentialists' Compelling Idea, and what is attractive in their Explanatory Thought.

Consequentializing does not provide support for consequentialism, nor does consequentializing augmented by appeal to the Compelling Idea and/or the Explanatory Thought. Rather, such arguments taken by themselves beg the question against many non-consequentialist alternatives. It is necessary to move beyond such arguments to adjudicate the core disputes among advocates of consequentialist, deontological, and other moral theories. In Section 6.6 I will suggest that it is the alternative evaluative framework that is neutral between consequentialism and many forms of non-consequentialism. It is this alternative that can function as a shared evaluative framework within which the merits of

[9] Although virtually all consequentialists adopt the consequentialist evaluative framework, some do not. Moreover, consequentialists who accept the consequentialist evaluative framework are not precluded from accepting the alternative evaluative framework as well. Indeed, I will suggest below that the consequentialist evaluative framework is most plausibly understood as a particular interpretation of the alternative evaluative framework. The relationship between these two evaluative frameworks will be explored in Section 6.6.

such alternatives can be considered without begging the question either way. I will close with a brief characterization of several of the central points of disagreement between consequentialists and their critics that come clearly into view within the context of this alternative evaluative framework.

6.2 CLOGGING THE EVALUATOR-NEUTRAL CONSEQUENTIALIST VACUUM

Traditional versions of consequentialism identify one ultimate value, such as pleasure, happiness, or well-being, and identify the right action as the action that maximizes the value in question overall, from an impersonal, evaluator-neutral standpoint. The position that there is one such ultimate value in terms of which outcomes are appropriately evaluated is often characterized as value monism;[10] the position that the moral relevance of this value is exhausted by its role in the ranking of states of affairs from an evaluator-neutral, impersonal standpoint is often characterized as evaluator or agent-neutrality. The form of consequentialism that will be considered in this section rejects monism in favour of pluralism, but maintains evaluator-neutrality. The form of consequentialism that will be considered in Sections 6.3 and 6.4 is committed neither to monism nor to evaluator-neutrality.

The traditional commitment to value-monism placed consequentialists in an awkward position—a position in which they seemed unable to take many important values seriously. Valuing rights, or autonomy, or respect for persons could only be vindicated on such an account to the extent that such values contributed to the single end identified as ultimate, such as happiness, pleasure, or well-being. But such a merely derivative role simply did not seem to do these values justice. The rejection of value monism by Peter Railton, Samuel Scheffler, and others appeared to eliminate this awkwardness, and seemed to breathe new life into evaluator-neutral consequentialism. If other values in addition to happiness or pleasure or well-being seem to be equally fundamental,

[10] See Peter Railton, "Alienation, Consequentialism, and the Demands of Morality," in Samuel Scheffler (ed.), *Consequentialism and Its Critics* (Oxford: Oxford University Press, 1988), for a characterization of this monistic feature of traditional consequentialist theories, and for a rejection of this feature in favour of value pluralism.

why not take all such values into account in the determination of the best overall outcome? The right action could then be recognized as the action that maximizes value thus pluralistically understood. Such a position provides the consequentialist with a powerful response to the challenge that she does not take values such as rights, autonomy, and respect seriously. After all, if rights are intrinsically valuable, is it not a better outcome, all other things being equal, upon which more rather than fewer rights are upheld? Why is not the position that takes rights the most seriously one which maximizes the extent to which they are upheld and respected? If you are really serious about rights, in short, should you not maximize the extent to which they are respected?[11]

Two claims are run together in this influential line of argument. The first is the evaluator-neutral consequential*izing* claim—that whatever intrinsic values there are are appropriately "consequentialized," taken into account in the evaluation of outcomes as better or worse overall. The second is the evaluator-neutral consequential*ist* claim—that the right action is the action that promotes the best overall outcome thus determined (or at least that such an action is always morally permissible). The first claim—that there is a plurality of intrinsic values, and that all such values are appropriately taken into account in the evaluator-neutral ranking of outcomes—has struck many as having considerable plausibility.[12] If rights or personal autonomy or respect for persons as ends in themselves has intrinsic value, an outcome that secures greater autonomy or respect for persons by others clearly seems better than one that does not. To simplify, let us grant this claim in what follows.

This first claim is often taken to support the second. Once we grant that rights, for example, are appropriately taken into account in the determination of the best overall outcome (evaluator-neutrally consequentialized) we should, it is suggested, recognize that the action that takes rights seriously—the right action—is the action that maximizes the extent to which they are respected and minimizes their violation (evaluator-neutral consequentialism). But the appropriateness

[11] See Robert Nozick, *Anarchy, State, and Utopia* (New York: Basic Books, 1974), Samuel Scheffler, *The Rejection of Consequentialism*, and David Cummiskey, *Kantian Consequentialism*, for discussions of such proposals to adopt a consequentialism of rights or Kantian respect.

[12] But see, for example, T. M. Scanlon, *What We Owe to Each Other* (Cambridge: Harvard University Press, 1998).

of evaluator-neutrally consequentializing values, by itself, provides no support for a consequentialist account of such values; indeed, the appropriateness of the evaluator-neutral consequentializing of certain values entails the straightforward rejection of evaluator-neutral consequentialism.

I believe that Bernard Williams can plausibly be understood as anticipating this point; indeed, as mounting an evaluator-neutral consequentializing argument against evaluator-neutral consequentialism.[13] He argues that evaluator-neutral consequentialism is caught between a Scylla and a Charybdis—either the evaluator-neutral consequentialist can have a plausible account of impersonal consequential*izing*, of the determination of better and worse overall states of affairs, or he can endorse evaluator-neutral consequential*ism*, but not both. Among agents' projects, Williams suggests, are those "which flow from some more general disposition towards human conduct and character, such as a hatred of injustice, or of cruelty, or of killing."[14] He elsewhere includes projects that ascribe "intrinsic and non-instrumental value to various activities and relations such as truth-telling, loyalty, and so on."[15] Williams then asks whether such paradigmatically moral, intrinsic, evaluator-relative projects are appropriately "counted among the projects whose satisfaction is to be included in the maximizing sum"—whether such moral projects are appropriately consequentialized. If the answer given is "no," then the consequentialist is "almost certainly committed to a version of utilitarianism as absurdly superficial and shallow as Benthamite versions have often been accused of being."[16] That is, if we exclude from the values to be consequentialized those picked out by projects "which . . . presuppose a reference to other people's projects," such a criterion would eliminate all that "was not blankly and in the most straightforward sense egoistic."[17] Here lies Scylla, the commitment to an implausible, because impoverished, account of intrinsic value, hence of better and worse overall outcomes.

[13] I explore Williams' argument in more detail in ch. 4 of *Beyond Consequentialism* (Oxford: Oxford University Press, 2009).

[14] *Utilitarianism For and Against* (Cambridge: Cambridge University Press, 1973), p. 111.

[15] *Making Sense of Humanity* (Cambridge: Cambridge University Press, 1995), p. 164.

[16] *Utilitarianism For and Against*, p. 111.

[17] *Utilitarianism For and Against*, p. 111.

The alternative is to countenance such values, such fundamentally agent-relative moral projects as staying loyal to one's friends, or being honest in one's dealings with others, or avoiding injustice. Williams argues that such a move will allow for a plausible account of better and worse overall outcomes, but it is also to encounter Charybdis: If it is such agent-relative values that it is appropriate for someone to take seriously at the deepest level, as what his life is about, then "it is absurd to demand of such a man . . . that he should step aside from his own project and decision and acknowledge the decision which utllitarian calculation requires."[18] The intrinsic agent-relative moral values that are consequentialized will ground actions that do not bring about the best overall outcome, such as telling the truth, even though more lies will be told by others, or remaining loyal to one's friends even though others will betray theirs. In short, the very recognition that there are such intrinsic agent-relative moral values, and that it is appropriate to consequentialize such values, is recognition that evaluator-neutral consequentialism is false. This is not, of course, to say that there might be no plausible way to steer between Scylla and Charybdis (to enrich the account of intrinsic value without embracing values that undermine consequentialism).[19] The point is that the consequentializing of many values leads consequentialism directly into Charybdis, and that the task confronting the evaluator-neutral consequentialist is far more daunting than has commonly been recognized.

Consider, for example, a deontologist who recognizes Kantian respect as an intrinsic agent-relative moral value: each person has a claim upon each other person to be treated as an end in themselves and not as a mere means to the promotion of that agent's good or the overall good, hence each person has an agent-relative moral reason to treat each other person with such Kantian respect. We are granting, if only for the sake of argument, that it is appropriate to consequentialize such an intrinsic agent-relative value; that, for example, it will be better overall if fewer rather than more people are treated as mere means to the promotion of

[18] *Utilitarianism For and Against*, p. 116.

[19] Frances Kamm explores one such possibility on pp. 263ff of *Morality, Mortality, Vol, II* (New York: Oxford University Press, 1996). Julia Driver has reminded me that Peter Railton ("Alienation, Consequentialism, and the Demands of Morality") is plausibly understood as exploring another such possibility, embracing a plurality of values, but rejecting Williams' claim that evaluator-relative *moral* values need be among them.

the overall good. Recognition that it is appropriate to consequentialize this intrinsic, agent-relative constraint on the promotion of the overall good just is recognition that there are moral constraints on the promotion of the overall good. But this is to acknowledge that to recognize such a value among those to be consequentialized is to be committed to the rejection of evaluator-neutral consequentialism. If rights and/or respect, properly understood, constrain the promotion of the overall good, then it may well be appropriate to consequentialize such values—to take overall goodness to be determined in part by what promotes the upholding of rights and respectful interaction overall. But *ex hypothesi*, the values that are consequentialized are values that constrain the promotion of the best overall state of affairs, including the maximization of the extent to which rights are upheld.

It may be objected that the inclusion of such a value among those to be consequentialized simply stacks the deck against the consequentialist. Of course it does! But that is precisely the point: the appropriateness of evaluator-neutrally consequentializing some value in no way indicates that such a value can be reconciled to evaluator-neutral consequentialism. Even if taking such a value into account presupposes the rejection of evaluator-neutral consequentialism, it may well be appropriate to consequentialize such a value—to take it into account in the evaluation of overall states of affairs. Indeed, for non-consequentialists the apparent appropriateness of consequentializing such values in certain contexts seems merely to demonstrate what common sense, and their own theories, have always suggested, that the determination of better and worse overall states of affairs can play the role at most as one of the factors that are relevant in the determination of what agents are morally required and permitted to do.[20]

[20] This problem is sometimes elided in consequentializing arguments for evaluator-neutral consequentialism by appeal to a form of the Explanatory Thought. Thus Scheffler, for example, suggests that "defenses of agent-centered restrictions in the literature often seem to take the form of an appeal to the disvalue of violation" (*The Rejection of Consequentialism*, p. 88), hence that any defense of the prohibition against treating persons as means only, for example, must appeal to the badness of such treatments happening. Notice, however, that to require such an evaluator-neutral consequentialist defense of all intrinsic agent-relative values is just to reject the claim that there can be any such agent-relative intrinsic values in the first place, since, *ex hypothesi*, to recognize such intrinsic agent-relative constraints on the promotion of the overall good is just to reject the claim that all relevant moral reasons must have an evaluator-neutral, value-based rationale.

6.3 THE EVALUATOR-RELATIVE VACUUM

The evaluator-neutral consequentialist vacuum is blocked by recognition of the appropriateness of taking certain agent-relative moral values into account in the determination of the best overall outcome. To maintain that it is appropriate to impersonally consequentialize such impartial but agent-relative values is to be committed to the rejection of an evaluator-neutral consequentialist account of moral value. The moral of this story, however, is taken by many consequentialist sympathizers not to lead away from consequentialism, but towards a more defensible form.[21] The evaluator-neutral consequentialist vacuum is blocked because, although the evaluator-neutral consequentializing operation can recognize intrinsic agent-relative moral values, evaluator-neutral consequentialism cannot. The solution, they argue, is not to abandon consequentialism, but to reject the traditional assumption that the relevant consequentialist standpoint is an evaluator-neutral, impersonal standpoint. If the best consequences are determined from an evaluator-relative rather than an evaluator-neutral standpoint, then the sucking up of agent-relative values into the consequentializing vacuum need not be an obstacle to maintaining that agents ought to promote the best consequences. Thus, although the claim that I should avoid violating someone's right, even when this will result in more rights being violated, cannot readily be reconciled to evaluator-*neutral* consequentialism, it can readily be reconciled to evaluator-*relative* consequentialism. The evaluator-relative consequentialist can take such a claim as indicating that violations of rights by me have a disvalue that is not entirely captured impersonally, and that an evaluator-relatively worse outcome can result if I violate a right even if fewer rights will be violated overall. It may be objected that such evaluator-relative consequentialism can nonetheless result in my violating someone's right to minimize my own rights violations overall, hence that someone who would reject such a claim will reject even evaluator-relative consequentialism. But as

[21] Amartya Sen, "Evaluator Relativity and Consequential Evaluation" *Philosophy and Public Affairs* 12 (1983), pp. 113–32; Jamie Dreier, "Structures of Normative Theories" *The Monist* 76 (1993), 22–40; Michael Smith, "Neutral and Relative Value After Moore" *Ethics* 113 (2003), 576–98; Jennie Louise, "Relativity of Value and the Consequentialist Umbrella" *The Philosophical Quarterly* 54 (2004), 518–36; Douglas Portmore, "Consequentializing Moral Theories," *Pacific Philosophical Quarterly* 88 (2007), 39–73, and *Commonsense Consequentialism* (Oxford: Oxford University Press, 2001).

Jennie Louise and others have argued, this objection can seemingly be addressed by taking the relevant evaluation of consequences to be not just relative to the evaluator, but relative in addition to the time. Such an account can maintain that it is evaluator-relatively worse at this time to violate someone's right, even if this will result in my violating more people's rights overall.[22]

Following Jamie Dreier, Douglas Portmore suggests that the consequentializing process, once freed from the limitation to determinations of evaluator-neutral goodness, is straightforward:

> Take whatever considerations that the non-consequentialist theory holds to be relevant to determining the deontic statuses of actions and insist that those considerations are relevant to determining the proper ranking of outcomes.[23]

Building upon their accounts of the consequentializing process, Portmore and Dreier put forward forms of a "deontic equivalence thesis":

> For any remotely plausible non-consequentialist theory, there is a consequentialist counterpart theory that is deontically equivalent to it[24]

Certain challenges confront such an evaluator-relative consequentializing strategy, particularly as a component of an argument that purports to support consequentialism. These problems can be highlighted by focusing upon certain features that have traditionally been thought to make consequentialism attractive—even unavoidable—as a moral theory. The first is that consequentialism can accommodate what has been labeled by Mark Schroeder as the "Compelling Idea," which he

[22] "Relativity of Value and the Consequentialist Umbrella," 533–5.

[23] "Consequentializing Moral Theories," p. 39. Portmore has more recently offered an alternative characterization of consequentializing in ch. 4 of his *Commonsense Consequentialism*. Dreier characterizes the consequentializing strategy as follows: "The main strategy for consequentializing any moral theory is simple. We merely take the features of an action that the theory considers to be relevant, and build them into the consequences" ("Structures of Normative Theories," p. 23).

[24] Portmore, "Consequentializing Moral Theories," p. 40. See also Dreier, "In Defense of Consequentializing," who deploys the phrase "deontic equivalence thesis." Campbell Brown has recently mounted a challenge to strong forms of this equivalence claim, arguing that not all non-consequentialist theories can plausibly be provided with a consequentialist counterpart. But such challenges are tangential to the arguments presented here. See his "Consequentialize This," *Ethics*, 121 (July 2011), 749–71.

characterizes as the idea that it is always at least permissible to bring about the most good.[25] The second is that consequentialism can accommodate what can appear to be a plausible constraint upon any adequate explanation of deontic values—that "facts about what people ought to do" ought to be explained by "facts about what is good."[26] This Explanatory Thought is sometimes characterized as the claim that an adequacy condition on any acceptable account of deontic evaluation of actions is that such claims about which actions agents ought and ought not to do should have a value-based rationale, a rationale that grounds such evaluations in an appeal to the good.[27] Evaluator-neutral consequentialism is taken to satisfy this Explanatory Thought because it takes acts to be right, or obligatory, or permissible *because* they bring about the best overall outcome: It provides a value-based rationale for deontic evaluations of acts.[28]

Evaluator-neutral consequentializing invokes an intuitively plausible notion of goodness, the evaluator-neutral goodness of overall states of affairs. But we have seen that it provides no support for evaluator-neutral consequentialism. Indeed, depending upon the values that are recognized as appropriate to consequentialize, it can entail the rejection of evaluator-neutral consequentialism. Evaluator-relative

[25] Scheffler characterizes this as the idea that "it is always permissible to do what would have the best outcome" (*The Rejection of Consequentialism*, p. 4), and Foot as the idea that "it can never be right to prefer a worse state of affairs to a better," or it must "not be irrational to prefer the worse to the better state of affairs" ("Utilitarianism and the Virtues," p. 227). See also my discussion in "Scheffler's Argument For Deontology," *Pacific Philosophical Quarterly* 74 (1993), 118–34. Such formulations of the Compelling Idea take it to relate the deontic status of actions to the value of outcomes. Douglas Portmore has recently argued for a variant upon this Idea upon which it relates the deontic status of actions to outcomes, but not to the value of outcomes.

[26] See fn. 6.

[27] Such formulations emphasizing the need for a value-based rationale are found in both Scheffler, *The Rejection of Consequentialism*, and Kagan, *The Limits of Morality*.

[28] Benjamin Sachs cashes out this Explanatory Thought in terms of "determination claims": "determination claims identify the kinds of fact that ground an action's rightness or wrongness" ("Consequentialism's Double-edged Sword," *Utilitas* 22 (2010), 264), while moral principles simply state "necessary and/or sufficient conditions for an action's rightness or wrongness" (264). Using the distinction, the Compelling Idea is taken to point in the direction of a consequentialist moral principle, that promotion of the good is a sufficient condition for an act's being permissible. The Explanatory Thought is taken to point in the direction of a consequentialist determination claim, that promotion of the good explains—provides a grounding rationale for—the judgment that the act in question is permissible. Why is it at least permissible? Because it promotes the good.

consequentializing avoids the particular problem that plagues the consequentializing argument for evaluator-neutral consequentialism, but it is less clear that it appeals to a meaningful notion of goodness, hence that it can meaningfully capture the traditional Compelling Idea behind consequentialism. We have seen that if rights are intrinsically valuable, there seems to be some intuitive plausibility to the idea that it is a better outcome overall upon which more rights are upheld. But in what sense is it a better outcome upon which I uphold rights even though this will result in fewer rights being upheld? Such an outcome will not be better overall. But it may also not be better *for me*—it may well be that I will suffer greatly by upholding such a right. In such cases it will be both worse for me and worse overall to uphold the right. If such an intrinsic value is nonetheless taken to be consequentialized into an evaluator-relative ranking of outcomes, there must be some evaluator-relative ranking of outcomes—as better or worse not *for* me but somehow relative *to* me—that is invoked by such a consequentializing move. Although there is intuitive support for some form of the evaluator-neutral consequentializing move, it is far less clear that there is any such intuitive notion of a ranking of outcomes relative-to-me to which the evaluator-relative consequentializing move can appeal. As Mark Schroeder has argued:

> This is where I get lost. *Good* and *good for*, after all, are concepts that I can understand…But since I do not understand what "good-relative-to" talk is all about, I do not understand how it could be appealing to think that you shouldn't do something that will be worse relative-to-you.[29]

To the extent that it is unclear whether there is a meaningful notion of evaluator-relative goodness of outcomes, it is unclear how an evaluator-relatively consequentialized theory can satisfy what consequentialists have typically taken to be their own Compelling Idea.[30]

In addition, such a consequentializing strategy appears to be in jeopardy of losing the Explanatory Thought.[31] The thought, recall, is that an adequate moral theory provides a value-based rationale for claims that

[29] Schroeder "Teleology, Agent-Relative Value, and Good," p. 291.

[30] See Dreier's "In Defense of Consequentializing," in which he attempts to address Schroeder's criticism.

[31] Schroeder "Teleology, Agent-Relative Value, and Good."

agents ought and ought not to perform certain actions. The consequentialist rationale purports to ground such evaluations of actions in the evaluation of states of affairs. But because such a consequentializing formula appears to allow for rankings of states of affairs that are themselves explained by evaluations of actions as permissible or impermissible, it appears to allow for consequentialized theories that do not capture the consequentialists' version of the Explanatory Thought. Consider, for example, an account that deontically ranks actions, identifies the outcomes of performing these various actions, ranks these outcomes so that they reflect the antecedent deontic ranking of the actions from which they result, and identifies the action that promotes the highest ranked outcome as the action that the agent is obligated to perform. It will be true on such an account that the agent in question ought to do what is evaluator-relatively best in this sense, and the resulting view will be in some formal sense consequentialized. But the ranking of states of affairs need not play any substantive explanatory role in such a theory at all—it can produce consequentialized theories upon which the ranking of states of affairs as better or worse is an explanatory fifth wheel.[32] For those consequentialists who take a defining feature of consequentialism to be the distinctive explanatory rationale that it offers for the rightness of actions through appeal to the goodness of outcomes, such an evaluator-relative consequentializing process can produce consequentialized theories that are not, in their view, consequentialist.[33]

[32] Portmore makes the epistemological analog of this point, arguing that if we adopt this straightforward consequentializing strategy, which he labels a "Footian procedure," there will be certain consequentialized theories upon which the "only way we can come to know that X's outcome is outranked by Y's outcome it to first know that X is wrong and Y is not." (See, for example, *Commonsense Consequentialism*, pp. 112–13). The ranking of outcomes will have become an epistemological fifth wheel. He has pointed out to me in correspondence that in earlier drafts I did not sufficiently distinguish such an epistemological claim from the explanatory or metaphysical claim that is central to the Explanatory Thought. An appeal to outcomes can be epistemically parasitic without being explanatorily or metaphysically parasitic—there could well be a consequentialized moral theory upon which the appeal to outcomes plays no epistemological role but plays a robust explanatory role. My point here is simply that such a straightforward consequentializing formula allows for consequentialized moral theories upon which the evaluation of outcomes plays no significant explanatory role.

[33] Although again see Dreier's argument in "In Defense of Consequentializing," in which he embraces the Compelling Idea but not the Explanatory Thought. My point here is that the fact that a theory can be consequentialized in no way establishes that it can avail itself of the consequentialist form of the Explanatory Thought.

6.4 THE DEONTOLOGIZING ALTERNATIVE

There are significant headwinds confronting the "good" to which the evaluator-relative consequentializer appeals in his proposal for satisfying the Compelling Idea; moreover, the consequentializing process that he proposes seems able to generate consequentialized theories that cannot satisfy the consequentialist's Explanatory Thought. Still, a wide range of plausible theories can, apparently, be consequentialized—placed within a framework that invokes a notion of evaluator-relative good and provides at least the formal structure for an answer along the lines suggested by the Explanatory Thought. What does this show? Does it provide support for consequentialism?

I will demonstrate in this section and the next that it does not. The formula for consequentializing apparently deontological theories is reflected in a roughly parallel formula for "deontologizing" apparently consequentialist theories. That all consequentialist theories can thus be in this sense "deontologized" may well provide no support for deontology, but for parallel reasons it will become clear that the consequentializing of all deontological theories by itself provides no support for consequentialism. I will then demonstrate, in Section 6.5, that the evaluative framework within which this "deontologizing" of plausible alternatives can be carried out can satisfy the legitimate demands that underlie both the Compelling Idea and the Explanatory Thought.

The deontologizing strategy that I have in mind can best be presented by way of an initial contrast with its consequentializing counterpart. The consequentializing strategy, recall, relates claims about what agents ought to do to *outcomes*, and takes considerations that appear to be relevant to the deontic status of actions to be considerations relevant to the ranking of the outcomes of such actions (for example, from better to worse). The corresponding deontologizing strategy, as I will understand it, relates claims about what agents ought to do to rankings of *actions* (as good and bad, better and best), and takes any considerations that appear to be relevant to the telic status of outcomes, for example, as better or worse overall or better or worse for me, to be considerations that are relevant to the ranking of actions as better and worse, and in some cases best.[34]

[34] Just as Michael Smith argues that there need be "no claims about the completeness or otherwise of rankings" of outcomes for consequentialists ("Two Kinds of Consequentialism," *Philosophical Issues*, 19 (October 2009), 257–72), so too there need be no claim about the

As deontology is often understood, particularly by consequentialists, it does not endorse an evaluative framework—an account of the relationship between what agents ought to do and evaluation as better or worse. It therefore must simply reject the Compelling Idea, maintaining that on the only relevant notion of goodness, it is often wrong to do what is best. It must also reject the Explanatory Thought, maintaining that there are non-value-based constraints upon doing what is best, constraints that clearly cannot have a value-based rationale. Obligations are thought somehow simply to constrain value. The deontologizing alternative that I have suggested, by contrast, invokes such an evaluative framework. On this alternative framework it is the goodness and badness of actions and reasons for action that determine deontic values of actions: agents ought to perform—are obligated to perform—the best action, understood as the action decisively supported by good reasons.[35] Such a framework need not deny that good, better, and best are properties of states of affairs; it need only claim that good reasons provide evidence for taking certain courses of action to be good, the best, or at least better than others, and that when such evidence for doing p is decisive with respect to doing any other q, we take p to be the best course of action available. It is the best action in this sense, according to such an alternative evaluative framework, that the agent ought to do. Stephen Darwall has suggested, developing a distinction introduced by David Velleman, that "we could say that the formal aims of belief and action are, respectively, to believe and act . . . as is best supported by normative reasons (and so, in this sense, as is best)."[36] The alternative evaluative

completeness of rankings of actions. For discussions of versions of Kantian deontology that explicitly relate claims about what agents ought to do to reasons and actions evaluated as better and worse, see Barbara Herman, "Leaving Deontology Behind," and Michael Ridge, "Consequentialist Kantianism," *Philosophical Perspectives* 23 (2009), 421ff.

[35] Derek Parfit (*On What Matters*, *Vol.* 1 (Oxford: Oxford University Press, 2011) identifies cases in which an agent has decisive reasons to perform some action. To have decisive reasons in Parfit's sense is just for an action to be supported by the best reasons in this sense, and for it to be the best course of action. Other accounts may differ from Parfit's in allowing that some course of action may be supported by the best reasons, but there may nonetheless be sufficient reasons to act some other way. (Gary Watson characterizes such reasons as "good enough"; "The Work of the Will," in Watson, G., *Agency and Answerability* (Oxford: Clarendon Press, 2004, p. 127)). In what follows I will follow Parfit in characterizing an agent as having decisive reasons to act in some way when she has the strongest reasons to act that way; that is, when it is the action supported by the best reasons—the best course of action.

[36] Darwall, *The Second-Person Standpoint* (Cambridge: Harvard University Press, 2006), p. 279.

framework maintains that an agent ought to do what is in this sense best—what it is best to do. Just as the evaluator-relative consequentializing formula relates obligation to act to the evaluation of outcomes as better and best, so the deontologizing formula relates obligation to act to the evaluation of actions and reasons for action as better and best. The former takes relevant deontic statuses of actions to be determined through an appeal to a ranking of outcomes, for example, as better or worse; the latter, by contrast, holds that relevant deontic statuses of actions are determined through appeal to an evaluation of actions and reasons for action as good, better, and best.

As I suggested in the opening section, this initial contrast between the two evaluative frameworks does not rule out the possibility that a particular moral theory could embrace both. A consequentialist could maintain, for example, both that an agent ought to do what it is best to do, what is decisively supported by good reasons, and that such decisively good reasons, properly understood, are provided entirely through appeal to the value of states of affairs.[37] But the alternative framework takes the relationship invoked by the Compelling Idea and the Explanatory Thought, the relationship in light of which they present themselves as virtual platitudes, to be that between deontic evaluation of actions and the evaluation of actions and reasons for action as good, better, and best. The consequentialist evaluative framework, by contrast, takes the relationship invoked by the Idea and the Thought to be between the deontic evaluation of actions and the evaluations of states of affairs as good, better, and best. The availability of such an alternative framework demonstrates that non-consequentialists need not eschew

[37] Indeed, as I will suggest in Section 6.6, consequentialist moral theories are best understood as accepting the alternative evaluative framework, and as offering an interpretation of what it is best to do upon which the goodness of actions is determined entirely through appeal to the relevant ranking of states of affairs, typically as better or worse. Non-consequentialists, by contrast, offer alternative interpretations of what it is best to do upon which the value relevant to the evaluation of actions is not limited to the value of states of affairs, and the reasons relevant to the evaluation of action are not limited to reasons to promote states of affairs. Douglas Portmore suggested to me in conversation that many of the consequentialists he queried in fact do readily accept the alternative evaluative framework, and recognize consequentialism not as denying that we ought to do what it is best to do, but as offering a distinctive interpretation of what it is best to do. Roger Crisp suggested to me in conversation that many virtue ethicists will also readily accept such an alternative evaluative framework, but see themselves as offering a more plausible alternative to the consequentialist interpretation of what it is best to do.

an evaluative framework in eschewing the consequentialist evaluative framework, and that the endorsement of a relationship between rightness and goodness by itself provides no support for consequentialism relative to its deontological rivals. Moreover, as I will argue in more detail below, the availability of such an alternative framework allows deontological, virtue ethicist, and other non-consequentialist theories to reject the consequentialist evaluative framework while accommodating both what is compelling in the Compelling Idea and what is intuitive in the Explanatory Thought.

Playing upon Dreier's and Portmore's consequentializing proposals, the proposal for "deontologizing" consequentialist theories within the context of such an alternative evaluative framework would have the following form:

> Take whatever considerations determine the telic statuses of outcomes, and insist that these considerations are reflected in good reasons, reasons that are relevant to identifying the statuses of actions as better and worse.

This deontologizing proposal suggests a telic equivalence thesis:

> For any plausible consequentialist theory, we can construct a version of deontology that is equivalent to it.

Deontologizing, thus understood, presents deontic statuses as determined not in the first instance through appeal to the ranking and evaluation of outcomes as good, better, or best, but through appeal to the ranking and evaluation of actions and reasons for actions as good, better, and, when applicable, best. Just as the consequentializing formula appropriates any considerations given by non-consequentialists for determining the deontic statuses of actions, the deontologizing formula appropriates any considerations given by consequentialists for determining the statuses of outcomes, treating them as providing good reasons that are relevant to the determination of the best course of action. Just as the consequentializing formula insists that any such considerations that determine deontic statuses be treated as considerations for determining how the outcomes of such actions rank, so too the deontologizing formula insists that any such considerations of the telic statuses of outcomes be treated as reflected in reasons that are relevant to the ranking of actions as better and best. The consequentializing formula insists that the considerations that are taken by non-consequentialists as obligating us perform actions should always be treated as relevant to the ranking of outcomes that we are obligated to promote; the deontologizing formula insists that the considerations that are taken by consequentialists as relevant to the ranking of outcomes that we are obligated to promote should always be treated as relevant to the ranking of actions that we are obligated to perform. Telic assessments of states of affairs are relevant on the deontologizing approach, thus understood, only insofar as they are reflected in good reasons for action (to the extent to which they count in favour of action). The best course of action is that decisively supported by good reasons, and the framework proposes that it is what it is best to do thus understood that the agent ought to do.

Thus, traditional rational egoism holds that agents ought to do what promotes the best outcome for the agent. The deontologized version of rational egoism takes such considerations about outcomes to be relevant only insofar as they are reflected in good reasons for action. Rational egoism, on such an approach, becomes the "deontologized" theory that the agent ought to do what it is best to do, what she has the best reasons to do, and what she has decisively good reasons to do is what promotes the best outcome for her. On such a deontologized account, the deontic status of the action (that the agent ought to perform it) is determined through appeal to the evaluation of actions and reasons for action as good and best. Rational egoism has been deontologized. As this example suggests, standard candidates for consequentialist theories can readily be deontologized. In their deontologized forms, rational egoism, welfare utilitarianism, and so on, are simply different accounts of the reasons that are relevant to determining the best course of action, the action that an agent ought to perform.

6.5 THE ALTERNATIVE EVALUATIVE FRAMEWORK: ALTERNATIVE INTERPRETATIONS OF THE COMPELLING IDEA AND THE EXPLANATORY THOUGHT

The ability to deontologize all candidate consequentialist theories suggests that the ability to evaluator-relatively consequentialize all candidate non-consequentialist theories itself provides no support for adopting the consequentialist evaluative framework. It may be thought, however, that such support for considering all theories in their consequentialized form can be provided by appeal to the traditional Compelling Idea or the Explanatory Thought. Beginning with the Compelling Idea, it may seem that despite the problems plaguing their appeal to goodness relative-to-me, such consequentialized theories can at least accommodate this idea that it is always right, or at least permissible, to promote the good. Non-consequentialist theories, by contrast, may seem to be precluded *ex ante* from incorporating this attractive Idea.

It seems clear upon reflection, however, that the consequentialist's traditional presentation of the Compelling Idea as the idea that it is always permissible to promote the good effectively runs together an idea that is indeed compelling, that it is always permissible to do what is best, with a particular consequentialist interpretation of the "best" to which this idea appeals, upon which the "best" in question is taken to be the best outcome. What it is permissible to do, on this interpretation, is to promote the outcome ranked best. But once the Compelling Idea is prized apart from this distinctively consequentialist interpretation of it, the alternative evaluative framework can plausibly be understood as offering an alternative interpretation of the "best" in the Compelling Idea that it is always permissible to do what is best. In particular, it suggests that "best" in this Compelling Idea is properly understood as the best course of action, the course of action decisively supported by good reasons.[38] Each framework, then, can accommodate the Compelling Idea that it is always right to do what is best, but they provide two potentially very different interpretations of the "best" that it is always right to do. One framework maintains that it is always permissible to pursue the best course of action, the course of action decisively supported by good reasons; the other maintains that it is always permissible to bring about the best outcome, the state of affairs that ranks the highest. That it is always at least permissible to do what is best is a very Compelling Idea. But which is more compelling: the interpretation dictated by the consequentialist evaluative framework, or that suggested by the alternative framework? Because the Compelling Idea is typically presented as constituted in part by the consequentialist interpretation of the idea, the alternative interpretation is effectively taken off the table by fiat: the

[38] Hurley, *Beyond Consequentialism*, 113–24.

consequentialist interpretation of the Idea is presented *as* the Idea itself. But the availability of the alternative evaluative framework demonstrates that there is an alternative interpretation of the Compelling Idea that provides no support, in the absence of additional argument, for consequentializing or consequentialism.

Moreover, once the alternative interpretation of the Compelling Idea is clearly in view, it becomes clear that it has considerable intuitive appeal. For instance, although it is not clear what it means to say that some outcome is good-relative-to-me, as opposed to good for me or good overall, it is intuitively clear what it means to say that some course of action is the best of those available to me, and that such an evaluation of actions is relative to each agent. Indeed, the impartial evaluation of actions seems to be fundamentally agent-relative. Consequentialists have long granted that intuitively it seems quite plausible, for example, to maintain that the best course of action for me, the action decisively supported by good reasons, is not to kill another person, even if doing so will result in two others not killing two other people. This seems true even though such a course of action may not bring about a state of affairs that is either better for me or better overall, and regardless of whether there is a clear independent sense in which the state of affairs promoted by such an action is evaluator-relatively best.[39] In short, there is an intuitive sense of good reasons for action, and of the best course of action as that decisively supported by such reasons, upon which the alternative evaluative framework can draw. These intuitive considerations sit comfortably with the idea that the "best" to which the deontic ought is linked in the Compelling Idea is "best" understood as a property of actions. It could, of course, simply be stipulated that the state of affairs that results from performing the best action, the action that the agent has decisively good reasons to perform, is the best state of affairs relative-to-the-agent. But this results in an account upon which such an evaluation of states of affairs as good is parasitic upon the evaluation of actions as good. It would lose the Explanatory Thought, and its interpretation of the Compelling Idea would be parasitic upon, rather than

[39] As Nagel points out in *The View From Nowhere*, in such cases it seems that although "things will be better, what happens will be better…I will have done something worse." (*The View From Nowhere* (Oxford: Oxford University Press, 1986), p. 180)

an alternative to, the interpretation of the Compelling Idea suggested by the alternative evaluative framework.

Perhaps, instead, it is the Explanatory Thought itself that might be thought to support the consequentializing of all alternatives. Here again, however, the alternative evaluative framework provides an interpretation of the Explanatory Thought that provides no support, absent additional argument, for such a consequentializing program. The Explanatory Thought is that there is some form of value-based rationale for claims that agents ought or ought not to perform actions. The evaluator-relative consequentialist framework, insofar as it seeks to preserve the Explanatory Thought, holds that it is the goodness of states of affairs relative to each of us that plays this role in the explanation of deontic evaluations. The suggestion is sometimes that to eschew the consequentialist framework is to eschew any such value-based rationale, hence to lose the Explanatory Thought. But no such difficulty confronts theories articulated within the alternative framework. They can readily be viewed not as eschewing a value-based rationale, but as maintaining that it is through appeal to the goodness of actions and reasons for action that we explain our deontic evaluations of actions. Properly understood, this alternative evaluative framework not only need not deny that we explain the deontic statuses of actions through appeal to goodness, it invites the view that a rationale for deontic evaluation is provided by appeal to the goodness of actions and reasons for action. Why do I believe that Smith ought to perform some particular action? Because it is the best course of action among those available to Smith, the action decisively supported by good reasons. I explain ought judgments, for example, by providing the decisively good reasons that favour the actions in question.

This is a very thin notion of a value-based rationale.[40] But in this context such thinness is a virtue rather than a vice. The Explanatory Thought is introduced as virtually a platitude, a claim that advocates of competing theories, whether consequentialist or non-consequentialist, will find powerfully intuitive. The rationale provided by the alternative framework is sufficiently robust to accommodate this thought, but sufficiently thin to capture something that each recognizes as intuitively

[40] This point, and many of its implications, was first made clear to me in conversation by Sarah Stroud.

compelling—that we explain claims about the deontic status of actions through appeal to the goodness of such actions and the reasons for performing them. What is intuitive about the Thought is precisely that in order to determine what I ought to do, I should determine which course of action, if any, is decisively supported by the weight of good reasons, hence is the best course of action. It is just such an explanation of deontic judgments through appeal to good and bad actions and reasons for action that is suggested by the alternative evaluative framework.[41]

As with the analysis of the Compelling Idea, this analysis of the Explanatory Thought suggests not only *that* the possibility of consequentializing candidate theories does not support consequentialism, but *why* it has mistakenly been thought to do so. The suggestion is that the advocate of the consequentializing argument for consequentialism has run together an attractive and plausible Explanatory Thought with a consequentialist interpretation of that thought. This creates the impression that to reject this consequentialist interpretation is to reject the Explanatory Thought itself. But once it becomes clear that the alternative evaluative framework offers an alternative interpretation of the Explanatory Thought, appeal to the Thought does not discriminate in favour of a consequentializing approach over its deontologizing counterpart.

The availability of the alternative framework suggests that much of the apparent appeal of the consequentializing argument for consequentialism comes from the assumption that only the consequentialist framework can capture the Compelling Idea and/or the Explanatory Thought. It suggests; moreover, that such an assumption is misguided—that there

[41] Within the context of this alternative framework, consequentialist and non-consequentialist theories can then be understood as offering rival thicker accounts of what these good reasons and actions are. For the consequentialist, good actions are determined entirely through appeal to good states of affairs, and good reasons to act are all fundamentally reasons to promote good states of affairs. Non-consequentialist theories reject such state-centered accounts of reasons and value, offering various alternative thick accounts of value that yield non-consequentialist accounts of these good reasons and actions. Thus Niko Kolodny, for example, follows Scanlon in advocating an account of value upon which "Things of value can provide us with reasons when we stand in relations to them other than being able to bring them about...and the reasons that they provide us with may be to do things other than to bring them about—such as to honor and respect them in suitable ways" ("Aims as Reasons," in *Reasons and Recognition*, ed. R. J. Wallace, R. Kumar, and S. Freeman (New York: Oxford University Press, 2011), p. 68). I will later consider briefly one such thick non-consequentialist account of good reasons and actions: the Kantian alternative as developed by Barbara Herman.

is an alternative evaluative framework that reflects an alternative interpretation of the Compelling Idea and the Explanatory Thought. I have suggested that this alternative framework has strong intuitive appeal along certain dimensions. In the absence of arguments for the adoption of the consequentialist framework either instead of or in addition to this alternative, the case for consequentializing all candidate theories, for articulating them within the context of this contested framework, simply begs the question against alternative theories articulated within this alternative framework.

My suggestion is that the apparent force of the consequentializing proposal arises from the implicit but mistaken assumption that the only plausible framework within which to articulate plausible theories of value is the consequentialist's. Such an assumption, for example, appears to structure Philip Pettit's attempt to mediate the conflict between consequentialists and their critics. He takes the conflict to be between two different ways of responding to goodness as the consequentialist framework understands it—as fundamentally a property of states of affairs: the consequentialist "maximizes its expected realization," while the non-consequentialist does what "would promote the value in a world, roughly, where others were equally compliant."[42] Each, on this view, takes the value central to the Explanatory Thought to be value of outcomes rather than actions, and attempts to unpack the role of this evaluation of outcomes in a different way. With the alternative evaluative framework in view, however, the contrast is not between what does promote and what would promote the value of outcomes, but between promotion of better and worse states of affairs and performance of better and worse actions—between the promotion of the best outcome and the performance of the best action, the action decisively supported by good reasons.

Pettit presupposes the consequentialist framework in the very articulation of purportedly deontological alternatives. The claim that an agent is obligated to perform the best action is transmuted into the implausible claim that the agent is obligated to promote the best outcome, but in a fastidious, dirty-hands-avoiding, counterfactual way. What seems fastidious and bizarrely agent-relative within the consequentializing framework—if

[42] Philip Pettit, "The Consequentialist Perspective," *Three Methods of Ethics*, with M. Baron and M. Slote (Oxford: Blackwell, 1997), pp. 127–8.

murdering is a bad thing to happen, why merely insist that each of us keeps our hands clean and minimizes our murdering?—seems appropriately agent-relative and not at all fastidious within the alternative framework.[43] If each agent has decisively good reasons in the relevant cases not to murder, not to steal, and so on, then the best course of action will be to avoid murdering in such cases, hence agents ought to pursue such a non-murderous course of action. If murdering and stealing are fundamentally bad things to do, courses of action that there are impartially good reasons for each person to avoid in her interactions with others, no appearance of arbitrary fastidiousness arises. Each agent has decisively good reasons not to steal, hence each is obligated not to pursue such a bad course of action.[44]

The availability of the alternative evaluative framework also provides a similar response to Michael Smith's consequentializing argument for consequentialism. He maintains that the plausibility of a "Moorean conception of obligation" favours the consequentializing of all plausible alternative theories. On the Moorean conception, "Ax (x ought to φ in certain circumstances C if and only if φ-ing is that action, of the actions that X can perform in C, that produces the most good or the least bad)." Smith argues that "once we buy into a Moorean conception of obligation . . . I see no way of analysing the stringency of an obligation except by way of considering the amount of good that acting on that obligation will produce."[45]

The alternative, he suggests, is to "look at the obligations themselves."[46] But once again this is to assume that there is no plausible alternative evaluative framework that can provide an alternative interpretation of the Compelling Idea and the Explanatory Thought. On such an assumption, the non-consequentialist can only be understood as putting forward constraints upon promoting valuable outcomes via

[43] I take this point from Mark Schroeder, "Ought, Agents, and Actions," *Philosophical Review* 120 (2011), 1–41.

[44] For development of this line of thought, see Hurley, *Beyond Consequentialism*, chs. 6 and 7.

[45] "Neutral and Relative Value After Moore," p. 587. See also Smith, "Two Kinds of Consequentialism," in which his commitment to the consequentialist evaluative framework leads him to understand a side-constraint such as a right as "a rational dictate telling each agent to aim at his own non-violation of anyone's rights." Although Smith takes such an account of side-constraints to be "natural," it seems far less natural than that available within the alternative evaluative framework, upon which a side-constraint is a "rational dictate" telling each agent that there are good and typically decisive reasons not to violate the rights of any other person.

[46] "Neutral and Relative Value After Moore," p. 587.

appeal to "obligations themselves" that are independent of any evaluative framework. The implication is that the only alternative to the Moorean theory of value and obligation flies in the face of both the Compelling Idea and the Explanatory Thought. But the alternative evaluative framework can plausibly be understood as appealing to an alternative to the Moorean theory of value and obligation, an alternative interpretation of the Compelling Idea that it is always permissible to do what is best.[47] Playing upon Smith's formulation of a Moorean theory of value and obligation articulated within the consequentialist evaluative framework, the following formulation can be offered for an alternative, "Kantian" conception of value and obligation that is articulated within the alternative evaluative framework: Ax (x ought to φ in certain circumstances C if and only if φ-ing is that action, of the actions that X can perform in C, that is best, the action that X has decisively good reasons to perform in C).

The "Moorean" maintains that agents ought to promote the best outcome; the "Kantian" maintains that agents ought to perform the best action. To accept the former is to have strong reasons to insist that any plausible theory must have a consequentialized form, but acceptance of the latter provides no grounds for preferring a consequentialist moral theory to any other non-consequentialist alternative.

Many non-consequentialist moral theories can plausibly be understood as deploying this alternative evaluative framework. For example, on one increasingly common interpretation of Kant,[48] defended forcefully by Barbara Herman, Kant is best understood as explicitly articulating his account within such an alternative evaluative framework. On Herman's reading Kant shares with advocates of the consequentialist framework the view that it is "implausible to suppose that a moral theory could persuasively do its work without a grounding concept of value."[49] However, the "domain of the 'good'" for Kant is not states of affairs and events, but "rational activity and agency."[50] On such an account

[47] For a recent deontologist who is plausibly read as rejecting both the Compelling Idea and the Explanatory Thought, however, see Andrew Schroeder, "You Don't Have to Do What's Best!," Mark Timmons (ed.) *Oxford Studies in Normative Ethics, Vol. 1* (New York: Oxford University Press, 2011), pp. 166–201.

[48] Michael Ridge, "Consequentialist Kantianism," p. 421. He offers a laundry list of Kant commentators who read Kant as defending "a distinctive kind of value," not merely as arguing for constraints on the promotion of consequentialist value.

[49] Herman, "Leaving Deontology Behind," in *The Practice of Moral Judgment* (Cambridge: Harvard University Press, 1993), p. 209.

[50] Herman, "Leaving Deontology Behind," p. 213.

of the good, "each agent, insofar as she is rational, acts in ways she takes to be (in some sense) good. She acts with and from the belief that her choices and reasons for choosing *are* good."[51] By contrast, "objects and events are…judged good just in case the determination to act for them is good." In short, Kant's moral theory, as Herman interprets it, is explicitly articulated within the alternative evaluative framework, and explicitly eschews a consequentialist interpretation of such a framework.

Such a Kantian account can readily accommodate the Compelling Idea. The best action is the action decisively supported by good reasons; moreover, any such decisively justified action, on such a Kantian account of reason, is at the very least both rationally and morally permitted. It also appears to accommodate the Explanatory Thought, and to do so at two different levels. First, it is the Kantian theory of value, Herman suggests, that provides a "rationale for moral constraint" generally. Without such a value-based rationale, deontological constraints would remain a "mystery."[52] Second, such recourse to value judgments is necessary "to support full deliberative judgment" in particular cases. When confronted with competing moral considerations, a theory of value is necessary "to make the reasoned comparative judgments necessary for deliberation."[53] Why should I do this rather than that, given that both are supported by conflicting moral considerations? An effective answer will demonstrate that one action rather than the other is the best available course of action, the course of action decisively supported by good reasons.

On Herman's interpretation, Kant's moral theory does not eschew an evaluative framework; rather, it deploys a version of the alternative evaluative framework. For those who accept such an alternative evaluative framework with the corresponding "Kantian" theory of obligation, whether or not they are, like Herman, actual Kantians, appeals to better and worse outcomes are relevant only insofar as they reflect good reasons for acting, and the action that an agent ought to perform is the best course of action, the action decisively supported by good reasons.[54]

[51] Herman, "Leaving Deontology Behind," p. 214.

[52] Herman, "Leaving Deontology Behind," p. 210.

[53] Herman, "Leaving Deontology Behind," p. 211.

[54] Korsgaard as well offers an alternative to the consequentialist account of goodness upon which "actions, acts-for-the-sake-of-ends, are both the objects of choice and the bearers of moral value"; "Acting for a Reason," in *The Constitution of Agency* (Oxford: Oxford University Press, 2008), p. 219.

6.6 THE DEBATE GOING FORWARD

I have demonstrated that it is simply mistaken to take the consequentialist evaluative framework, and with it the case for consequentializing all relevant values, to be dictated by the Compelling Idea and/or the Explanatory Thought. The alternative evaluative framework, with its "Kantian" conception of value and obligation, can readily accommodate both this Idea and this Thought, and versions of most non-consequentialist theories can readily be articulated within this framework. That such arguments for consequentialism and consequentializing are mistaken, however, does not rule out the possibility that the case can be made on other grounds. Consequentialists can argue against the alternative evaluative framework, rejecting the claim that agents ought to do what it is best to do. A strategy that is perhaps both more illuminating and more initially promising, however, is to embrace this alternative evaluative framework, and deploy an argument for consequentialism within the context of such a framework.

Such a consequentialist approach would accept, along with many Kantians, Aristoteleans, and others, the claim that agents ought to do what it is best to do, what they have decisively good reasons to do. But it would attempt to make the case, *pace* these other approaches, that what agents have decisively good reasons to do is always to promote states of affairs (typically evaluated as better or worse). On such an approach, consequentialists agree that agents ought to do what it is best to do. Like many of their opponents, they are offering theories within an evaluative framework that relates the deontic status of actions to the goodness of actions. But this consequentialist argues, in addition, for a particular interpretation of what it is best to do upon which what agents have decisively good reasons to do is to promote the highest-ranked state of affairs. On this approach the argument for consequentialism, and for consequentializing all relevant values, is an argument within the context of the alternative evaluative framework, an argument for one rival account of the evaluation of actions as better or worse.

The central disagreement between consequentialists and their critics, thus understood, concerns the relative merits of competing accounts of what it is best to do—of what agents have decisively good reasons to do. Many of the central debates in ethics and metaethics bear directly upon this question; indeed, the resolution of such central debates concerning

1) the nature of normative "ought" claims, 2) the nature of practical reasons, and 3) the nature of desire, seems likely to pave the way for a resolution of the central disagreement between such consequentialists and their deontological and other critics. In what follows, I will briefly indicate the potential relevance of each of these debates to this central disagreement.

Some normative oughts appear to be fundamentally "oughts to do" rather than "oughts to be," but arguments have been made that all "oughts" that appear to relate agents to actions, all "oughts to do," are fundamentally "oughts to be" evaluating propositions that "would obtain were things to be ideal."[55] If it is plausible, as Mark Schroeder suggests, "to suppose that whatever the deliberative 'ought' relates agents to is the same sort of thing as whatever reasons relate agents to," then if all "oughts" are fundamentally "oughts to be" that relate agents to propositions, all reasons will ultimately count "in favour of propositions," [56] of states of affairs to be promoted. Such a result would considerably strengthen the case for a consequentialist interpretation of what it is best to do within the alternative evaluative framework. If, by contrast, some "oughts" are, as Schroeder argues, fundamentally "oughts to do" that relate agents to actions, this suggests that some reasons will fundamentally count in favour of actions rather than propositional contents (to be promoted), a result that would favour non-consequentialist interpretations of the alternative evaluative framework.

Similarly, if a case can be made that all practical reasons are, as Nagel suggests in *The Possibility of Altruism*, fundamentally reasons "to promote A," where A is an event or state of affairs that the reason counts in favour of,[57] then it is natural to view all reasons as fundamentally facts that are relevant to the ranking of events/states of affairs, typically as better or worse. This account of reasons would appear to count strongly in favour of a consequentialist interpretation of the alternative evaluative framework. If, by contrast, at least some reasons are fundamentally reasons to perform A, where A is an action that the reason counts in favour of, such an account of reasons would appear to favour

[55] Schroeder, "Ought, Agents, and Actions," p. 2.

[56] Schroeder, "Ought, Agents, and Actions," p. 36.

[57] Nagel, *The Possibility of Altruism*, p. 47.

non-consequentialist interpretations of the alternative framework.[58] Again, if all desires are properly characterized fundamentally as pro-attitudes towards propositional contents that the desiring agent is disposed to promote, and as judgment-sensitive, the judgments to which they are sensitive would seem plausibly to be judgments concerning the states of affairs that make up the content of such propositions.[59] Such an account of desires would seem to provide support for a consequentialist interpretation of the alternative framework. If, by contrast, many desires are properly understood fundamentally as attitudes towards actions to be performed, and as judgment-sensitive, the judgments to which they are sensitive would seem plausibly to be judgments concerning the actions that are the content of these attitudes.[60] Such an account of desire would seem to provide support for non-consequentialist interpretations of the alternative evaluative framework.

The alternative evaluative framework suggests that agents ought to do what it is best to do. Standard arguments for consequentializing all candidate moral theories are profitably understood as arguments for a particular interpretation of what it is best to do, upon which it is best to do what promotes the best state of affairs. But appeals to the Compelling Idea and the Explanatory Thought, we have seen, fail to privilege this consequentialist interpretation over its non-consequentialist rivals. The most promising strategy for consequentialists is to adopt explicitly the alternative evaluative framework, and to argue for a consequentialist account of what it is best to do as what promotes the highest ranked (typically, the best) state of affairs. It may be thought that arguments for such a consequentialist interpretation of the alternative evaluative framework can be grounded in accounts of the nature of evaluative oughts (as all fundamentally "oughts to be"), practical reasons (as all fundamentally counting in favour of events or states of affairs), and/or desires (as all fundamentally having states of affairs as their contents).

[58] See, for example, Portmore's arguments that all practical reasons are teleological reasons in chapter 3 of *Commonsense Consequentialism*, and arguments by Scanlon and Korsgaard against such an account of practical reasons in *What We Owe to Each Other* and "Acting for a Reason" respectively.

[59] See, for example, Parfit's account of desire in *On What Matters*, ch. 1.

[60] Such an account is sketched in Hurley, *Beyond Consequentialism*, ch. 8, and defended in Michael Thompson, *Life and Action* (Cambridge: Harvard University Press, 2008), and Talbot Brewer, *The Retrieval of Ethics* (Oxford: Oxford University Press, 2009).

But many philosophers (myself included) reject these accounts of the evaluative ought, practical reason, and the nature of desire. They offer rival accounts that would appear to be better suited instead to non-consequentialist interpretations of what it is best to do—what agents have decisively good reasons to do. Within the context of these alternative interpretations, a requirement to consider all theories in their consequentialized form would be perverse and counter-productive. My suggestion, then, is that it will be productive going forward to focus on this question of whether a consequentialist or non-consequentialist interpretation of what it is best to do is more plausible. In part, this is to ask which interpretation can be supported by a more plausible account of the evaluative ought, practical reason, and the nature of desire.

7

On Criminal and Moral Responsibility

DAVID SHOEMAKER

My aim in this essay is to investigate the relation between criminal and moral responsibility. Among those few theorists who talk explicitly about it, very little is actually said either to explain or to defend what they believe the relation consists in. There are not many views on the table, and they are fairly similar in the end. Some simply take for granted that the two types of responsibility are intimately related, with the former entailing the latter: "The practice of holding an agent criminally responsible for breaching the criminal law is a specific instance of the more general practice of holding agents responsible for what they do," the latter of which is called "the ordinary moral practice" (Tadros 2005, 23). Others describe the entailment relation in normative terms: "It is often thought that criminal responsibility should track moral responsibility: I should be held criminally responsible for V's death, and liable to conviction for criminal homicide, only if I can be held morally responsible for V's death" (Duff 2009a, 978). Still others acknowledge that, while there are some differences between criminal and moral responsibility in practice, these are just a function of there being different criteria specific to the content of the criminal and moral *law* (Hart 2008, 226), and anyway, when it comes right down to it, at least "the doctrines that excuse or mitigate criminal responsibility . . . closely track the variables commonly thought to create moral excuse or mitigation" (Morse 2008, 208).

There are two related tenets revealed here that together make up what I will call the *Standard View*, assumptions that seem to be held by most Anglo-American legal and moral theorists.[1] The first tenet is that moral

[1] Those working in the Continental tradition seem to have a very different take on matters. For explanation and development of the foundations of the Continental tradition, with the Anglo-American tradition as its foil, see Brudner 2009. See also Husak 2010 for helpful discussion of the contrasting traditions.

responsibility (MR) is a necessary condition of criminal responsibility (CR) normatively understood, that for me to be genuinely or legitimately criminally responsible for Φ, I must be morally responsible for Φ. Why is this? The connective tissue is most often thought to be desert: the criminal law imposes sanctions on people—punishment which must be deserved to be morally justified. But one deserves punishment only if one is "a responsible agent who performs a culpable wrong prohibited by law in the absence of justification or excuse" (Husak 2010, 842). The second, related, tenet is that essential elements of CR are structural and functional analogs of essential elements of MR. More specifically, as Douglas Husak puts it, the "legal analyses of responsibility, desert, wrongdoing, justification, and excuse are closely related to their counterparts in the domain of morality" (Husak 2010, 842).

In this essay I will show why both tenets of the Standard View are false, or at the very least quite misleading. The relation between CR and MR is actually far more complicated—and interesting—than the Standard View would have us believe.

7.1 WHAT WE ARE TALKING ABOUT WHEN WE TALK ABOUT RESPONSIBILITY

On one plausible formulation, to be criminally responsible is for one's conduct to be "of the kind that warrants conviction in a criminal court" (Tadros 2005, 1). CR is thus a matter of being worthy of a particular sort of response specific to the criminal justice system. In line with the ambiguity of "conviction," we may distinguish between two sorts of relevant response: (a) judgments of guilt, and (b) their associated sentences. This distinction corresponds to a distinction between (a) being judged *to be* criminally responsible, and (b) being *held* criminally responsible. Very generally, then, to be criminally responsible for some action is to be worthy of at least one such response.

There is also rough agreement on the general concept of MR: to be morally responsible is for one's conduct to be worthy of a certain sort of response particular to our moral practices (Eshleman 2009, and Fischer 1986, 12). To remain as ecumenical as possible, we may say here too that the sorts of response relevant to the practices include both judgments and holdings. To be morally responsible for some action, then, is to be the worthy target of at least one of the following responses: (a) being judged to be morally responsible, or (b) being held morally responsible via blame or praise (in the most general terms).

The analogy between the concepts of CR and MR suggests a very general umbrella concept of capital-R Responsibility: to be Responsible (*simpliciter*) for Φ is to be the worthy target of certain sorts of responses, responses specified by the normative domain(s) (criminal, moral, or other) in which Φ occurs. One way to put our overall questions, then, is this: (a) does the criminal instantiation of Responsibility entail its moral instantiation, and (b) is the difference in the content of their respective normative domains their only significant difference? My answers will be no, and no.

7.2 THREE CONCEPTIONS OF RESPONSIBILITY

The general concept of Responsibility has been put in inclusive terms deliberately. But while many MR theorists think there is only one true conception of this concept (disagreeing, however, about the details), I believe that our responsibility practices actually track three different conceptions. I have argued for this pluralistic view in great detail elsewhere,[2] so I will not revisit those arguments here. But given that the remainder of the essay is structured on top of those distinct conceptions, I need to say something briefly to explain and motivate each. I will focus solely on responsibility for *actions*, but that will just be shorthand for the variety of things we may be responsible for, including, of course, attitudes.

The first conception is *attributability*. For me to be attributability-responsible for Φ is for Φ to be properly attributable to me for purposes of appraisal (moral or criminal, depending on the normative domain).[3] This conception is about one's volitional *structure*: what it means for Φ to be properly attributable to me is just that it flows from or expresses the ends, commitments, or cares that generally make up my character as a practical agent.[4] The fact that I am tall, hirsute, or good-looking is thus not attributable to me in this sense because such features are unconnected to any of my psychological characteristics. But not all features

[2] See Shoemaker 2011b, and Shoemaker Forthcoming. For a two-conception forerunner, see Watson 2004, 260–88.

[3] See, for example, Scanlon 1998, chapter 6; and Smith 2005, esp. 238.

[4] See, for example, Tadros 2005, 31; and Watson 2004, 264–71.

that do flow from someone's psychological elements are attributable to that person *qua* practical agent, given that they may be unconnected to her ends, cares, or commitments (sometimes called her "deep" self; see Wolf 1987 for this terminology). This seems true of those who suffer from Obsessive Compulsive Disorder or Tourette's syndrome, for instance.[5]

The second conception of responsibility is *answerability*. This conception captures our different aim to assess agents' *reasons* for action—reasons they took to justify those actions. An answerability demand is a demand for explanation (in the sense of justification): "Why did you do that?" Here we are often measuring what the agent took to be his justifying reasons against some standard or other. Notice that the standard, and the action, may be positive or negative: I may ask you the answerability question with a tinge of admiration or a tinge of accusation. You are answerability-responsible for some action, then, just in case you can, in principle, answer that question; that is, when you judged some reasons to be more worth acting on than others.

The third conception of responsibility is *accountability*. This is a response-dependent form of responsibility: I am accountability-responsible for Φ just in case I am susceptible to being appropriately *held to account* for Φ.[6] To be held to account for Φ often involves sanctions or rewards for Φ, so considerations of fairness are thought to apply to these appraisals in a way they do not for attributability or answerability appraisals. And it could not be fair to blame or punish someone, it is often thought, unless the target understands, and could have avoided, such responses. Accountability thus seems to implicate capacities that answerability and attributability may not.

I will develop each of these barebones sketches of the three conceptions in what follows, but for now their content and differences may be brought out more clearly by thinking about how they are typically instantiated in the criminal justice system. The prosecutor's burden is to establish *attributability*, that the offense with which the defendant has been charged is properly attributable to him *qua* practical agent.[7] This

[5] Scanlon 2002 is a rare holdout to this view, however.

[6] This is one plausible way to interpret the famous view of P. F. Strawson 2003, taking him both to be talking solely about the accountability conception of responsibility and to be maintaining that being accountable is a function of being *held* accountable.

[7] See Tadros 2005, chapter 1, for this way of putting the matter.

ordinarily involves establishing the defendant's *mens rea* with respect to the crime; that is, establishing whether the action he performed depended in the right way on his mental states—most often his knowledge and intentions—such that it counts as the charged offense. Once the burden is met, the defendant is invited to *answer* for his actions by offering a defense. In this stage, defendants are expected to cite their evaluative reasons in acting. A successful defense is standardly constituted by either a justification or an excuse. A defendant is justified in doing what he did when his reasons for doing so were valid; that is, when what he judged was a good reason for performing the action in seeming violation of the law, all things considered, really *was* such a reason. A defendant is excused from doing what he did when he was justified in *thinking* that his reasons were valid, despite the fact that they were not (Gardner 2007, for example, 109–10). An agent is *answerable*, then, when he is capable, at least in principle, of citing such reasons. Agents exempt from this sort of responsibility include those who are insane or cognitively disabled. Finally, defendants to whom some criminal offense is attributable, and who are answerable but unsuccessful in mounting a defense, are thereby *accountable* for that offense; that is, liable for punishment in response to the criminal deed. To be accountable, they must have been able to avoid the punishment coming their way for the criminal violation (by, say, having been able to avoid performing the criminal action in the first place).

Attributability, answerability, and accountability are all familiar as well in our interpersonal moral lives, in ways that I will make explicit throughout the essay. Our question, then, is whether the relation between these conceptions of responsibility in the criminal and moral realms is as the Standard View affirms. We can best address this question by examining each conception of responsibility individually.

7.3 ATTRIBUTABILITY: THE CASE OF STRICT LIABILITY

In this section I will explore one reason to doubt the second tenet of the Standard View by exploring a disanalogy between CR and MR with respect to the attributability conception of responsibility, in particular with respect to its relation to accountability. It is a basic assumption of moral theorists that *attributability is necessary for accountability*, that for one appropriately to be held accountable for Φ, Φ must be properly attributable to one, such that it is an expression of one's ends, cares, or commitments (Watson 2004, 263 and 278).[8] The relation between attributability and criminal liability (accountability) seems quite different, though, given the inclusion in the law of a certain familiar and entrenched subset of criminal offenses; namely, *strict liability*.[9]

Strict liability attaches to certain criminal actions performed absent any requirement of *mens rea* at all; that is, performed without intention, knowledge, or foresight, where it is not even true that these conditions should have been met, or would have been met, by a "reasonable man." There are strict liability laws against selling bad milk or tainted meat, selling alcohol to those who are intoxicated, being found drunk on the road (even if the police carried one there), being in possession of drugs or a forged passport, engaging in statutory rape, and more. As should be clear, removing *mens rea* in such cases also removes attributability: if I sold you alcohol but genuinely had no idea that you were intoxicated (and could not have found this out or been reasonably expected to know this given your remarkable skill in concealing it), there is no coherent

[8] In Shoemaker 2009 and Shoemaker Forthcoming, I mark a distinction between thin/shallow and thick/deep attributability to show that, while the latter may not be necessary to moral accountability in whim cases and cases of mildly mentally retarded adults, at least the former is. But even the thin/shallow relation is absent in CR, given what I am about to say in the text regarding strict liability. On a related point, one might think that in Sher 2006 his cases illustrating MR without control or certain types of knowledge actually belie the claim that moral accountability presupposes attributability. Nevertheless, as he puts it explicitly, on the proposal he favours what still obtains in these cases is that the "wrongness of [the agents'] acts and omissions [as well as, presumably, their accountability for them] can be traced to sets of beliefs, desires, and so forth, not all of which are conscious but all of which are *fully their own*" (297; my emphasis).

[9] There is a delicate issue here about whether criminal *negligence* might also be an instance of accountability without attributability. Negligence requires the barest condition of *mens rea*, namely, the capacity (shared by the "reasonable man") for *foresight* of the damage the agent's negligence caused, even if the agent as she was did not want the damage to occur. This sounds as if the negligent agent lacks attributability, for the negligent "action" did not flow from her ends. But described differently (or more fully?), negligence might seem to flow from the agent's ends after all, as it could be construed as reflecting her general lack of concern for certain things. She may, for example, deem certain matters as not being worth her attention. When these matters do result in damages, though, and where she could have judged that they were worth her while after all (that is, where a reasonable man would have so judged), then the negligence could be deemed to be attributable to her. Nevertheless, this is a complicated, and perhaps controversial, view. It might easily be thought instead that certain acts of negligence might meet this reasonable man test without being genuinely attributable to the agent (see, for example, Alexander and Ferzan 2009, ch. 3, for what might be suggestions along this line). To sidestep this controversy, I will focus in the text just on strict liability as it is traditionally conceived. (Thanks to an anonymous referee for flagging this issue.)

sense in which this action *qua* criminal offense—selling alcohol to an intoxicated person—could be said to depend on or express my ends, cares, or commitments *qua* practical agent.[10]

The justification for including these crimes in the pantheon of law is purely pragmatic: if we required *mens rea* in such cases, many guilty people would escape justice, insofar as it would be extremely difficult, if not impossible, for prosecutors to prove the relevant intention or foresight. Consequently, there would be more "Not Guilty" pleas, more acquittals, more cases dropped by prosecutors pre-trial, and so, in the end, the law would be "less effective in deterring the dangerous or harmful conduct at which it is aimed: more people will acquire drugs, more shopkeepers will fail to take adequate precautions to ensure that the food they sell is safe, since they will know that even if they are caught, they have a good chance of avoiding conviction" (Duff 2009a, 983).

H. L. A. Hart puts a common attitude toward this body of law succinctly: "Strict liability is held in some considerable odium by most academic writers and by many judges" (Hart 2008, 176). The primary reason is, to put it in my terms, that punishment for strict liability undermines the presumed entailment relationship between accountability and attributability. There seem to be only two ways to preserve this relationship and keep CR analogous to MR (on this score, at least). First, and most obviously, one might simply get rid of strict liability laws. If, as Duff puts it, "[N]on-culpable ignorance of relevant facts makes blame illegitimate" (Duff 2009a, 982), then blame for strict liability would be illegitimate.[11] But abandoning strict liability would require a radical revision of the criminal law—half of all cases brought to the courts are for strict liability offenses[12] —and it would leave us grossly unprotected with respect to a large variety of activities we deem public wrongs. The pragmatic reasons for preserving strict liability are quite compelling.

The second approach would keep strict liability laws and then just divorce the justification of punishment from responsibility altogether.[13]

[10] I will say more about the nature of the relevant action description in such cases later.

[11] One could of course preserve trials and judgments of guilt for violations of strict liability laws, but without any (legitimate) punishment attached such judgments, criminal justice would be both toothless and pointless.

[12] See <http://sixthformlaw.info/01_modules/mod3a/3_10_principles/16_principles_strict_liability.htm>. If this sounds surprising, keep in mind that possession of certain guns or drugs is a strict liability offense.

[13] Hart 2008, 176–7, lays out this possibility.

But allowing that punishment could be justified without responsibility is not only conceptually jarring but also seems to require adopting an objectionable utilitarian justification in its place, such that "an important principle has been sacrificed to secure a higher measure of conformity and conviction of offenders" (Hart 2008, 20). The majority of legal theorists—retributivists, mostly—would find this second response to be no less odious than the first.

Given these worries, one might simply abandon the attempt to preserve this analogy between CR and MR, admitting that accountability in the criminal law just does not presuppose attributability the way it does in interpersonal morality. But doing so yields a different worry: it could well eliminate *desert* from the criminal landscape, the element taken by most retributivists as the sole justification for punishment itself.[14] The reason is simple: how could I deserve punishment for some action if that action were not truly *mine*? In other words, if some action does not flow from my practical agency, then there would seemingly be nothing to distinguish punishing me for it from punishing me for my being tall, hirsute, or good-looking, none of which could sensibly involve desert. None of the ways of dealing with strict liability, therefore, looks very attractive.

Antony Duff has recently proposed a very interesting way out of this jam. It involves preserving all three desired elements: (a) strict liability in the criminal law, (b) accountability's presupposition of attributability, and (c) the presumed conceptual connection between being liable to punishment and being responsible generally. The key is to distinguish between what Duff calls "answerability" and what he calls "liability." Unfortunately, he uses the term "answerability" in a way that is different from my usage, so in order to avoid confusion I will try to explicate his version as carefully as possible and then refer to all his mentions of the term as "answerability(D)."

On Duff's usage, for me to be *answerable(D)* for some untoward behavior Φ in general is for me to be the target of an accusation that Φ is attributable to me ("as its author"; Duff 2009a, 980) and to be subject to a demand to answer for Φ. For me to be *liable* for Φ is for me to be answerable(D) for Φ, lack sufficient exculpation (justification

[14] See, for example, Moore 1997, 87. For Moore, retributivism just *is* the view that "punishment is justified by the desert of the offender."

or excuse) for Φ, and consequently be the target of blame for Φ (so Duff's "liability" is more or less my "accountability"). This structure is mirrored in a criminal trial, where the burden is on the prosecution to prove both *actus reus* (guilty act) and *mens rea* (guilty mind), and such proof establishes answerability(D): it involves attributing the action/event to the defendant and then demanding that he answer for it. The defendant then attempts to provide a defense, offering either justification or excuse, and if these are insufficient he is liable to blame; that is, punished (Duff 2009a, 980–1).

Given this important distinction, Duff suggests we transform strict liability into strict *answerability*(D). As things currently stand with ordinary laws, in a criminal trial we have non-strict answerability(D): to meet her burden the prosecutor must prove both *actus reus* and *mens rea* before defenses are necessary. But one might instead require (and some laws already do require) the defendant to answer for his deeds after only *actus reus* has been established; that is, after it has been established only *that he voluntarily did the deed*, with the burden of proof then shifting to the defendant to show, as part of his attempt at exculpation, that he lacked the relevant *mens rea*. Strict answerability(D) of this sort could thus be built into the relevant criminal offenses, which would themselves be put in strict terms but which would also include a description of possible *defenses* for the accused, for example, to "prove that he neither knew of nor suspected nor had reason to suspect" that what he was doing or had in his possession was illegal (Duff 2009a, 983; the quote is from the (UK) Misuse of Drugs Act 1971, s. 28).

Bolstering the plausibility of this move, suggests Duff, is that it would actually make CR *more closely* track MR, insofar as MR is allegedly already strict in just this way. When you knock over my vase, for example, I begin only with the assumption that you indeed knocked over the vase, and I then demand an answer from you for it. You then are expected to provide exculpatory reasons for your action, among which may be that you did not intend it, or did not know (or non-culpably lacked foresight) that it would happen. In other words, when I make the answerability demand in the moral realm, the burden of proof is not on me to prove *mens rea*; the burden to *deny* it is on you. So if answerability is already strict in the moral realm, transforming strict liability into strict answerability(D) actually resolves an important *disanalogy* between CR and MR as they are currently structured (Duff 2009a, 983–4).

I do not believe that this sort of move is viable. The reason is that strict answerability(D) is either not strict or not answerability.[15] The problem stems from the ambiguity of "authorship." Again, Duff includes two conditions for answerability(D): "we attribute an action or event to the person *as its author*, and request (or demand) that she answer for it…" (Duff 2009a, 980; my emphasis). What does it mean to be the author of an action, however? On the one hand, it might mean attributability in the sense I have detailed. If so, then it requires *mens rea*: if the action ultimately flowed from, or was expressive of, me *qua* practical agent, then I intentionally brought it to life (in line with my ends, cares, or commitments). But if a determination of whether I have met this condition is required prior to my accuser's demand for an answer, then whatever sort of liability is involved from that point on, it cannot be strict.

Clearly, then, Duff has in mind a different, looser sense of "authorship." But if my being the author of some action does not require attributability in the above sense, what could make it mine? Duff tells us only that it must be the case that "the defendant's conduct caused a harm or evil of a kind that concerns the criminal law" (Duff 2009b, 308), and his doing so included a "voluntary act" (Duff 2009b, 303). So suppose the *actus reus* is selling tainted meat (currently a strict liability offense). Suppose further that I did this unknowingly and without any way of knowing. Presumably, my voluntary—uncoerced—act was "selling meat," but insofar as the meat was in fact tainted, I have caused a harm or evil that concerns the criminal law, despite lacking any intention to sell *tainted* meat. The problem with making this move, however, is that it abandons any plausible notion of answerability. Let me explain.

For Duff, I am answerable(D) to the extent that I may "properly be called to answer or to account for" (Duff 2009b, 298) some result. How should we understand the "properly" here, though? There is a sense, after all, in which anyone may be called to answer for anything. I could even call you to account for the crimes of Hitler. And your response, "I did not do it," *is* an answer of sorts to my query. Does this make you answerable(D)—responsible—for those crimes? Duff admits that

[15] I am grateful to Nicholas Sars and David Sobel for discussion of some of the following points.

it does not.[16] But why not? He does not say, but the most plausible explanation is that "I did not do it" is not an answer of the right *kind*, for it fails to implicate any of the agent's justifying reasons. The relevant answerability demand here is "Why did you Φ? Explain yourself!" This is a demand to hear the reasons the agent took to justify her Φ-ing. "I *did not* Φ" rejects the premise of the question, though, so it obviates the need to make reference to any of her reasons at all: it "answers" the question only by saying that there just is no answer of the sort being sought.

Duff insists, though, that answerability(D) can be strict, and a voluntary action directly causally implicated in the risk or production of some harm or evil in the absence of *mens rea* falls under that rubric. But is mere causal responsibility without *mens rea* sufficient for any kind of answerability? It is not part of moral answerability, despite Duff's assumption otherwise. If I knock over your vase because I had no idea (and could have had no idea) that you had just moved it by the door I was about to go through, your accusatory "Why did you do that?" will be answered by my horrified "Oh no, I did not even know it was there!" But this "answer" serves a similar function as saying "I did not do it;" namely, it *undercuts* the question by rejecting one of its premises. In the case of "I did not do it" I cannot cite any reasons for the action because I had none relevant to it. In the case of "I did not know" I cannot cite the right *sort* of reasons for the action because I had none relevant to *them*. I am no more morally answerable in the accident case than I am in the non-identity case.

To explain, in moral answerability there is an implicit completion of the accusatory question "Why did you Φ?" which is rarely made explicit but is nevertheless crucial: "Why did you Φ *in light of the reason(s) not to Φ*?" My having a genuine answer to *this* question depends on my having access to those reasons for not Φ-ing. If I did not know you had placed the vase outside the door I was about to go through, I lacked access to that reason for not opening the door, so despite the fact that I did have a reason for voluntarily opening the door, I had no reason to *refrain* from doing so. Similarly, even if I can give you an answer for why I sold meat voluntarily, I cannot give you an answer for why I sold meat voluntarily *in light of the reason not to do so* (that is, it was tainted), because from my

[16] He states that it would be a "contrast" case to the relevant form of answerability if "I avert blame by denying responsibility[, saying] it was Jones, not I," who did the deed (Duff 2009a, 980).

position at the time I had access to no such reason; that reason certainly shone no light on me. But in saying that I had no access to reasons not to sell the meat, I have only "answered" the question by saying that there just is no answer of the sort being sought. If the relevant reasons could not have been part of my deliberative set, I am as incapable of answering your charge as I was when I just did not do the thing in question. If Duff disallows "It was not me" from counting as an answerability(D) answer in virtue of the agent's inability to cite any relevant justifying reasons in his defense, then it would be arbitrary not to disallow "I did not know" from counting as an answerability(D) answer in virtue of the agent's inability to cite any relevant justifying reasons in his defense. If so, then there could be no strict *answerability*(D), for the strictness is precisely what would render the accused agent *not* answerable.[17]

To recap, if Duff's strict answerability(D) in the criminal realm is to preserve the analogy with MR, it must incorporate attributability, which renders it not strict after all; if it remains strict, then it is just not answerability of the type found in the moral realm. More generally, then, either we include strict liability in our account of CR or we do not. If we do not, then, while we may preserve the entailment relation between accountability and attributability in both CR and MR, we must make radical revisions to current legal practice. If we do include strict liability, and we still want to preserve the entailment relation, then we must either show how strictness—lack of *mens rea*—is somehow compatible with attributability (authorship) in CR, or deny that attributability is necessary for accountability in either CR *or MR*. There are serious difficulties attached to both horns, though—so much so that perhaps just giving up the entailment relation is the least implausible option, though I take no stand on that here. My primary aim has just been to show that the so-called analogy between CR and MR on this point is far

[17] Here is another way to put the point. Duff suggests that "I did not do it" excuses one from *responsibility* (answerability(D)), whereas "I did not know" merely excuses one from *blame* (while preserving one's responsibility/answerability(D)). Perhaps we could allow him merely to stipulate this distinction? (I thank Drew Schroeder for this suggestion.) My point has been that he could do so only at the cost of rendering *answerability* disanalogous between CR and MR. In MR, the point of the answerability demand is to get at the target's reasons for Φ-ing *as opposed to not Φ-ing*. If this is removed as the point of the answerability demand in CR, then while the entailment relation between accountability and attributability could remain analogous in CR and MR, it could do so only at the cost of making answerability different in each realm.

more complicated and tenuous than the Standard View would have us believe, and that there would be serious costs and difficulties in trying to maintain it, including the creation of different disanalogies elsewhere.

7.4 answerability: the "reasons gap" and good will hunting

Being answerable consists in being able in principle to cite the reasons one took to justify one's actions. Call these one's *motivating reasons*. The difference between CR and MR that I now want to explore is the function played by motivating reasons for answerability in each realm. Here again, the Standard View's so-called analogy between CR and MR is much more complicated than it would have us believe.

What is criminal answerability all about? Victor Tadros articulates a plausible, widely accepted view, according to which what is expected from an answerability demand is a rational explanation of the defendant's actions, which involves citation of his motivating reasons, and these reasons are then compared to the *normative reasons*; that is, the reasons there actually were in his circumstances. Where, as Tadros puts it, "normative and motivating reasons come apart, it is often appropriate to react to the agent negatively…" (Tadros 2005, 30). In other words, being punishment-worthy is a function of a *gap* between the reasons the agent took himself to have in acting and the reasons there actually were.

But to what end? The location for the citing of motivating reasons in a criminal trial is in the defense phase, where they are relevant to establish whether the defendant was justified or had an excuse. Offering them is necessary only to the extent the prosecution has already established that the criminal action is indeed attributable to the defendant. If his motivating reasons were good enough—revealing no gap with the normative reasons—then he is justified in what he did and may get off the hook. If he had good reason to *think* his motivating reasons were good enough, even if they actually were not, he has an excuse and may get off the hook. However, if his motivating reasons fall into neither category, then, in light of the gap between his motivating and normative reasons, his answers to the answerability demand are deficient, and he is therefore punishment-worthy. In other words, people are punished for *crimes*, when criminal actions are properly attributable to them without justification or excuse. Their motivating reasons matter only insofar as they might establish one of the latter.

This is not the case for moral answerability, however. In interpersonal morality, the point of making the answerability demand is not just to hear the target's motivating reasons in order to compare them to normative reasons for purposes of excuse or justification; it is instead (or in addition) to determine from them the *quality of the target's will*.[18] While there are different interpretations of "quality of will" in the literature (see Shoemaker Forthcoming), the basic idea on which they all agree is that motivating reasons reveal the actual target of our responses to those who are morally answerable; namely, "the good or ill will or indifference of others towards us, as displayed in *their* attitudes and actions" (Strawson 2003, 80; emphasis in original). Scanlon puts this idea in terms of the *meaning* of the target's action, which is "about what [the action] shows about his attitude toward me…and about the significance of that attitude for our relationship" (Scanlon 2008, 129). On this understanding, then, the moral answerability demand is ultimately a probe to find out what the agent *meant* by doing what he did, a probe for his quality of will, and the reasons he offers are either constitutive of, or evidence for, that quality. We morally blame and praise others for their quality of will as revealed by their motivating reasons. By contrast, we criminally punish others for their *actions*, given that their motivating reasons do not establish a justification or excuse. In the law, the only time we respond directly to the quality of one's will as revealed by one's motivating reasons is when one's particularly nasty attitude provides aggravating circumstances to an already established criminal action. But even then, the crime for which one is punished remains distinct from one's motivating reasons in doing it, the latter of which contribute only to the *degree* to which one is punished.[19]

[18] The first official quality of will theorist was Strawson 2003. Other contemporary theorists who fall under this rubric include Scanlon 1988, Wallace 1994, Watson 2004, and McKenna 2012, among others.

[19] What, though, of hate crimes or the varying degrees of murder? Are not the defendant's motivating reasons of primary importance there in determining his punishment-worthiness? (I am grateful to Tim Scanlon for raising this concern.) No. Instead, all that matters in these cases are the defendant's degree of foresight and the nature of his intentions; that is, *mens rea*. And these factors matter for a determination of whether the action that is attributable to him counts as the *criminal* action for which he is ostensibly answerable. His motivating reasons with respect to these criminal actions (such as the reason he performed what now counts as a hate crime in virtue of his meeting the *mens rea* condition), however, are not relevant to a determination of whether he is worthy of punishment for that crime. I discuss these ideas in much more detail in Shoemaker 2013.

Nevertheless, are not justifications and excuses of central importance to moral answerability too? And if so, do not the agent's motivating reasons serve the same function at least in *that* regard as they do in criminal answerability? I do not believe so. Consider the function of justifications in moral answerability (and from here on I will focus solely on justifications; excuses are more complicated and may not even be theoretically unified, so they deserve a much more expansive treatment than I can provide here; see Tadros 2005, ch. 11). In that realm, justifications function as a plea for the accuser to withdraw any negative response in light of the fact that the accused's quality of will was not ill after all. My default reaction to someone who has taken my car without authorization will be anger and condemnation at what I believe to be his horrible attitude to me, but when the purported thief explains that he did so in order to rush my neighbor to the hospital after her heart attack, my anger will be suspended in light of the fact that, as his justification reveals, he had no ill will to me at all.[20] On the other hand, what makes for an unsuccessful justification is precisely that a case for there having been no ill will cannot successfully be made. Were the purported thief to say that he took my car because he had a craving for a Whopper and so needed to get to Burger King as fast as possible, this surely would not be good enough to displace my judgment of his insufficient regard for me, and it is that perceived lack of regard to which I will respond in holding him responsible.

Contrast this function with justifications in the criminal realm, where there are actually two distinct types of reasons relevant to various offenses, what Tadros calls *content reasons* and *prohibition reasons* (Tadros 2005, 267–73). Content reasons are those about the wrongness of some action that obtain independently of whether it has been rendered criminal; prohibition reasons are those that obtain with respect to its criminalization. So there are reasons against assault independently of whether assault is illegal, but the fact that the criminal law prohibits it provides an additional reason against it. Indeed, prohibitions go beyond content reasons insofar as they get citizens to expect that the prohibited actions will not be performed (Tadros 2005, 268). Defendant's justifications thus must counterbalance *both* content and prohibition reasons. But an unsuccessful justification may therefore not reveal anything

[20] See, for example, Wallace 1994, 127–36; and Watson 2004, pp. 223–4.

whatsoever about the defendant's quality of will simply because it might have involved only the inability to surmount a prohibition reason, and prohibition reasons do not implicate motives (for good reason). As prohibition reasons are irrelevant to moral answerability, by contrast, justifications necessarily serve a different function in that domain.

An example is in order. Suppose that you and I are at a bar talking about my mother when suddenly I slap you. To both your and the state's answerability demand "Why did you do that?!" I respond that you provoked me by questioning my mother's virtue. We may well think this constitutes a sufficient moral justification for the slap. But in court, while the justification of "defending my mother's honor in response to provocation" may well counterbalance the content reason for not slapping someone ("it causes victims minor harm," say), it will probably not be sufficient to counterbalance both that content reason *and* the prohibition reason against assault (that it is illegal and so has thereby created a social expectation that people will not be slapped).[21] Suppose, indeed, that in the criminal case my proposed justification as a defendant fails. What has been shown about my quality of will as a result? *Nothing*. Or at least nothing to which the state could sensibly be responding in punishing me. This is because my justificatory appeal has already revealed my quality of will not to be ill: maternal defense is, to me (and many others), an extension of self-defense, so I was simply acting to protect my mother and meant you no ill will in so reacting (I just wanted you to stop the attack). The fact that this justification may be insufficient to counterbalance the prohibition reason on top, though, reveals nothing *in addition* about the quality of my will, about the attitude I have toward my fellows. At most, it reveals an attitude I have about certain laws, a kind of insufficient regard for their bindingness in circumstances of minor provocation. But insufficient concern for some legal norm *qua* legal norm implies nothing at all about my concern level for those of my fellows ostensibly provided greater security by the existence of that norm, simply because it is only *the existence of the norm* (publicly acknowledged) that has an effect on these particular security interests of my fellows, and not any individual violation of it. The failure of my criminal justification means only that my motive was not good enough to counterbalance the

[21] A more abstract case along these lines is discussed in Tadros 2005, 268.

reasons in favour of the criminal prohibition, and this may be for reasons entirely independent of any (dis)regard for the interests of others.

In sum, to be answerable, generally, is just to have motivating reasons for one's actions that are citable in principle. But those reasons serve different functions in CR and MR. In MR they function to reveal quality of will: the reasons on which you acted (or the reasons you ignored in acting) expose just what your attitude toward me was. It is this attitude for which I am probing in making the answerability demand of you, and it is this attitude to which I respond in blaming or praising you. In CR, on the other hand, motivating reasons are relevant only for determining whether the defendant should get off the hook for an action that the prosecution has *already* established is attributable to him (strict liability aside); that is, the offense for which he is to be punished. This is one important difference between them. The thought, then, was that perhaps where motive *does* matter for criminal answerability (in the defenses of justification and excuse) it at least functions analogously with moral answerability in that location. I have thus further argued that even there (at least with respect to justifications) it does not, because the addition of prohibition reasons in the criminal law means that many (most?) unsuccessful justifications may reveal no ill will or lack of regard toward others at all, whereas what *defines* an unsuccessful justification in the realm of moral answerability is precisely the persistence of ill will.[22]

[22] Jerry Gaus has raised an important point to me that should be addressed here; namely, that the notion of "moral answerability" I have been discussing may not be at all unified in the way I seem to have been assuming it is. For instance, we often become (morally) angry at strangers who do immoral things (such as tossing trash out of their cars onto the road), and it may seem strained, if not false, to think that our answerability demands of them are probing for quality of will, or that any justifications they give are meant to reveal said quality. I am not so sure, however, as their reasons typically do reveal a lack of regard for the interests of others, and that sounds like an objectionable quality of will. Nevertheless, we may have reason to think Gaus is correct. After all, the analysis I am giving with regard to moral answerability is explicitly modeled on moral reactions between those who are in more "close-up," interpersonal relationships, and despite Scanlon's optimism that there is a "moral relationship" in which we all (strangers and intimates alike) stand to one another that is analogous to the interpersonal relationships of friends and intimates (Scanlon 2008, chapter 4), this may well be doubted (see, for example, Wallace 2011). And in many ways, our dealings with strangers do take on a more legalistic cast: there are various (informal) norms in place that simply are not to be violated, seemingly independently of one's quality of will in so doing. These norms may thus provide reasons in the way that criminal prohibitions do, and so answerability with respect to their violations may well serve the same functions that it does with respect to criminal violations. Nevertheless, even if Gaus is right, there would at least be a large subset of interpersonal morality (with which we are all quite familiar) in which answerability functions differently than criminal answerability. And in addition, as Drew Schroeder has suggested to me, if moral answerability is indeed not unified, but criminal answerability is, this fact itself would provide yet another disanalogy between the two.

7.5 ACCOUNTABILITY: NORMS AND META-NORMS

To be accountable for Φ is to be appropriately susceptible to being held to account for Φ, where this may involve sanction or reward.[23] The first tenet of the Standard View takes criminal accountability to entail moral accountability, so that agents are appropriately susceptible to being held criminally accountable only if they are appropriately susceptible to being held morally accountable. I believe that this tenet is false.

As discussed earlier, insofar as holding someone accountable often involves communication about or expression of the norms that were violated, the fairness of any associated sanctioning in cases of norm violation depends on the targeted agent's ability to avoid the sanctions,[24] and so crucial to this ability is *the ability to recognize and respond to considerations against performing the action in question*. If you could not see why you should not Φ, then you had no genuine opportunity to avoid Φ, and so it would not be fair to hold you accountable in a way that would involve sanctioning you for Φ. Of course, the relevant sense of "could not" here is a longstanding source of controversy, but it is one I hope to sidestep, given that my only point is that, assuming *some* minimal capacities for the ability to avoid, accountable agents must have them.

What count as the considerations against performing the action in question, however? In answering this question, we will see that, on this score, moral accountability is more demanding than criminal accountability. This, I want to suggest, is because sanctions for morally impermissible actions are fair only as long as the targeted agent was able to avoid the sanctions via her ability to recognize the considerations that made the action impermissible in the first place. Sanctions for the performance of criminally impermissible actions, however, are fair just as long as the targeted agent was able to avoid the sanctions via her ability

[23] While I advocated a necessary connection between moral accountability and sanctions for a long time in several papers, I no longer believe it, and I have explicitly rejected it in Shoemaker Forthcoming. Nevertheless, moral accountability is still a matter of the appropriateness of certain sorts of attitudinal responses, essentially variations on anger and gratitude, and these are of a sort where sanctions or rewards may be unintended *side-effects* of their expression (see McKenna 2012, chapter 6, for discussion). Consequently, the worries about fairness I discuss herein are still relevant.

[24] A point brought out nicely by both Wallace 1994, for example, 199–202; and Watson 2004, 279–80.

to recognize considerations against the performance of the impermissible action, period. What matters in the realm of moral accountability is *responsiveness to the reasons* of moral demands, whereas what matters in the realm of criminal accountability is merely the regulation of one's *behavior in conformity with* the legal demands.

The issue here is complicated, but one way to see what is going on is to imagine an agent living with you who fails to be able to see the reasons that make stealing wrong—for example, that it will set back your interests or that it will cause you pain—but who is nevertheless able to recognize and respond to the fact that some things she does—such as annoying you or making you angry—will make her own life miserable. Suppose that this agent steals from you and you find out both about her action and about her "moral reasons disability." She could have avoided any sanctions that came her way only to the extent she could have responded to the prudential reasons against performing the action. Would actively blaming her be fair? It seems not. "The fact that it would make me angry" does not get to the source of the wrongness in stealing from me. But part of the point of our expressing resentment and general blame to one another is that it communicates the terms of our *ongoing* interaction, and to the extent we are to live together, there are certain sorts of considerations we need to trust to which the other will respond. If my roommate can respond only to prudential considerations about avoiding my wrath, however, there are two problems: (a) my wrath is merely contingent and unpredictable, and so too would be the fairness of sanctions in light of her wrongdoing; and (b) we could not be in the relevant sort of relationship to render my wrath appropriate.

To explain the point of (a), suppose I were to get worn out with my wrathful responses and so cease to respond that way any more. Then there would no longer be a prudential reason for her to avoid the action; she would thus lack access to *any* reason to avoid it. Were I suddenly to regain my wrathful energy and express my resentment to her once more, she would then have lacked any opportunity to avoid its associated sanctions. But how different individual people will in fact respond to wrongdoing is rather unpredictable, given differences in mood, willingness to confront, desires to maintain the peace, and so forth. The fairness of my wrathful response to my morally blind roommate would thus depend on what I knew about what *she* knew about me and the sorts of responses I might have. Indeed, it would further depend on whether she knew that *I* knew about what she knew about me and my responses. This is a sort of knowledge missing from most individual moral interactions. So if she has no fair opportunity to avoid my wrath for her wrongdoing given either (1) her lack of access to the moral reasons against the action, or (2) her failure to meet the strong knowledge conditions just detailed, then my wrathful response in any individual case will rarely, if ever, be fair.

To explain the point of (b), any relationship we have (friendship, professional, moral) is going to be defined in terms of norms or expectations (Scanlon 2008, ch. 4). What is essential to any genuine, non-exploitative moral relationship, however, is that both parties not only recognize and accept these norms but also recognize and accept a meta-norm; namely, a norm stating that the reasons for the relationship-defining norms *are the very reasons for adhering to them*. In other words, these norms set the terms of the parties' relationship for various reasons, and it is these reasons that are to guide the parties' deliberations in following them. For example, a marriage relationship will typically include a "be sexually faithful" norm at its foundation; violating this norm will fundamentally impair the relationship. The reason for the norm is presumably to preserve love and trust. The meta-norm, then, states that both parties ought to be faithful *for the reason that* that norm was instituted; namely, to preserve love and trust in the relationship. One may thus be subject to sanctions in such a relationship if one's partner finds out one is staying faithful only in order not to get caught. The wrath here will be in response to the latter's violation of the meta-norm, for while he is indeed conforming his behavior to the relationship-defining norm, he is not doing so for the right reasons.

Now suppose, as in the original case discussed previously, that one party to a moral relationship is simply blind to the meta-norm. If she cannot recognize it she cannot incorporate it into her moral deliberations. But then, insofar as she has no opportunity to avoid any sanctions based on a violation of that meta-norm, resentment for her "violation" of it would be neither fair nor appropriate. Indeed, the reason for the inappropriateness is precisely the fact that the parties would not even be *in* a genuine moral relationship in the first place, given my roommate's meta-norm disability, and so she *could not* impair a relationship (in a way grounding blame-sanctions) that did not exist.[25]

[25] To the extent that what I say here is taken to be about "the" moral relationship, it is subject to the concerns raised in fn. 22.

What we find when exploring criminal accountability, however, is that *neither* of the two worries about moral reasons-blindness is relevant. Consider again our case of someone who steals but is capable only of recognizing and responding to prudential reasons—reasons that sometimes will counsel stealing. I have argued that she would not be morally accountable because the contingency and unpredictability of individual moral responses to her, as well as her inability to be in the right sort of moral relationship, renders any blaming sanctions unfair and/or inappropriate. But when it comes to the sanctions of criminal accountability—punishment—(a) state responses are neither contingent nor unpredictable (in a way undermining their fairness), and (b) the relevant relationship conditions *can* be met.

To explain the point of (a), it is built into the machinery of the state and the criminal law to punish criminals. This makes the state sanctioning response neither contingent nor unpredictable. To be clear, the relevant datum here is *that the state will respond to criminal accountability with punishment*. Sure, some actual *sentences* are unpredictable—judges sometimes exercise judgment—but that is irrelevant to the main point, which is that citizens can count on the fact that convicted criminals pay (indeed, this is the central social expectation produced by the prohibition reasons discussed in the previous section). This is, in fact, the hallmark of impartial justice: the different moods or desires of individual state executors simply do not blunt the fact that there will, invariably, be a state sanction if one is convicted. The strong knowledge constraints that cannot (more than rarely) be met in the moral realm are thus easily met in the criminal realm. The prudent criminal knows well what awaits.

To explain the point of (b), the relevant relation obtaining between parties sufficient to render state sanctions appropriate would at most be a *citizenry* relation, not necessarily a moral one. What defines this relation are norms about various sorts of behavioral expectations, but there simply is not a meta-norm in place demanding that one adhere to these expectations *for the reasons that give rise to those expectations*. Certain conduct has been rendered criminal—a violation of our citizenry relations—in order to protect citizenry interests deemed to have political worth. In order to pursue various goods, for example, every citizen has an interest in not being harmed that needs protection.[26] Now I may

[26] The arguments of the next few paragraphs borrow from Shoemaker 2011a.

refrain from harming you because I pay attention to, and guide my deliberations by, the reasons grounded in your interests. But I may also refrain from harming you solely because I am afraid of getting caught (and so pay no attention whatsoever to the interest-based reasons). Contrary to the above discussion of the moral relationship and the non-cheating spouse, here the difference between the two cases is *irrelevant* to the criminal law. The fairness of state sanctions thus rests entirely on the defendant's abilities to avoid the sanctions by having access to *any reasons whatsoever* that can steer him away from the criminal act. If these are merely prudential (as in our imagined case), then they are sufficient. What really matters for the criminal law is not that those under its rubric actually track the reasons underlying it and take the politically protected interests of others as reason-giving; rather, all that matters is that those under its rubric are able to act *as if* they took the interests of others as reason-giving.

Now, it is true that most of us adhere to the criminal laws because we take them to articulate moral obligations, and we often do adhere to a meta-norm about these obligations (obeying them for the reasons that ground them); the penalties for violating the criminal law rarely even enter into our deliberations.[27] But for others, it is *only* consideration of these penalties that prevents their wrongdoing, and this is surely enough to render them members of the legal community. But this requires only the capacities for prudence and practical rationality generally. Consequently, it is possible for there to be those who are criminally accountable (possessing this latter capacity) without being morally accountable (not possessing the capacity to recognize moral reasons or meta-norms generally); that is, criminal accountability just does not entail moral accountability.[28]

7.6 CONCLUSION

My aim has been to explore and critically evaluate the alleged relation between CR and MR—in particular, the two tenets of the Standard

[27] For Tadros 2011, this is probably a function of the general deterrence effects of a system of punishment (282).

[28] This is why, as I argue in Shoemaker 2009 and Shoemaker 2011a, psychopaths may be criminally, but not morally, accountable.

View, that (1) CR entails MR, and (2) the constituents of CR are analogous to those of MR. What I have found is that these tenets are either false or far more complicated than has previously been thought, once we articulate and explore the three conceptions of Responsibility that occur in both realms.

First, a key role attributability plays in MR is as a necessary condition for accountability. In the criminal realm, however, strict liability threatens this presupposition: if we preserve such laws in our criminal justice system, it looks like we have to accept that one may be criminally accountable for actions not properly attributable to one. Duff's notable recent attempt to resist this implication by showing how what he calls strict answerability(D) is analogously present in both CR and MR is actually either not strict or not (moral) answerability, and so cannot provide a feasible way out of the jam without creating another disanalogy.

Second, the role played by motivating reasons is quite different in moral and criminal answerability. In the interpersonal moral realm, answerability is fundamentally about the quality of the agent's will, with motivating reasons important for revealing this quality. In the criminal realm, however, answerability is important only after attributability has been established, and so only with respect to establishing possible justifications or excuses (where this involves determining whether a gap exists between motivating and normative reasons). A related difference is found in the role played in each realm by justifications (where there is no gap between motivating and normative reasons): for moral answerability, justifications serve to undermine suspicions of ill will, whereas for criminal answerability, given the additional presence of prohibition reasons, justifications may obtain utterly independently of the quality of the defendant's will.

Third, the criminal law does not demand that one be sensitive to, or even capable of grasping, the meta-norm crucial to interpersonal moral accountability, the norm that one act for the reasons grounding the norms that define one's interpersonal moral relationships. Criminal accountability thus does not entail moral accountability.

The Standard View seems mistaken, then. There are deep structural differences between CR and MR, and the former, at least with respect to accountability, does not entail the latter. Whether this lack of entailment and the other differences I have articulated are worth preserving in the two realms is another matter entirely. At the very least, though, they are factors worth recognizing.[29]

REFERENCES

Alexander, Larry, and Ferzan, Kimberly Kessler (2009). *Crime and Culpability: A Theory of Criminal Law*. Cambridge: Cambridge University Press.

Brudner, Alan (2009). *Punishment and Freedom: A Liberal Theory of Penal Justice*. Oxford: Oxford University Press.

Duff, Antony (2009a). "Legal and Moral Responsibility." *Philosophy Compass* 4/6: 978–86.

—— (2009b). "Strict Responsibility, Moral and Criminal." *Journal of Value Inquiry* 43: 295–313.

Eshleman, Andrew (2009). "Moral Responsibility." In Edward Zalta, ed., *The Stanford Encyclopedia of Philosophy*. <http://plato.stanford.edu/entries/moral-resopnsibility/#RecWorConRes>

Fischer, John Martin (1986). "Introduction: Responsibility and Freedom." In John Martin Fischer, ed. *Moral Responsibility* (Ithaca, NY: Cornell University Press), pp. 9–61.

Gardner, John (2007). *Offences and Defences*. Oxford: Oxford University Press.

Hart, H. L. A. (2008). *Punishment and Responsibility*, 2nd edition. Oxford: Oxford University Press.

Husak, Douglas (2010). "Review of Alan Brudner's Punishment and Freedom: A Liberal Theory of Penal Justice." *Ethics* 120: 841–6.

McKenna, Michael (2012). *Conversation and Responsibility*. Oxford: Oxford University Press.

Moore, Michael S. (1997). *Placing Blame*. Oxford: Oxford University Press.

Morse, Stephen J. (2008). "Psychopathy and Criminal Responsibility." *Neuroethics* 1: 205–12.

Scanlon, T. M. (1988). "The Significance of Choice." In Sterling McMurrin, ed. *The Tanner Lectures on Human Values*. Salt Lake City, UT: University of Utah Press, 1988, pp. 149–216.

—— (1998). *What We Owe to Each Other*. Cambridge, MA: Belknap Press of Harvard University Press.

—— (2002). "Reasons and Passions." In Sarah Buss and Lee Overton, eds., *Contours of Agency*. (Cambridge, MA: MIT Press), pp. 165–83.

—— (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Belknap Press of Harvard University Press.

Sher, George (2006). "Out of Control." *Ethics* 116: 285–301.

Shoemaker, David (2009). "Responsibility and Disability." *Metaphilosophy* 40: 438–61.

[29] For helpful advice on earlier drafts of this essay, I am grateful to Nicholas Sars, Victor Tadros, audience members at my presentations of this material in 2011 at the Tulane Law School, the 2011 Ohio St/Dubrovnik conference, the 2011 International Society of Utilitarian Studies conference, and the 2012 Normative Ethics Workshop in Tucson. I also thank the student members of my seminars on criminal and moral responsibility at Tulane University and Bowling Green State University during spring and summer 2010. Thanks also to those who contributed helpful comments when I posted on related topics in May 2011 on the ethics blog PEA Soup; for example, here: <http://peasoup.typepad.com/peasoup/2011/05/scanlon-on-blame-part-3-criminal-blame-and-meaning.html> (including Mark van Roojen and, again, Victor Tadros). Finally, for their extremely helpful comments, questions, and discussion I am very grateful to Justin D'Arms, John Deigh, Jerry Gaus, Doug Husak, Dan Jacobson, Michael McKenna, Connie Rosati, Drew Schroeder, an anonymous referee for Oxford University Press, and, most especially, David Sobel.

—— (2011a). “Psychopathy, Responsibility, and the Moral/Conventional Distinction.” *The Southern Journal of Philosophy* 49, Spindel Supplement: 99–124.

—— (2011b). “Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility.” *Ethics* 121: 602–32.

—— (2013). “Punishment and Blame.” In Neal Tognazzini and Justin Coates, eds., *Blame* (Oxford: Oxford University Press).

—— (Forthcoming). “Qualities of Will.” In *Social Philosophy & Policy*.

Smith, Angela M. (2005). “Responsibility for Attitudes: Activity and Passivity in Mental Life.” *Ethics* 115: 236–71.

Strawson, P. F. (2003). “Freedom and Resentment.” In Gary Watson, ed., *Free Will*, 2nd edition (Oxford: Oxford University Press), pp. 72–93.

Tadros, Victor (2005). *Criminal Responsibility*. Oxford: Oxford University Press.

—— 2011. *The Ends of Harm*. Oxford: Oxford University Press.

Wallace, R. Jay (1994). *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

—— (2011). “Dispassionate Opprobrium: On Blame and the Reactive Sentiments.” In R. Jay Wallace, Rahul Kumar, and Samuel Freeman, eds. *Reasons and Recognition: Essays on the Philosophy of T. M. Scanlon*. Oxford: Oxford University Press, 2011, pp. 348–72.

Watson, Gary (2004). *Agency and Answerability*. Oxford: Oxford University Press.

Wolf, Susan (1987). “Sanity and the Metaphysics of Responsibility.” In Ferdinand Schoeman, ed. *Responsibility, Character, and the Emotions*. Cambridge: Cambridge University Press, pp. 46–62.

8
Consequentialism, Cognitive Limitations, and Moral Theory[1]

DALE DORSEY

Consider the moral theory called "objective act-consequentialism" (hereafter, "consequentialism"). This view holds that the moral quality of a given action φ in comparison to another action ψ is determined by the quality of φ's actual consequences in comparison to the quality of ψ's actual consequences. There are many reasons people have found to reject this sort of view. Some have claimed that it is too demanding.[2] Some have claimed that it disrupts a moral agent's integrity.[3] Some suggest that it inappropriately denies that there are no constraints on the pursuit of the overall good.[4]

Other objections focus on consequentialism's ill fit with the fact of human cognitive limitations. Traditionally, such objections focus on our general ignorance of the consequences of our actions, and hence their moral valence.[5] However, a recent and penetrating objection notes that the problem of cognitive limitations runs much deeper. Consequentialism seems to require us, at virtually all times, to perform actions that, though strictly possible for us to perform, we will not perform *simply* given our everyday limitations as cognitive agents. This is

[1] Special thanks are due to Doug Portmore, whose generous comments radically improved an earlier draft of this essay. I would also like to thank John Bricke, Richard DeGeorge, Julia Driver, Ben Eggleston, Erin Frykholm, Scott Jenkins, Sean McKeever, Howard Nye, Mark van Roojen, Drew Schroeder, Dave Shoemaker, Daniel Star, and participants in the 2012 Arizona Normative Ethics Workshop for helpful conversations.

[2] Samuel Scheffler, "Morality's Demands and their Limits," *The Journal of Philosophy* 83 (1986): 531–7.

[3] Bernard Williams, "A Critique of Utilitarianism," in Smart and Williams, *Utilitarianism: For and Against* (Cambridge: Cambridge University Press, 1974).

[4] Frances Kamm, "Non-consequentialism, the Person as an End-in-Itself, and the Significance of Status," *Philosophy and Public Affairs* 21 (1992): 354–89.

[5] Consider, for instance, James Lenman, "Consequentialism and Cluelessness," *Philosophy and Public Affairs* 28 (2000): 342–70.

true even *if* we know that the action in question will have tremendously good consequences, better than any other alternatives.[6]

In this essay I argue that this problem is not isolated; it applies not just to consequentialism, but to virtually every plausible moral theory. I argue that the root of this problem is to be found in what might be called a *theory of moral requirability*. I propose an independently plausible alternative to the traditional theory of moral requirability, and argue that this alternative successfully avoids commitment to a view that renders morally acceptable behavior out of reach for virtually all normal human beings.

8.1 COGNITIVE LIMITATIONS AND MORAL REQUIREMENTS

It is well-known that consequentialism has some sort of problem with actions one can, strictly speaking, perform, but actions that one will not perform given normal cognitive limitations.[7] Examples of this kind abound. Take, for instance:

Wilda: Wilda stands before a safe. Were she to open the safe, overwhelmingly good consequences would result. However, she has no idea what the combination of the safe is.

Despite the fact that Wilda does not know the combination of the safe, consequentialism would appear to require her to open the safe. After all, she *can* open the safe, by entering the right numbers in the right order. But it would seem odd to say that she behaved in a morally impermissible manner in failing to do so. After all, she does not know how.

Some consequentialists might feel comfortable simply biting this bullet. As Wiland notes, to say that Wilda is morally required to open her safe seems more or less innocuous, insofar as her case is pretty unusual,

[6] Eric Wiland, "Monkeys, Typewriters, and Objective Consequentialism," *Ratio* 18 (2005): 352–60.

[7] See, for instance, Francis Howard-Snyder, "The Rejection of Objective Consequentialism," *Utilitas* 9 (1997): 241–8; Erik Carlson, "The Oughts and Cans of Objective Consequentialism," *Utilitas* 11 (1999): 91–6; Mozaffar Qizilbash, "The Rejection of Objective Consequentialism: A Comment," *Utilitas* 11 (1999): 97–105; Dale Miller, "Actual-Consequence Act Utilitarianism and the Best Possible Humans," *Ratio* 16 (2003): 49–62, at, 50; Eric Wiland, "Monkeys, Typewriters, and Objective Consequentialism," *Ratio* 18 (2005): 352–60.

and, furthermore, consequentialism generally refuses to index an individual's moral requirements to his or her epistemic states.[8] Somewhere along the line, however, this sort of bullet-biting starts to sound a bit goofy. Take another example:

> *Jeffrey*: Jeffrey is spending a leisurely afternoon reading a titillating detective novel. He could, alternatively, sit down at his computer, type the cure for cancer, and send it to a top medical journal. The consequences of performing this action would be overwhelmingly good. It does not occur to Jeffrey to do this not because he believes that curing cancer would have suboptimal consequences, but rather because he has no clue what the cure for cancer is, or how to write it down.

Jeffrey, in a perfectly sensible sense of "can," *can* widely disseminate the cure for cancer. All he needs to do is to sit down, type it, and disseminate it widely. Jeffrey can type. No one is stopping him. Assuming such a cure exists, if we also make the relatively innocuous assumption that curing cancer would produce overwhelmingly good consequences,[9] it would appear that he morally ought to do so.[10] Of course, there surely are additional senses of "can" that entail that Jeffrey cannot cure cancer.[11] However, the sense of "can" I work with is the sense of "can" that is implied by "did": if I φ-ed at *t*, it follows that I could have φ-ed at *t*. (In essence, we do not want to say that it could occur that someone performs an action at *t* that he or she *cannot* perform at *t*.[12]) Insofar as there is nothing stopping Jeffrey from having cured cancer at *t*—all he has to do is hit the right keys in the right order—it follows that he *can* cure cancer at *t*. Given the overwhelmingly good consequences of curing cancer, it would appear that simply continuing to read his detective novel is *far* below the moral ideal.

But this is surely absurd. Though one might be willing to bite the bullet in Wilda's case (though this does not strike me as particularly plausible in any event), Jeffrey's case illustrates that the problem at hand is not confined to the stranger than fiction. In fact, as Wiland puts

[8] Wiland, 355–6.

[9] For a more pessimistic view of the good generated by a cure for cancer, see Albert Brooks's novel *2030*.

[10] See Wiland, 356–7.

[11] See Howard-Snyder, 243–4.

[12] See Miller, 53–4.

it: "once we become literate creatures, the number of action-options usually available to us is astronomically high . . . The chance you ever do what in fact has the best consequences is laughably small. If you act rightly only if you do what in fact has the best consequences, then you can be fairly certain that you never (or almost never) act rightly."[13] And, or so it would appear, there is good reason to reject (objective act-) consequentialism.

It is worth saying a little about what the objection in question amounts to. Wiland and Miller focus on the suggestion that "[objective consequentialism] implies that our actions are *almost always* wrong."[14] And this is bad enough. But I think the objection is broader than this. Recall that even if we focus strictly on Wilda's unusual case, the fact that she does not know how to open the safe seems to exempt doing so from the set of those actions that can be sensibly morally required of her. Thus the problem is not—or not only—that according to consequentialism we almost always act wrongly. Rather, the problem with consequentialism appears to be that it—however rarely or often—requires us to perform actions not that we cannot perform, but that we *will not* perform *simply* as a result of our limitations as cognitive agents (perhaps along with a failure to have cosmically good luck). This is true even if we have perfect knowledge of the relevant consequences. But this is absurd. Call this the "problem of cognitive limitations."

8.2 COGNITIVE LIMITATIONS AND NON-CONSEQUENTIALISM

As stated, the objection from cognitive limitations is an objection to consequentialism. But the fact of cognitive limitations causes problems not just for consequentialism but, as I shall argue, virtually every moral theory. Take the following principle:

Minimal Consequentialism (MC): that φ-ing will produce good consequences is a moral reason to φ, a reason that strengthens as the quality of the consequences of φ-ing increases.

[13] Wiland, 359.

[14] Wiland, 353; Miller, 55.

Minimal Consequentialism is shared by consequentialists and many non-consequentialists. For consequentialists, MC is sufficient to fully catalog moral reasons. However, a moral theory can accept *Minimal Consequentialism* without accepting consequentialism. After all, MC says only that good consequences of an action generate *a* moral reason for performance, not that this reason is overriding, or even comparatively very important. For instance, one could accept MC and a host of agent-centered restrictions, permissions, and so on. However, any theory that accepts this general principle will face the problems noted in Section 8.1.[15] Take Jeffrey's case. Curing cancer by means of typing out the formula and widely disseminating it will produce overwhelmingly good consequences. But given that good consequences generate moral reasons, and given that the good consequences of curing cancer are *overwhelming*, one would expect that no matter how comparatively unimportant the reason to promote good consequences is, the sheer magnitude of good caused by such an action would be enough to entail that the reason to do so is sufficient to morally require doing so. To put this another way, it would appear that (a) the enormity of the good involved should, on any plausible view, morally override any non-consequentialist permissions or restrictions, and (b) even if not, it is very unlikely that typing out the cure for cancer would be particularly burdensome for Jeffrey (beyond slightly delaying the *denoument* of his gooseflesher), nor would it require him to commit an unpardonable moral sin. Even if the reason to promote goodness is very weak compared to all other moral reasons, any MC-accepting theory will face the problem of cognitive limitations.

One possible solution, whether for explicitly consequentialist theories or for theories that treat consequences as only one factor among many, is to reject the claim that moral reasons are a product of *actual* consequences. Rather, one might hold that moral reasons are generated by *expected* consequences, expected, presumably, by the agent in question. Correspondingly, such a view would accept:

[15] An anonymous reviewer notes that some moral theories will hold that moral reasons must also be paired with, say, additional "requirement generating" factors to generate a moral requirement, which, in Jeffrey's case, say, will not be present. But this suggestion is just another way of getting at the central problem of this essay: which acts are morally *requirable*, and what makes them so?

> *Minimal Subjective Consequentialism* (MSC): that φ-ing will produce good expected consequences is a moral reason to φ, a reason that strengthens as the quality of the expected consequences of φ-ing increases.[16]

MSC suggests that the moral valence of action is, at least in part, determined by the action's expected, rather than actual, consequences. MSC permits of varied interpretation; some will say that expected consequences are those that can be expected given the agent's actual epistemic states. Others will say that expected consequences are determined by what an agent would expect under certain idealized or counterfactual epistemic conditions.

Offhand, however, it is a little difficult to see why any view that accepts MSC rather than MC could avoid the problem of cognitive limitations.[17] After all, the expected consequences of my curing cancer are certainly extremely good; the expected consequences of any other action I could perform at the time pale in comparison. If so, MSC is no fix: curing cancer not only has the best *actual* consequences, but also the best *expected* consequences. This is true even if we hold that expected consequences are determined from the perspective of the agent's actual epistemic states. As I sit, right now, I certainly have sufficient information to determine that the expected consequences of curing cancer are extraordinarily good. If so, or so it would seem, I am morally required to do so.

A response is worth considering. One might say that though the expected consequences of *curing cancer* are good, the expected consequences of taking the steps necessary to cure cancer—that is, typing a set of random characters on one's keyboard—are not particularly good. After all, the chances that I will actually cure cancer by sitting down and typing out a bunch of random words or characters are pretty slim. And if so, I cannot be required to do so: the expected consequences are of very low comparative quality. But this response does not solve the problem. While I may not be required to sit down at my computer and

[16] Wiland, for instance, explicitly focuses on what he calls "objective" consequentialism—a view that accepts MC rather than MSC, though he does not consider whether subjective consequentialism would be able to avoid the problem at hand.

[17] Dale Miller, in proposing a version of the problem discussed here, suggests that "even some utilitarian or non-utilitarian forms of probable-consequence act consequentialism" may succumb to it (Miller, 50).

punch keys randomly—which is a means to curing cancer—I am nevertheless required, given its foreseeable consequences, to *cure cancer*. To claim that I am not required to cure cancer given that I am not required to sit down at my computer and type a bunch of characters randomly is to treat as identical two distinct acts: the act of curing cancer and the act of typing a bunch of random keys.[18] And insofar as we can distinguish these actions, as we surely can, any view that accepts MSC will generate a moral reason to cure cancer. Thus MSC does not avoid the problems cited here.

Here is a promissory note: I argue in the next section that *any* halfway plausible moral theory will face the problem of cognitive limitations. However, even if we allow that entirely non-consequentialist moral theories can avoid these problems, MC and MSC are critical features of a very wide range of moral theories. Any moral theory that cares about the consequences of our actions will accept one or the other. Insofar as there appears to be very good moral reason to care about consequences (even if just a little bit), a fix for the problem of moral obligation in light of cognitive limitation is thus of the first importance not just for consequentialism, but for moral theory more generally.

8.3 MORAL REQUIRABILITY

The reflections just offered seem to indicate that a solution to the problem at hand cannot be found simply by subbing out one moral theory for another. Rather, insofar as any theory that accepts MC or MSC seems to succumb to the problem of cognitive limitations, it would appear that the solution must be found elsewhere.

To see this in more detail, I want to introduce a bit of terminology. Call an action φ "requirable" for a person x at time t if and only if x's φ-ing is suitable to be assigned a deontic valence at t. Every moral theory must offer what might be called a "theory of moral requirability," that

[18] See, for instance, Alvin Goldman, *A Theory of Human Action* (Englewood Cliffs, NJ: Prentice Hall, 1970), chs. 1 and 2. I should note here that I will assume a fine-grained theory of action-individuation. This is, of course, controversial. But nothing substantive rides on this. One could say also that Jeffrey expects very good consequences to result from his action of curing cancer *by* typing the right characters in the right order, which is surely correct. I will assume a fine-grained scheme in this essay, and will put arguments in terms of such a scheme, but the arguments can be translated to a coarse-grained scheme *mutatis mutandis*.

is, a method by which to determine, for any person at any time, what set of actions are eligible to be assigned a deontic valence. But, in addition, every moral theory must offer what might be called a "theory of moral valence." A theory of moral valence takes the requirable acts and determines their deontic valence; some will be required, and the others will have some different valence (permissible, impermissible, and so on). As a theory of moral valence, consequentialism will assign deontic categories to requirable actions as a result of the quality of their actual consequences. Other theories of moral valence will assess requirable actions differently, according to a set of specified moral factors.

Notice, however, that substantive theories of moral valence need not differ on the proper theory of moral requirability. A strict Kantianism and a strict act-utilitarianism will certainly differ concerning what morally requirable action, for *x* at *t*, is morally required, and so on. But they need not differ about which actions are *requirable*, or which actions can be properly assessed for a particular comparative deontic valence.

By far the most popular theory of moral requirability is:

The Traditional View (TRAD): φ is requirable for *x* at *t* if and only if *x* can φ at *t*.

TRAD is compatible with any theory of moral valence. TRAD does not say how one ought to evaluate the actions it identifies as morally requirable. Rather, it simply restricts the set of actions that can properly be assigned a moral deontic valence to those that can be performed by the agent.

The root of the problem of cognitive limitations lies in accepting TRAD, which holds that any action one can perform is in principle eligible to be assigned a deontic valence. But this yields that, for Jeffrey, curing cancer will be among those actions that will be evaluated by a theory of moral valence. But no halfway plausible theory of moral valence, whether consequentialist or non-consequentialist, will hold that *assuming that curing cancer is morally requirable for Jeffrey*, Jeffrey is not morally required to cure cancer. To say so is to render one's theory of moral valence too absurd for consideration *if* we assume that to cure cancer is an action that can be assigned a moral valence. But if this is correct, the problem of cognitive limitations lies in TRAD, not in any particular theory of moral valence. Hence it would seem sensible, if we are concerned about the puzzle of cognitive limitations, to narrow the range of morally requirable actions whatever one's preferred theory of moral valence.

8.4 ALTERNATIVES

But how? Examining the cases of Wilda and Jeffrey might lead to an obvious possibility:

The Knowledge View (KNOW): φ is morally requirable for *x* at *t* if and only if *x* knows how to φ at *t*.

Insofar as Wilda does not know how to open the safe, and Jeffrey does not know how to cure cancer, it would appear that KNOW entails that they cannot be morally required to do so even given the quality of the resulting consequences. So far so good.

However, there are good reasons to reject KNOW, which is both under-inclusive and over-inclusive. First, it is under-inclusive. Consider:

Sondra: Sondra stands before a safe that, were she to open it, cancer would be cured. Also imagine that Sondra believes that entering the numbers 867–5309 will open the safe. Furthermore, her belief is justified and true—this number is, in fact, the safe's combination. However, imagine that Sondra's justified true belief was a product of a Gettier-like scenario,[19] and hence that this particular number will open the safe does not count, for Sondra, as knowledge.

Whatever else is true about the relationship between knowledge-that and knowledge-how, it is *surely* the case that to know how to open the safe requires one to know its combination. But in this case, Sondra does not know the combination. Her justified true belief is not knowledge. Hence, she fails to know how to open the safe. Nevertheless, it would seem absurd to say that she could not be required to open the safe. After all, she believes she knows how to do so, that belief is justified, and, were she to actually enter the number she believes will open the safe, cancer

[19] Imagine, for instance, that Sondra overheard the owner of the safe declare that anytime he had a safe, he would set the combination to the title of his favorite Tommy Tutone song. But also imagine that the owner never actually set the combination, and it just happened to have been set to that number randomly in the factory.

would be cured. Not only can she be required to open the safe, her failure to do so is a grave moral wrong.

Second, Know is over-inclusive. Some actions we know how to perform are inappropriate targets of moral obligation even if we know how to perform them. Assume, for instance, that the correct combination to Wilda's safe is 867–5309. Wilda certainly *knows how to enter* 867–5309. Insofar as she knows how to work a safe, she knows how to enter any particular number into that safe.[20] And insofar as performing this action entails *curing cancer*, many theories of moral valence will, if doing so is requirable, require it.[21] But requiring Wilda to enter 867–5309 is certainly no more plausible than requiring Wilda to *open the safe*.[22] After all, if the problem of cognitive limitations is that people, like Wilda, will fail to perform morally required actions *simply* given the standard cognitive limitations of human beings, this problem surely applies to a moral requirement to enter 867–5309 no less than to a requirement to open the safe. Parallel reasoning applies in Jeffrey's case. We can assume that there is some string of characters, *S*, such that were Jeffrey to type *S*, he would cure cancer. Insofar as Jeffrey knows how to type, he knows how to type *S*. Know implies that Wilda can be morally required to enter 867–5309, and Jeffrey can be morally required to type *S*. But it seems implausible to say that, for example, Wilda should be required to do so; the reason surely is, at least in part, that she fails to recognize entering 867–5309 as a method by which to open the safe, just as Jeffrey fails to recognize *S* as a method by which to cure cancer.

[20] See Carlson, 92.

[21] This might not be a problem for Know on the assumption of MSC. It is certainly the case that entering 867–5309 does not have very good expected consequences. Nevertheless, Know seems to me to fail for two reasons, and the former is applicable even if we accept MSC rather than MC.

[22] An anonymous reviewer disputes this claim. The contrary proposal is that to require Wilda to enter 867–5309 is *much* more plausible, insofar as this is an open option for Wilda. After all, *she knows how to do it* and *to do it would be to produce the best consequences*. But this seems to me not enough; this reasoning begs the question in favour of Know. A fan of Trad might say that to open the safe is also an open option. She *can* do it. And so we need some way to settle what it means for a particular action to be an open option. This can only be settled, as far as I can tell, by figuring out which is the most plausible theory of moral requirability. For my money, it simply is not plausible to say that Wilda should be required to enter 867–5309, for every reason we find it implausible to say that Wilda should be required to open the safe. And hence we should look beyond Know. I admit, however, that this is simply my considered judgment, which just like all the others is subject to dissent.

I think we should reject Know. It seems wrong to say that the central cognitive limitation involved in the problem of cognitive limitations is knowledge of how to perform the action in question, for the reasons just cited. But we might do better by indexing requirability not to one's epistemic stance *vis-á-vis* how to φ, but rather to one's ability to successfully carry out an intention to φ. Take:

The Intention View (Intend): φ is morally requirable for *x* at *t* if and only if *x* can intend to φ at *t*, and were *x* to intend to φ at *t*, *x* would φ at *t*.[23]

This view seems to avoid holding Wilda and Jeffrey morally accountable for failing to open the safe, and failing to cure cancer, respectively. Given their cognitive limitations, we can assume that, were they to intend to cure cancer or open the safe, they would (under most normal conditions) fail to do so. Intend also avoids the problematic suggestion that Sondra cannot be morally required to open the safe. However, Intend also seems to imply that Wilda can be morally required to enter 867–5309, and Jeffrey can be morally required to type the string of characters designated here as *S*. Were they to intend to perform those actions, they would do so.

An important response arises if we conjoin Intend to MSC rather than MC, we can hold that, because the expected consequences of entering 867–5309 are not particularly good, Intend + MSC avoids the problematic claims that we have discussed so far. However, I remain unconvinced that Intend + MSC can solve the problem in a satisfactory way. It seems right to say that Wilda could intend to open the safe *by* entering 867–5309.[24] Of course, the expected consequences of opening the safe by entering any number at all are extraordinarily good. And were it the case that Wilda were to intend to open the safe by entering 867–5309, she would succeed at so doing.[25]

[23] See Howard-Snyder, 244.

[24] Two notes. First, the fine-grained theory of action will refer to this as a "longish" action; see the next section for more discussion of this idea. Second, this claim depends on a controversial view of intention, as any interpretation of Intend surely does. Some hold that one cannot intend to φ unless one believes one will be successful in φ-ing. But there are well-rehearsed reasons to doubt this claim. As Bratman notes, I can intend to rescue someone, despite the fact that I have very little confidence that I will actually succeed at so doing. See Michael Bratman, *Intention, Plans, and Practical Reason* (Cambridge, MA: Harvard University Press, 1987), 38. See also Hugh J. McCann, "Settled Objectives and Rational Constraints" in *American Philosophical Quarterly* 28 (1991), 27.

[25] Even if we reject the suggestion that Wilda could intend to open the safe by entering 867–5309, there remains a problem: Intend + MSC seems to get the wrong answer in Joey's case, which I explore in Section 8.6.1.

I regret not having the space to consider a number of other potential theories of moral requirability; I have surely not yet scratched the surface of the sheer volume of potential iterations of the various views.[26] Many alternatives, some perhaps plausible, may remain, and so I leave open that any or all of the previous views might be tweaked here or there to arrive at better results. However, I believe that there are special reasons to accept the proposal I advocate here (even leaving aside the problem of cognitive limitations), and so I will focus on reasons for and against my preferred alternative from here on out.

8.5 THE AGENCY VIEW[27]

To introduce my proposal, let us consider just what it is that Sondra maintains, but that Wilda lacks. What *cannot* Wilda do that Sondra can? Though there may be many such things, one thing that speaks to me is that Wilda cannot *reason her way* to opening the safe. Sondra can, on the basis of reasons she recognizes, open the safe, or enter the right numbers, and so on. For Wilda, whether she opens the safe or enters the right numbers is the result, quite simply, of a lucky guess.

If this is correct—and I think it is—one might consider the following alternative theory of moral requirability:

The Agency View (AGENT): φ is morally requirable for x at t if and only if x can φ at t as an agent.

To see AGENT in more detail, and why it offers a solution to the problems I investigate, I must first explore the notion of "agency" with which I seek to work. My intention is not to provide a full theory of agency or any cognate concepts. I remain neutral on such topics as whether moral deliberation is a result of cognitive, conative, or affective pro-attitudes, the nature of autonomy or autonomous action, whether autonomous agents act under the guise of the good, or whether normative judgments are necessarily motivating, and so on. Rather, I seek a minimal theory of what it means to perform a particular act at a particular time *as an*

[26] In particular, I would like to than Sean McKeever and Mark van Roojen for proposing a variant of INTEND that I simply do not have space to discuss.

[27] This section was improved by helpful conversations with Julia Driver.

agent—one that I hope will be broadly ecumenical when it comes to substantive controversies such as those noted here.

In assessing whether a particular action can be performed as an agent in the sense I mean, it is insufficient to say that the action was (a) performed by *x* and (b) *x* is a moral agent. Moral agents can sometimes fail to perform actions *as* agents. For instance, moral agents can decide based on whim or caprice, they might on occasion govern their actions by the simple flip of a coin. Doing so does not entail that they are not moral agents. But doing so does entail that the resulting actions are not performed *as agents*. To perform an action as an agent is to perform that action on the basis of one's own deliberative agency. It is, in other words, to *see a reason to perform that action*, and to *perform the action on the basis of that reason*.

Two qualifications should be made clear immediately. To perform a particular action as an agent, it is insufficient simply to see *some reason or other* to perform that action. To perform an action as an agent, the reason one sees, or is responsive to, must be of the right *kind*. To see what I mean, take the following case.

Roulette: I stand at a roulette table, and I see reasons to play roulette: it will be thrilling, it will satisfy a desire of mine, etc. However, I see no reason to bet on any *particular* number. I bet on 22.

In *Roulette* I have performed a number of different actions. I have played roulette. I have bet on 22. I have bet on a black number, an even number, and a number in the second third of the table. However, if we ask "have I bet on 22 as an agent? " it seems to me that the answer is clearly *no*. Of course, I have played roulette as an agent, insofar as I deliberated on the basis of reasons I recognize to play roulette. But I have not bet on 22 on the basis of my deliberative agency. This is true despite the fact that I see clear reasons to bet on 22. Insofar as betting on 22 is a way to play roulette, this is certainly a reason I will recognize to bet on 22. Insofar as betting on 22 is one way to not cut my own legs off, which I would prefer to avoid doing, I certainly recognize a further reason to bet on 22. But importantly, none of these considerations will tell in favour of betting on 22 rather than placing any *other* bet on the roulette table. The reasons I recognize do not support betting on 22 *per se*, but rather a disjunction of potential actions I might perform.

Thus to perform an action as an agent it is not sufficient to see some reason or other to perform the action. Rather, one must see what I shall dub a "contrastive" reason. A contrastive reason to φ is a reason to φ *rather than* any other action ψ one might perform. The reason(s) one sees cannot simply support a disjunction of potential actions. They must support *that action in particular*.[28]

The second qualification runs as follows. To φ as an agent it is not sufficient to φ on the basis of some contrastive reason I see to φ. Rather, it must be the case that the difference between φ-ing and merely *trying* to φ is under one's deliberative control. To see why this is important, take the following example:

Sarah: Sarah, a scientist, believes that there are good reasons to produce organism X in a certain petri dish. However, though Sarah knows that she must, at the very least, leave open the petri dish, she does not know what to add to it, where to place the petri dish, at what temperature the petri dish must be kept, or any other details about the process by which organism X is grown. On a whim, she places it at a certain spot and temperature, adds random chemicals, and leaves for the night. When she returns in the morning, she has grown organism X.

We would say that Sarah grew organism X, surely. But would we say that she grew organism X *as an agent*, as a result of her deliberative agency? Surely not. This is true despite the fact that she sees contrastive reason to grow organism X, and, one can assume, threw a bunch of chemicals together partly on the basis of this contrastive reason. How, then, to explain what Sarah lacks?

To explore this point in more detail, it is helpful to make use of the idea of a "longish" action.[29] Some actions unfold over time, and are made up of individual sub-acts. Sarah has deliberative control over

[28] One might say that this radically restricts the range of actions that I might perform as an agent. Imagine, for instance, that I recognize a reason that tells in favour only of *two* potential actions φ and ψ; though it rules out all others I might perform. On this view, if I see no reason to φ rather than ψ I cannot φ as an agent, despite the fact that I might deliberate, and on the basis of my deliberation, decide to either φ or ψ. But it seems to me that this verdict is exactly the correct one. If I cannot see a reason to φ rather than ψ, we might say I "decided, as an agent, to φ or ψ." If we allow the possibility of disjunctive acts, we may say I φ-or-ψ-ed as an agent. But we would not say that I φ-ed as an agent.

[29] Goldman, ch. 2.

some of the individual sub-acts and not others. Hence we can make a distinction between the sub-acts of growing organism X that Sarah can, and those that Sarah cannot, perform as an agent. Some of these sub-acts will spell the difference between trying to grow organism X and actually succeeding at doing so (which she does, but not as an agent). Sarah has deliberative control over whether she washes up, uses a petri dish rather than a dirty lunchbox, places it in a temperature-controlled incubator rather than in a hot iron skillet, and so on. These are all sub-acts for which Sarah sees contrastive reason. Nevertheless, these sub-acts are not sufficient to succeed at growing organism X. The petri dish must be placed at the *right* temperature, with the *right* chemicals. But Sarah can see no reason to select the right chemicals rather that any other set of (wrong) chemicals. To be clear: I do not mean to say that any action one could perform as an agent must be "longish" in this sense, or that for every such action there is a distinction between trying on the basis of contrastive reasons and succeeding on the basis of contrastive reasons. Sometimes one has deliberative control over whether one φ-s (rather than merely tries to φ) in simply seeing contrastive reason to φ; if we assume that there is such a thing as a "basic action" it could be that seeing contrastive reason to perform it may be enough to perform it as an agent. But sometimes one can see contrastive reason to perform an action without having success rather than a simple try under one's deliberative control. This is illustrated in Sarah's case. And in such cases, to perform the action in question as an agent, one must see contrastive reasons to perform the individual sub-acts that spell the difference between success and failure.

Thus my account of what it means to perform an action as an agent runs as follows. One φ-s as an agent to the extent that (a) one φ-s, (b) one sees contrastive reason to φ, and (c) one sees contrastive reason to perform any individual sub-acts that constitute the difference between merely *trying* to φ and φ-ing (if φ-ing is a "longish" act, *a lá* Sarah's case). Once these conditions are in place, one φ-s as an agent.

One note before I move on. Some will complain that my account of what it means to φ as an agent does not match their considered judgments. Fair enough. However, I argue in the next section that there is independent reason to believe that morality should not treat as requirable actions that could not be performed as an agent *in my sense*. If that is right, then not much will ride on whether my account of the performance of an action on the basis of deliberative agency is the true account. Readers can feel free to treat my account as stipulative, as an account of "schmagency," as it were, if they so choose. The reasons in favour of Agent (understood in my way) will remain the same.

To begin, note that Agent—understood in my way—avoids the problem of cognitive limitations. Recall that it seems implausible to say, of Wilda, that (a) she can be morally required to enter 867–5309, and (b) she can be morally required to open the safe. Take, first, entering 867–5309. In this case it is relatively clear that Wilda cannot see any contrastive reason to enter that number. Were she actually to enter the right number, she would not do so on the basis of deliberation, but simply as the result of a guess, a random selection, and so on. This is because, as noted before, she does not recognize 867–5309 *as* the safe's combination. Parallel reasoning applies in the case of Jeffrey's moral requirement to type *S*.

Take now Jeffrey's purported moral requirement to cure cancer. Though Jeffrey can see contrastive reasons to cure cancer, he cannot do so as an agent. Recall that, as in Sarah's case, to perform the action as an agent, one must see a contrastive reason to perform the individual sub-acts that spell the difference between trying to φ and φ-ing. But insofar as Jeffrey is unaware of the proper procedure to cure cancer, he is unable to see reasons to take the *proper* procedure rather than some alternative *improper* procedure. Because Jeffrey has not the foggiest idea how to cure cancer, he cannot *cure* cancer as an agent insofar as he sees no contrastive reason to take the necessary steps rather than any other set of steps. If, against all odds, he actually cures cancer he will not have done so as an agent. He will have *tried*, perhaps as an agent, to cure cancer but—as in Sarah's case—he will not have seen contrastive reason to perform the crucial sub-acts that constitute *curing* (rather than merely trying to cure) cancer. His performance of them would have been the result of mere guesses. Parallel reasoning applies in the case of Wilda's moral requirement to open the safe.

I think a general conclusion is worth stating here. There are many cognitive limitations that we may possess. Some of these limitations will, if we accept Trad or other theories of moral requirability, render it the case that we will not perform actions we are morally required to perform. But if we accept Agent, the problem of cognitive limitations is eliminated; given that morally requirable actions are those that an individual can perform as an agent—that is, as a result of deliberative agency—cognitive limitations *of themselves* will not determine whether we do or do not conform to our moral obligations. The set of actions I can perform as an agent is determined at least in part by what actions I will not perform *given* my cognitive limitations. This set of actions surely excludes any actions that can be performed by me simply as a matter of luck *given* my cognitive limitations. And thus the problem of cognitive limitations—whatever the relevant cognitive limitations are—is dissipated. Under Agent, for any morally required action I fail to perform, it remains the case that I *could have* performed that action *as a deliberative agent*.

8.6 objections

In the final section of this essay I consider three important objections to Agent. The first argues that Agent is *ad hoc*, especially in light of the specialized theory of what it means to act as an agent offered in the previous section. The second argues that my view unduly restricts the range of objective moral reasons. The third argues that Agent is under-inclusive: it rules out the possibility of morally requiring particular actions that really should be requirable.

8.6.1 Ad hoc

Perhaps the most important objection to Agent runs like this: why believe that this account of agency has anything to say about what is or is not morally requirable? Even if, or so it may be claimed, Agent is successful at turning back the problem of cognitive limitations, what independent reason is there to believe it? Is not this simply an *ad hoc* addition to our understanding of the boundaries of moral assessment?

I do not think so. Indeed, to recognize contrastive reason to perform an action seems a perfectly sensible constraint on moral obligation independently of the problem of cognitive limitations. Consider the following case:

Joey: Joey is the victim of a dastardly neuroscientist, who has implanted a device in his brain. This device shuts off Joey's ability to see contrastive reasons for or against committing certain acts of aggression for a limited period of time (about a half-hour) on Thursday evenings. Before and after this period, his deliberative capacity operates in perfect working order.

Imagine that Joey kicks someone in the shins at the specified time on Thursday evening. Would we be tempted to say that Joey behaved in a way he morally ought not to have? If Joey has the ability to, say, perform some good deed on Thursday evening, but does not, would we say that Joey has failed in his moral obligations? These suggestions sound wrong to me.[30] And the reason stems from a range of quite general assumptions we make about the role of moral requirement and the nature of moral assessment. Plausibly, moral requirements apply to individuals who have some degree of second-order reasons-based control of their first-order motivations. I take this thought to be roughly platitudinous. But if this is the case, it is incongruous to say that moral requirements cannot apply to individuals who have no reasons-based control of their motivations, but that individuals *can* have a moral requirement to φ, even though they have no *reasons-based* control over whether they φ rather than not φ-ing (given that they do not recognize contrastive reason to φ rather than to $\sim\varphi$). But if this is correct, AGENT—in my sense—is a natural result. After all, if Joey cannot see contrastive reason to φ rather than ψ, he cannot control whether he kicks or does not kick *on the basis of reasons*. And hence whether he kicks rather than performing the good deed in question will be the product, at most, of a bare first-order motivation, whim, caprice, or happenstance. And hence AGENT is not *ad hoc*. It is a natural result of a plausible thought concerning the application of moral concepts.

[30] Notice that these cases also demonstrate the problem with INTEND + MSC (whether in its revised iteration or not). In this case, Joey could certainly intend to kick someone in the shins (perhaps as a means to a relevant secondary act), and could certainly see that this had worse expected consequences, and so on. But it remains implausible to say that he is morally required to refrain from so doing, simply because in this case *whether* he will do so rather than not is not under his deliberative control.

8.6.2 Objective moral reasons[31]

To see the second objection, take the following case:

> *Roger*: Roger is a member of a crack bomb-squad and is in the middle of defusing a bomb that, if left undefused, will destroy a large office building, killing dozens in the process (as well as himself). Roger knows that to defuse the bomb he must either cut the red wire or the green wire. But he has no idea which wire he should cut. He can see no reason to cut the red wire rather than the green wire, or *vice versa*. As it happens, the red wire defuses the bomb.

Several things seem natural to say about this case. First, given that the red wire defuses the bomb, Roger morally ought to cut the red wire. In addition, it seems natural to say that, despite his lack of evidence, there is an *objective* moral reason for Roger to cut the red wire: doing so will save dozens of lives.

But there are two problems for AGENT. First, it would appear that AGENT holds that Roger has no objective moral reason to cut the red wire, insofar as he sees no reason to cut the red wire rather than the green wire. Second, and perhaps more importantly, AGENT seems unable to accommodate our commonsense idea of what it means for a person to have an *objective* moral reason rather than a *subjective* moral reason. To see this, consider the following account of objective moral reasons:

> *Objective Moral Reasons* (OMR): if *r* is an objective moral reason reason for *x* to φ at *t*, the true moral theory holds that *r* is a reason for *x* to φ at *t* irrespective of *x*'s epistemic states at *t*.

According to the objection at hand, AGENT treats the class of objective moral reasons as empty. Insofar as AGENT treats moral reasons as indexed to a person's recognition of reasons, they cannot fail to be indexed—at least to some degree—to an individual's epistemic states; whether *x* recognizes a reason to φ will necessarily depend on such states. And if this is the case, there will be no reasons that will satisfy the consequent, and hence no reasons that will satisfy the antecedent.

[31] Thanks to Doug Portmore for Roger's case, and a very helpful exchange on this topic.

I will take these in reverse order. To hold that Agent cannot accommodate the everyday conception of objective moral reasons relies on one way of reading a scope ambiguity in OMR. One might ask: what does "irrespective of *x*'s epistemic states" mean? On the wider scope interpretation, for a reason to be objective, this reason can make no mention of *x*'s epistemic states at all. In other words, if the fact or proposition that constitutes this particular reason includes reference to anything about *x*'s epistemic states, this is not an objective reason. But there is another reading. Call the "narrow scope" reading the suggestion that objective moral reasons can, in fact, include reference to epistemic states, but that the *normative significance* of these reasons are not further indexed to *x*'s recognition, or understanding, or knowledge of, these reasons. In other words, objective reasons are determined by "what the reason-constituting facts about [*x*'s] choice happen[s] to be, and so irrespective both of what [*x*] *take[s]* those facts to be and of what [*x*'s] evidence suggests that those facts might be."[32] This reading takes the true reason-constituting facts as primitive, and hence allows that the "reason-constituting facts" could be determined, in part, by facts about *x*'s mental states, including *x*'s epistemic states. This reading notes only that objective moral reasons are not indexed to what *x* takes these reasons to be, or what *x*'s evidence suggests these reasons are. Put this another way, the narrow reading holds that *r*: "*x* believes *q*," could be (part of) an objective reason for *x* to φ. But it would not, then, further index the significance of *r* to *x*'s recognition of *r as* a reason to φ.

Agent can accommodate the existence of objective reasons if read in the latter way. For instance, it could be that *x* sees a reason to φ, but only a prudential reason to φ. *x*'s evidence suggests no moral reason to φ, and yet Agent allows that there could be an objective moral reason, even a decisive objective moral reason, to φ. Furthermore, the subjective moral significance of φ-ing—which is certainly determined in part by an individual's actual or counterfactual epistemic states—could be substantially different than the objective moral significance of φ-ing. *x*'s evidence might suggest that *r* is a very strong moral reason to φ, despite

[32] Douglas Portmore, *Commonsense Consequentialism* (Oxford: Oxford University Press, 2011), 12. My emphasis.

the fact that the "true reason-constituting facts" present a weaker moral reason to φ. In this way, we can preserve the distinction between *reasons the theory says there are* and *reasons in light of a person's evidence or epistemic states* that is crucial to the distinction between objective and subjective moral reasons at least on the narrow reading.[33] Of course, my view cannot preserve a distinction between objective and subjective moral reasons if OMR is read in the wider scope way. But it seems to me that there is little reason to believe that the conceptual distinction between objective and subjective moral reasons must be read in this way.

Even if Agent can make a sensible distinction between objective and subjective moral reasons, however, we would like to say that Roger has objective moral reason to cut the red wire. I humbly admit that this seems, at first glance, like the right answer. But should we treat this as dispositive reason to accept Trad rather than Agent? Consider the following. Imagine that Roger faced not only two wires, but 1,000. It seems to me that there is no principled reason to say that Roger has objective reason to cut the red wire in the case above, but does not have objective reason to cut the 456th wire (which is the correct one) in the current case. But if this is correct, Roger's case simply becomes (more or less) a version of Wilda's, and hence in suggesting that Roger has objective moral reason to cut the red wire one seems committed to the claim that Wilda has objective moral reason (and hence objective moral requirement) to enter 867–5309. Hence it seems to me that to accept Roger's objective reason to cut the red wire is to commit to a theory of moral requirability (Trad, or perhaps Intend) that we should reject for reasons already explored. This is, of course, not to gainsay the contrary intuition. But I think we should regard this considered judgment a remnant of a theory of moral requirability that, on reflection, we do better to reject.

8.6.3 Bob

Finally, Agent might be under-inclusive for the following reason. Consider:

[33] For more on this topic, see Dale Dorsey, "Objective Morality, Subjective Morality, and the Explanatory Question," *Journal of Ethics and Social Philosophy* 6 (2012): 1–24.

Bob: Bob is a member of a hard-core Satanic cult. Bob believes that drowning children is morally required, and seeks to subject children to drowning whenever possible. In fact, Bob's Satanic indoctrination is such that Bob can see no reason not to drown children.

Now imagine that Bob is faced with the choice of costlessly saving a child from drowning, or letting that child drown. Most would say that Bob is morally required to save the child. But the question now arises: given the structure of Bob's attitudes, could he save the child *as an agent*? Not on my account. After all, *given* his attitudes, if he sees a reason to save the child at all, it will certainly not be a *contrastive reason* (that is, a reason to save rather than not save). If he ends up saving the child, he will not be doing this as a result of deliberation—given his psychology—but as a matter of mere accident. And if Bob cannot save this child as an agent, surely Agent is under-inclusive. We certainly would not want to say that Bob can only be morally required to perform actions he sees reason to perform, given his wacko normative beliefs.

But I think this verdict is not as implausible as might first be thought. First, Bob's status *as* an individual who has deliberative control over thought and action is severely compromised given his indoctrination. One might think, even in Bob's case, that moral categories are somewhat less than appropriate given his utter failure as a deliberative agent. In this way, we might compare Bob to Joey: Bob, like Joey at the relevant time on Thursdays, has no capacity to perform the action of saving the child *by means of reasons*. Although Bob can see contrastive reasons to perform some actions (that is, to drown the relevant child rather than not doing so), a range of actions is simply barred to him *as* a deliberative agent, and this includes the action of saving the child. And hence, like Joey at the relevant time on Thursdays, it seems implausible to say that he could be morally required to save the drowning children, given that *whether* he does so is beyond Bob's capacity to control by means of reasons. Of course, morality will require Bob to perform the morally best of those actions Bob *can* perform as an agent. Nothing about Agent requires us to assign the moral valence of morally requirable actions for Bob on the *basis* of Bob's chilling normative beliefs.

Of course, we certainly want to say that there is something very morally wrong with Satanists who refuse to save children, even if their beliefs are so entrenched that they cannot do so as agents. But there is a very important sense in which we can accept this verdict even if we accept Agent. We would want to say that Satanists behaved badly, for instance, if they *developed* Satanic beliefs—beliefs they could have refrained, as agents, from developing. We can morally criticize them for developing their Satanic principles had they the deliberative option to choose other principles, principles the choice of which on their parts would have led to better consequences, or a greater achievement of whatever other moral factor the substantive theory in question deems relevant. Furthermore, if it is within Bob's deliberative control to develop *new* attitudes, he seems morally required to do so. But assume that the fact that these individuals are Satanists is simply beyond their deliberative control. I find it difficult to criticize such individuals morally speaking, so long as they perform actions that are morally best among those they can actually direct themselves to perform as a result of deliberative action rather than luck or mere happenstance.

Of course, we can still make a number of plausible claims about Bob's refusal. We can say it is bad, regrettable, that Bob did not save the children in question. We will certainly seek to retrain Bob or to get him to change his normative beliefs. We will even *despise* Bob, or seek to *blame* him. It may be that any or all of the *reactive attitudes* are appropriate to direct toward Bob. But I see little ground to be gained in saying that Bob is morally required to perform an action he could not have performed as a result of his own deliberation. To say so, it seems to me, is to confuse two very different questions: whether a particular action at a particular time possesses certain properties that we regard, in the abstract, as morally valuable (given a theory of moral valence), and whether those properties generate moral reasons or requirements for particular individuals given the proper theory of moral requirability (which must necessarily depend on a range of factors about the individual in question, including a capacity to perform such actions on the basis of reasons).

I humbly submit that these are my intuitive reactions, but I am certainly not convinced that they will be widely shared. It may very well be that I have uncovered a seriously implausible verdict of Agent. But the mere fact that some results of Agent may be counterintuitive does not entail that we should reject Agent forthwith. Rather, this consequence should be weighed against the implausible verdicts of competitor theories of moral requirability. In particular, one might wonder whether it is worse or better to accept a theory of moral requirability that entails that on virtually any theory of moral valence, one can be doomed to moral failure simply given the hum-drum fact of cognitive limitations. For my money, even if Agent delivers implausible results in the case of Bob or anyone else, it is far less implausible than simply accepting a view on which we are required to perform actions we will not, given cognitive limitations, perform. And for this reason, it seems to me, we should accept Agent rather than Trad, warts and all.

8.7 CONCLUSION

The fact of cognitive limitations has very serious consequences for virtually every moral theory if we accept a traditional theory (and some non-traditional theories) of moral requirability. I have argued that a fix for this problem is to adopt an alternative: to insist that x can be morally required to φ if and only if x can φ as an agent. This proposal solves the problem of cognitive limitations: in determining what someone can do as an agent, his or her cognitive limitations are *factored in*. In addition, Agent permits of a compelling rationale: Agent is a natural extension of the plausible thought that moral requirements apply to individuals who have reasons-based control of their first-order motivations.

Of course, the points in favour of Agent must be weighed against the points against. Such points there are. However, we must be careful not to jettison Agent too quickly. First, it is likely that many of our considered judgments, especially considered judgments about just what sorts of reasons we have, are heavily influenced by the tacit assumption of Trad rather than a more restrictive theory of moral requirability. Second, and more importantly, on virtually any theory of moral valence, Trad, and a wide range of potential alternatives, seems to succumb to the problem of cognitive limitations. Although I am perfectly willing to admit that there may be as yet unexplored options, in weighing the pluses and minuses, Agent must be taken very seriously indeed.

9

They Can't Take That Away from Me: Restricting the Reach of Morality's Demands*

SARAH STROUD

> No, no, they can't take that away from me.. .
>
> George Gershwin (music), Ira Gershwin (lyrics),
> "They Can't Take That Away from Me" (1937)

The topic of this essay is moral demandingness, and, especially, one distinctive way in which moral theorists have tried to limit it. Barbara Herman nicely limns a familiar dialectic when she writes:

> Once the claim of need is acknowledged, it is not easy to see what, morally, can constrain its demand . . . If it is allowed that the duty [of beneficence] might be even somewhat demanding, might impose real costs on our activities and plans, then given any reasonable account of the need that might trigger beneficence, there is no well-founded stopping point on the demand up to the point of reducing the aid-provider to comparative neediness. (Herman 2001, 227–8)

The scenario Herman describes will be familiar to anyone who has stuck more than a toe into moral theory. You accept some apparently innocuous moral claim or principle about beneficence or the claims of need, and before you can even catch your breath you see that it has taken more or less everything away from you. That toothless-sounding

* I benefitted greatly from presenting earlier versions of this essay at Princeton University, the American Philosophical Association Central Division meeting, the University of Vermont, and the Murphy Institute of Tulane University, as well as from the lively discussion at the Arizona Workshop in Normative Ethics.

principle you endorsed turns out on reflection to threaten to sunder you from your family, your loved ones, your career, your pursuits, and your enjoyments—not to mention almost all your money. Indeed, its demands seem liable to swallow up your whole life. The moral demands you acknowledged seem quickly to have spread like kudzu; the transition from modest shoot to aggressive, voracious weed is vertiginous, even bewildering.

I want to explore one type of response to the possibility that morality imposes extreme demands on us: one way of trying to stop that vertiginous transition. The arguments I want to examine all claim, contra Herman's gloss above, that there *is* a "well-founded stopping point on the demand": that we have principled reasons for thinking the elevator cannot go express all the way to the basement in the way it seems to. With a nod to George and Ira Gershwin, I shall call arguments of the kind that interest me here "they can't take that away from me" arguments, for they all claim to identify something which morality cannot legitimately ask me to give up—something which morality cannot take away from me. A successful argument of this kind would set a definite limit to morality's demands, by establishing a "hands-off" or "no-fly" zone which moral requirements could not penetrate. It would guarantee that no moral principle could end up swallowing up our whole life. I think this represents a distinctive and interesting strategy for resisting extreme moral demands—a strategy that so far as I know has not previously been singled out and given its own treatment—and I want to explore its prospects in this essay. Careful examination, I shall suggest, reveals those prospects to be less bright than we might have hoped, and at the end of the essay I propose that we look for a different way of objecting to the arguments which threaten to take us express to the basement.

9.1 POTENTIAL SOURCES OF EXTREME MORAL DEMANDS

It is reasonable to begin by reviewing why we might be worried by the possibility that morality presents us with extreme demands. (If there is no real chance of moral demands spreading past their designated flowerbed, there is not much need to prepare a counter-response to the kudzu.) Where could such extreme moral demands come from? I want to highlight three potential sources of extreme moral demands, in descending order of strength. As we shall see, it is worrying how little it seems to take for the kudzu to take root in our garden, with predictable consequences.

Consequentialism. It is easy to see how consequentialism could place demands on us which we would consider to be extreme. Consequentialism, in its traditional and most standard formulation, bases the rightness and wrongness of actions wholly on the comparative *agent-neutral* value of states of affairs. Therefore, in considering whether I am morally required to do A, consequentialism will treat any benefits or costs *to me* of doing A simply as benefits or costs *to someone* of my doing A. And because traditional consequentialism is maximizing, requiring agents to bring about the *best* states of affairs they can, it will always be wrong to fail to provide someone with a benefit whenever we could do so at a cost that from an agent-neutral perspective is smaller than the magnitude of the benefit we could provide.

Consequentialism, it seems, requires us to do literally *everything we can* to benefit others or to alleviate bad states of affairs, even if our own lives are thereby ruined. It asks, to put it mildly, a lot of us; it is likely to condemn much of what we currently do as wrong. Its demands are not absolutely limitless, but they are limited only by a) the limits on your powers to aid others or otherwise bring about better states of affairs, and b) the constraint that you cannot be required to bring about a benefit when the agent-neutral magnitude of the cost to you would be greater than that of the benefit in question. Consequentialism, then, is definitely a possible source of extreme moral demands. But we could evade this potential source of extreme demands by rejecting consequentialism (which we might well do for reasons independent of its extreme demands). So in this case there is a clear route to preventing the kudzu from taking hold.

There are, however, other potential sources of extreme demands that are more theoretically modest, and therefore harder to keep from taking root in the garden.

Singer's principle. According to a principle Peter Singer famously advocated, "if it is in our power to prevent something bad from happening, without thereby sacrificing anything of comparable moral importance, we ought, morally to do it" (Singer 1972, 231). This principle is distinct from (and indeed weaker than) consequentialism, but nonetheless broadly consequentialist in spirit. Its antecedent hypothesizes an ability on our part to prevent a bad state of affairs, and Singer's main gloss on what he means by the key qualification "without sacrificing anything of comparable moral importance" is "without causing anything else comparably bad to happen" (Singer 1972, 231). As I read the principle, Singer's moral "ought" therefore reposes principally, as does consequentialism, on a comparison of the agent-neutral value of states of affairs.

This feature makes Singer's principle extremely demanding, even if it is weaker than full-fledged consequentialism and indeed sounds innocuous and hard to reject.[1] In order to see why Singer's principle is nonetheless very demanding, it is important to focus on the fact that Singer's notion of "comparable moral importance" does not track what we might call *importance to me*. Something can be, in our ordinary way of speaking, very *important to me* without being important in the sense Singer intends. For example, it might be extremely important to me to be at my daughter's graduation piano recital, or to finish a paper that I have been working on for a long time, or to run well in the marathon for which I've been training; but whether or not these events occur is of comparatively minor *moral* importance in Singer's sense. From an agent-neutral perspective, *someone's missing her daughter's graduation piano recital* is not of comparable importance to *someone's dying prematurely of malaria*. So according to Singer's principle, if ever I can prevent an instance of the latter by bringing about an instance of the former, I am obligated to do just that.

Singer's principle, then, can easily ask me to give up something that is very *important to me*. This needs to be underlined, because the example Singer uses to motivate his principle—the famous pond case—conveniently obscures this fact. In the pond case, you are the only person who can rescue a child drowning in a shallow pond right in front of you—at the cost of having to get your suit dry-cleaned. Singer takes it

[1] Singer's principle does not quite amount to an embrace of consequentialism across the board, for at least two reasons. First, unlike consequentialism, Singer's principle confines its demands to preventing bads: it does not demand, as does full-fledged consequentialism, that we increase the magnitude of goods. In addition, Singer's principle seems not to require you to prevent a bad B if the cost C to you of doing so is lower than, but nonetheless *comparable to*, the magnitude of B (that is, if the difference between the respective magnitudes of B and C is small). A strictly maximizing theory such as traditional consequentialism would demand that you prevent B whenever C is of lesser magnitude than B, no matter how small the difference between the two magnitudes.

as obvious that you must rescue the child, and proposes his principle as the appropriate generalization of that verdict. But in the pond case, what you are being asked to sacrifice is not even something *important to you*, let alone important from the point of view of the universe, as it were; and it is actually the latter which is operative in Singer's principle, not the former.

It is worth noting, too, that while Singer discusses the (depressing) implications of his principle for where you are required to direct the money you have, he does not underline that his principle would also call for you to work more, or indeed to change to a more lucrative career, in order to have more to spend on aiding the hungry and needy. (His principle will require this as long as it is true that by doing this you could prevent something bad from happening that is of greater agent-neutral significance—that is, which has a greater magnitude of agent-neutral value or disvalue—than the loss to you of making such a drastic career or lifestyle change.) We should not labor under the illusion that the demands of Singer's principle are limited to how we spend our excess disposable income.

To sum up, Singer's principle may sound plausible and unobjectionable when you first read it. But when you consider its implications, including those which Singer does not explicitly draw, you may well conclude that it is not so plausible after all. You might therefore refuse to give it a home in your garden.

Cullity. Garrett Cullity offers an argument in Part I of his book *The Moral Demands of Affluence* (Cullity 2004, hereafter cited simply by page number) which builds on the basic instinct behind Singer's but which is, crucially, more theoretically modest in its commitments. This makes it especially dangerous for present purposes.

Cullity's argument has two main elements. Like Singer, Cullity takes as his "base case" opportunities for what he calls "direct" rescue, such as we saw in Singer's pond case. Cullity makes a modest proposal, which I label "BC" (for "base case"), about cases like these:

> (BC) "If I could easily save someone's life right in front of me [at negligible cost to myself] . . . then . . . not doing so would be wrong" (11; text in brackets from 10).

Cullity adduces the following line of thought in favour of (BC). In cases in which you have the opportunity to effectuate a direct rescue, 1) you have a powerful reason to save the person, a reason stemming from his strong interest in being saved. 2) You have no powerful countervailing reason or defeater: only the very trivial or minor cost to you of doing so, which is not a *powerful* countervailing reason or defeater. 3) Given the enormous disparity between what is at stake for him and what is at stake for you, to refuse to save in such a case would manifest an insufficient regard for others' interests. It would constitute a *failure of beneficence*, and would for that reason be wrong.

It is important to keep in mind that (BC) is limited, first of all, to cases of "direct rescue," as (BC) defines them, and second, to cases in which the cost to you of effectuating a "direct rescue" is slight or negligible. Crucially, Cullity seems to mean by the latter that you yourself would assess the cost to you as trivial—as was true in Singer's original pond example, but was not guaranteed to be the case in all situations falling under Singer's more general principle. This fact makes Cullity's starting point considerably more modest than Singer's principle. Unlike Singer's principle, (BC) does *not* entail that you must save at the cost of missing your daughter's piano graduation recital, or at any other cost which you would consider at all grave or serious.[2] Because of the modesty of his starting point, it seems difficult not to let Cullity's argument get a foot in the door.

However, Cullity argues that even this very modest foothold marks out a clear path toward extreme moral demands. For (BC), whose scope is limited in a number of respects, functions only as the "base case" in Cullity's argument. In the second part of his argument, Cullity subjects his apparently uncontentious "base case" to a kind of mathematical induction that will take us far beyond the boundaries of cases of direct rescue. Cullity calls this second part of the argument "the life-saving analogy": it aims to apply the moral we took from the base case to the seemingly quite different issue of our obligation to give money to aid agencies. Through deployment of his "life-saving analogy," Cullity argues that giving trivial amounts to aid agencies is in morally relevant respects analogous to rescuing children who are drowning in shallow ponds right in front of you. If he is right, we can drop one of the two limitations on

[2] Of course, Cullity is not claiming that in those cases you ought not to save—he is simply not saying, with regard to those cases, that you must.

(BC) that we noted above, namely the one which restricted its demands to cases of "direct rescue."

Cullity wants to convince us that situations in which you have the opportunity to give to an aid agency which saves lives possess the same basic moral structure that generated the conclusion about cases of direct rescue which (BC) articulated. That structure, recall, had two elements: a strong reason to help, and a lack of sufficiently powerful countervailing reasons not to. Cullity says that both are equally present here. When you have the possibility of giving to an aid agency which saves lives, you are confronted with the same kind of powerful reason to help that is present in a case of "direct rescue," namely the very strong interest the beneficiaries have in being saved. And Cullity argues that there are no new countervailing reasons present here that are sufficiently powerful to alter the moral verdict we drew in the base case: neither distance, nor immediacy, nor the fact that you do not even know *whom* you are saving, genuinely countervails the strong reason to help with which you are presented. So for Cullity, these cases possess essentially the same moral profile as cases of direct rescue. Cullity concludes that "not giving to aid agencies is wrong: it is wrong for the same reason that it is wrong not to save a life [right in front of you] when you could easily have done so" (11). Both of these "exhibit…a failure of beneficence, and that makes [both of them] morally wrong" (32).

Cullity's conclusion about our obligation to give to aid agencies sounds modest. Since it is a kind of generalization of (BC), it is, like (BC), limited to cases in which the cost to you of giving is negligible or trivial by your own lights. Indeed, to highlight this, we might formulate his conclusion (which I call "AA" for "aid agencies") in a way that explicitly takes over the relevant caveat from (BC):

> (AA) If I could easily and at negligible cost to myself give to aid agencies that save lives, then not doing so would be wrong.

We noted above that in moving from (BC) to (AA) we dropped (BC)'s limitation to cases of direct rescue, since (according to Cullity) that limitation played no role in making (BC) true. But we cannot similarly drop (BC)'s limitation to cases in which the cost to you is trivial or slight, since that was an essential element of the original case for (BC). (AA), then, is weaker than the principle from Singer which we discussed above, as it will never ask us to make a donation of money or time when doing so would generate a cost to us that is non-trivial by our own lights. It might therefore be mysterious why I have classified Cullity's argument as a possible route to *extreme* moral demands.

I so classify it because Cullity shows that if we apply even this very modest principle *iteratively*, it will yield extreme demands.[3] For even if the cost to you of giving, or saving, is trivial in any individual case, you will in fact face many, many situations in which you *could* give some time or money to an aid agency at trivial cost to yourself. Because (AA) is sensitive (on the iterative reading) only to the *marginal* or *incremental* cost to you of your next small donation, it will, it seems, require you to give in each of those many situations. Cullity's argument thus appears to lead to the conclusion that I must continue to give small increments of time or money to aid agencies until either they have run out of potential beneficiaries, or my own condition has been so much reduced that the very next increment I could give would be so costly to me as to excuse me from being required to give it under (AA). According to this Extreme Demand, as Cullity calls it, we affluent people act wrongly unless we lead intensely altruistically-focused lives in which we keep personal attachments and pursuits to whatever strict minimum is psychologically necessary in order to keep our beneficent activities going. The Extreme

[3] As Frances Kamm and others have pointed out to me, Cullity's argument here structurally resembles Warren Quinn's "paradox of the self-torturer" ("The Puzzle of the Self-Torturer," *Philosophical Studies* 59 (1990), 79–90). Quinn's self-torturer has repeated opportunities to turn up a pain dial an infinitesimal amount in return for a payoff that makes each such individual transaction advantageous. Note, however, that even if each turn of the dial increases the self-torturer's pain so little that he hardly notices the increase, he will eventually (after some number of turns of the dial) be in absolute agony. If we suppose that the man would never have chosen to put himself in such agony for the aggregate return he receives, we reach the puzzling result that although each turn of the dial is apparently rational when viewed as a discrete action, the larger action of turning the dial that number of times is not rational.

That the man in the puzzle can apparently be rationally driven to self-torture through this process seems to suggest that something must be wrong with the argument that one ought to turn the dial this time because the cost of doing so is outweighed by the benefit, a point that could also be applied to Cullity's strategy of argument in Part I. Those who pointed out this structural similarity to me thus advanced it as a general reason to think there is something fishy about Cullity's argument. While I appreciate that intuitive reaction, which is nicely bolstered by pointing out the resemblance to Quinn's self-torturer, I would hope that we might be able to move beyond the conviction that there *must* be something wrong with a certain argument (even if we cannot say, or even see, what it is) to an at least tentative identification of *what* is wrong with that argument. I take some steps in that direction at the end of the essay.

Demand “allow[s] me to spend practically no time or money on my own personal fulfillment” (80).

As we saw, Cullity's argumentative route to the Extreme Demand started with a modest claim about his base case, (BC), and then used the life-saving analogy to extend its implications to (AA). Because the base case, at least, is significantly more modest than either consequentialism or Singer's principle—and, indeed, (AA) seems to be as well—Cullity's route to extreme moral demands will be *prima facie* harder for us to resist by simply rejecting the starting point. Cullity himself is obviously worried that many will be inclined to reject, not the base case, but the life-saving analogy; he devotes much of Part I of his book to attempting to block the exits from that analogy, so that those of us who accept the base case will be forced to feel the pull of the Extreme Demand. If we set aside possible worries about the life-saving analogy, Cullity's route to extreme moral demands seems the most dangerous of the ones we have considered.

9.2 “THEY CAN'T TAKE THAT AWAY FROM ME” RESPONSES

We have looked at some potential sources of extreme moral demands. In response to possibilities like these, some philosophers have attempted to erect a bulwark against the potential for extreme and voracious moral demands by identifying some thing which morality allegedly cannot take away from me. This response is a version of, but is more specific than, the general counter-claim that a given moral theory or principle is “too demanding.” To offer a “they can't take that away from me” response is to claim that a principle or theory which would deny us *x* in particular is too demanding: it is to contend that morality cannot legitimately require that we sacrifice—cannot leave us without—*x* in particular.[4] Let us look at some manifestations of this form of argument.

Williams. Some of Bernard Williams' remarks about “ground projects” could well be interpreted as making a claim of the “they can't take that away from me” variety. As we have seen, consequentialism

[4] “They can't take that away from me” responses differ in this respect from some other well-known maneuvers to mitigate moral demands, such as the “agent-centered prerogative” introduced by Samuel Scheffler (1982) and the kind of “fair share” view defended by Liam Murphy (2000). Neither of those approaches claims that there is a particular *x* which morality cannot legitimately take away from me.

requires you to forgo the pursuit or the satisfaction of one of your projects whenever, were you to do so, you could bring about a state of affairs that is even slightly better overall from an agent-neutral perspective. As Williams puts it in his critique of utilitarianism, "[your] own substantial projects and commitments come into it, but only as one lot among others" (Williams 1973, 115). This means, of course, that they are liable to be outweighed. If you complain about this, the utilitarian moralist has an answer: the cost to you of having to give up your project has already been taken into account, and has been overridden. All of this is familiar.

In a passage that is often overlooked, Williams makes the rather surprising concession that "in the case of many sorts of projects, that is a perfectly reasonable answer" (Williams 1973, 116). But—as is better known—Williams insists on one subclass of projects for which (he maintains) this kind of answer is *not* acceptable. "A man," he writes elsewhere, "may have . . . a *ground* project or set of projects which are closely related to his existence and which to a significant degree give a meaning to his life" (Williams 1976, 12); and a demand to give up a *ground project*, Williams strongly intimates, cannot be justified in the way just described. He writes:

> How can a man . . . come to regard as one satisfaction among others, and a dispensable one, a project or attitude round which he has built his life, just because . . . that is how the utilitarian sum comes out? . . . It is absurd to demand of . . . a man, when the sums come in from the utility network . . . that he should just step aside from his own project. (Williams 1973, 116)

As so often, it is not obvious what point Williams is making here. On one way of reading Williams—and I think he is often read this way—he is claiming that morality simply cannot ask me to give up a ground project: that it cannot take a ground project away from me. That would indeed be one version of a "they can't take that away from me" argument. If that is what he is saying, though, it seems to be false. Consider the case of Frau Paul.[5]

[5] The following is adapted from Funder (2004), chs. 21–23.

9.3 INTERLUDE: FRAU PAUL

Frau Paul gave birth to her first child in a Berlin hospital in January 1961. There were serious complications, and after the birth the child spat up blood and was unable to feed. Doctors at her hospital in the eastern part of Berlin could not figure out what was wrong. At a hospital in the western part of the city, however, they were able to make a diagnosis right away. The baby had suffered a ruptured diaphragm during delivery—a life-threatening condition requiring immediate surgery. Baby Torsten recuperated at the Westend Hospital until July 1961, at which point he was discharged to his parents' care. He needed special medicines and formula, however, which Frau Paul would fetch at the Westend Hospital once a week.

During the night of 12–13 August 1961, the Berlin Wall appeared. When Frau Paul awoke she saw that she was cut off by barbed wire from the western part of the city. She requested permission to cross to West Berlin to receive the formula and medicines Torsten needed, but permission was denied. Torsten began to spit up blood again, and his condition deteriorated. One night his parents took him in desperation to the emergency room of the east end hospital. The doctors told them to go home and rest while they examined Torsten. When Frau Paul returned the next morning, her son had vanished. During the night, realizing that they could do nothing to help Torsten, the doctors had managed to spirit him across the international border to the Westend Hospital where he could be cared for properly. Their action saved his life.

Once back at the Westend Hospital, however, Torsten continued to suffer many complications, and his condition did not improve. He underwent four surgeries and was fitted with an artificial esophagus and diaphragm. Because it was feared that he would die, an emergency christening was held when Torsten was nine months old. Frau Paul was granted a day pass to cross to West Berlin to attend the christening, but her husband was not: the Stasi were afraid that if both parents were granted passes they would simply stay in the West. Frau Paul came back to her home and her husband at the end of the day thinking she might never see Torsten alive again.

Over the coming months Frau Paul and her husband came into increasing contact and, eventually, collaboration with a network of East and West Germans who were trying to help people trapped behind the

Wall cross to the West. This group exploited whatever weak points they could find in the international border, having people cross platforms at the Ostbahnhof railway station, for instance (you could do this with a West German passport, which you would have been lent by an obliging West German who resembled you). The leader of this little band was a student, Michael, who was from the eastern part of the country but had been studying at the Free University when the Wall went up; he decided to stay in the West, and was now working to help his parents and others who wished to cross over. Frau Paul and her husband did not think of themselves as political people or as opponents of the regime, but they began to contemplate seriously escaping to the West to be with Torsten. They wanted more than anything else to see him and care for him; as it was, their only news was an occasional letter from the hospital staff.

They helped Michael and the band in whatever way they could up until the day for which their own escape was planned, in February 1962. In preparation for their departure, they gave away their car, and discreetly sold some of their possessions. But their escape via the Ostbahnhof had to be aborted at the last moment: the Stasi had cottoned on to that particular escape strategy, and had arrested the previous group of East Germans aided by Michael's band earlier that same day. Frau Paul and her husband burned their borrowed West German passports and definitively gave up the idea of crossing to the West.

Even though they had given up any hope of leaving themselves, they continued to assist Michael and the other students over the next year. The students began working on a new escape route: a tunnel under the Wall. But the Stasi discovered the tunnel on its completion, and from that day on Stasi men followed Frau Paul in the street. One day in February 1963 some plain-clothes Stasi agents plucked Frau Paul off the street in broad daylight and shoved her into a large black limousine. They took her to Stasi HQ and interrogated her for 22 hours. The following exchange took place during one of those sessions.

"We understand," said the Stasi interrogator, "that you have a son in enemy territory." Frau Paul agreed. "Would you like to see him?" Frau Paul's heart started to pound; she wondered what the man was driving at. "That can be arranged," the Stasi man went on. "If you would like to visit your son in enemy territory, we would ask only that, while you are there, you arrange to meet up with your young friend Michael. The two of you could go for a stroll. For instance, in the grounds of

Charlottenburg Castle." Frau Paul was confused. "You can leave the rest to us," the Stasi man said. Then Frau Paul understood: "they were going to use me as bait in a trap to kidnap Michael." "Frau Paul knew Michael would trust her to come to a meeting in the park, and when they came to bundle him into a vehicle she would have to turn her back and walk away" (Funder 2004, p. 220). Frau Paul understood that once she did this, the Stasi would feel they "owned" her and would therefore permit her to cross over for regular visits with Torsten.

Frau Paul, however, refused the deal the Stasi man was offering her, and as a result spent over four years in a secret East Berlin prison for political prisoners (whose horrors are chillingly described in *Stasiland*). The Stasi never did get its hands on Michael. Torsten, meanwhile, was brought up by the staff at the Westend Hospital until he was five. When he finally came home to his parents, he did not recognize them, and he addressed them with the formal *Sie*. He had survived, and had been reunited with his parents, but—as is clear from this book written many years later—his relationship with his mother was forever marked by her absence from his life during those difficult early years.

There are many things we could take from this rich real-life case. For present purposes I wish to propose the following trio of takeaways:

1) Being able to see her son qualifies as one of Frau Paul's ground projects.
2) In present circumstances, Frau Paul can only pursue or continue that ground project if she does *x*. If she does not do *x*, she will have to abandon pursuit of that ground project—forever, for all she knows.
3) It would be wrong for Frau Paul to do *x*.

By "*x*" I mean, of course, betraying her friend to the Stasi in the way the Stasi agent proposed. While I think we have tremendous sympathy for Frau Paul on hearing this story, I do not think we waver in our judgment that Frau Paul must not betray Michael, that to do so would be a grave moral wrong. If this is so, however, then Frau Paul is morally required to abandon pursuit of one of her ground projects in the present situation. It would be wrong of her to pursue it, given what that would entail; in that sense she is required to give it up. And this is true even though the project itself is morally inoffensive (to put it mildly).

If Williams was saying that morality simply *could not* ask an agent to give up a ground project, period, then Williams was, I think, wrong. (He is also wrong if he was saying, more modestly, that morality cannot ask an agent to give up a ground project that is morally unobjectionable in itself.) But there are other ways to understand Williams' point. Notably, one could read him, not as claiming that it is unacceptable for a moral theory to ask me to give up my ground project, period, but as making the narrower point that it is unacceptable for morality to ask me to give up my ground project *for a certain sort of reason*.[6] Since, in the Frau Paul case, we are not appealing to the kind of reason which Williams finds insufficient or even objectionable, it will not follow that morality's demand *there* is illegitimate. The moral requirement not to do *x* which is generating the demand that Frau Paul give up her ground project is, it seems, some sort of deontological constraint against betraying your friend to a brutal regime. It is very different from the kind of reason for which utilitarianism would or could ask an agent to give up a ground project.

One might find support for the idea I am attributing to Williams in the very contrast between our reaction to the Frau Paul case and our reaction to the idea that utilitarian morality might require us to sacrifice all our ground projects. For *our judgments about whether an extreme demand is acceptable seem to be sensitive to the reason for which the demand is being imposed*, just as this point would predict: we may be prepared to accept that reason R1 generates a genuine moral demand D without being willing to accept that reason R2 generates the same demand. Moreover, consulting again our intuitive reactions, the reason utilitarianism offers seems to be one whose sufficiency we are more inclined to doubt. Williams does not spell out his objection to that reason, and there is a range of possible interpretations. On perhaps the narrowest, Williams may simply be suggesting that reasoning which treats *my* ground project simply as *someone's* ground project cannot provide sufficient grounds for asking me to give it up. While this idea seems plausible to me, note that if we read Williams this way, his point has force against consequentialism (which of course does treat *my* ground project just as *someone's*), but not against any view which eschews such reasoning, such as

[6] In that case Williams' claim would no longer strictly speaking be of the "they can't take that away from me" variety, since morality could indeed take even one of my ground projects away from me.

Cullity's.[7] On this interpretation, Williams' point, however plausible, cannot neutralize the case Cullity mounts for the Extreme Demand. Alternatively, we could interpret Williams as making a somewhat broader claim: perhaps he is suggesting that reasons of *beneficence* cannot require me to give up a ground project. In that case his argument does engage with Cullity's Extreme Demand, which asked us to abandon all but the most minimal set of ground projects in the name of beneficence. On the other hand, this view would seem to owe us an explanation of why it is objectionable for beneficence in particular—as opposed to other kinds of moral considerations—to compel such a sacrifice.[8]

What is clear, at any rate, is that Williams here attempts to block the force of moral reasons by appealing to something which is not itself a moral reason in the narrow sense, namely the significance to an agent of his ground projects. Williams is suggesting that something exogenous to moral reasons can limit or block the force of the latter.

9.3 OTHER EXAMPLES

Herman. Our second illustration of a "they can't take that away from me" argument takes a quite different tack. In the paper from which I quoted earlier, Barbara Herman discusses the demands of beneficence from an avowedly Kantian perspective. Her way of blocking a kudzu-like takeover of our lives by those demands is to argue that whatever is necessary for my own moral development is simply not available for distribution to others:

> At issue is why, in the face of need, a human life is not to be regarded as a warehouse of potentially distributable skills and possessions... Many of the resources that support successful or developing agents cannot be made available for use by others without undermining the agency from which they would be withdrawn... *If* education [for instance] is a necessity... for effective agency, then it (or the wherewithal to support it) is not available for distribution to others. (Herman 2001, 241, 244)

[7] Recall that Cullity's (BC) and (AA) effectively privileged our personal projects, by asking us to take action only if the effect on our projects would be minimal.

[8] Liam Murphy persuasively and insistently presses this general idea (not specifically against Williams) in ch. 3 of his (2000).

Herman's strategy—similar, as we shall see, to Cullity's (to be discussed)—is to use resources *internal* to morality in the narrow sense to place limits on the scope of the demands of beneficence. Herman rejects the strategy of limiting the latter by appealing to something exogenous to morality narrowly construed: she says that if we view the pull toward meeting need and the conviction that morality must leave our lives livable as "reflecting independent values," then our only option will be "some sort of balancing. But it is hard to imagine striking a balance that will not seem or be arbitrary" (228). By contrast, Herman thinks that the very logic of the source of the demands of beneficence implies limits to their scope.

For Herman, the demands of beneficence stem from the portion of a Kantian theory which holds up certain ends—namely my own moral perfection and the happiness of others—as obligatory. From this Kantian perspective, it would exhibit a kind of incoherence for morality to ask me to give up something that is necessary for my own moral development in order to advance the happiness of others: that would be to pillage one obligatory end of moral agency (my own moral perfection) in order to advance another end with the same status, namely the happiness of others. From within the very moral perspective which gives rise to the demands of beneficence, it would be self-defeating for morality to try to take this away from me. We might wonder what, concretely, will fall within the "protected zone" which Herman finds within the Kantian moral framework. That is, *what* is necessary for my own moral development and therefore not available for distribution to others? Herman suggests, unsurprisingly, that education will qualify, and also—more surprisingly—that even some enjoyment will qualify (on the grounds that I will not become a developed moral agent if no enjoyments propel me forward).[9]

Herman's argument merits more attention than I have space to give it here. I do, however, want to note several limitations on her argument which seem to me serious, and which hold even if her argument is successful on its own terms. First, Herman's argument shows at best that *Kantian morality* implies certain limitations on the demands of beneficence. That is, if Kantian moral theory is correct, we can legitimately

[9] It is interesting to note the similarity of this idea to Williams' point that ground projects are a condition of my going on living and therefore my having any reasons for action at all.

restrict the reach of the demands of beneficence in the way Herman describes. But that antecedent, in my view, greatly limits the dialectical usefulness of Herman's conclusion. Kant's conception of our obligations of beneficence appears non-maximizing on its face; so it is likely to come as no surprise when we are told that the Kantian perspective would not require that we give over our lives to good works. Moreover, to offer Kantian morality in support of the conclusion that the demands of beneficence are not after all limitless may seem to be a case of attempting to justify the more certain by the less certain. We may well have more doubts about whether Kant's moral system is overall correct than we do about whether the demands of beneficence could really be virtually limitless.

A different worry about Herman's argument concerns not its specifically Kantian character, but what might be termed its *moralistic* character. Herman argues that we cannot be obligated to give certain things up because for us to give them up would be to default on another important moral obligation. Herman thus offers what might well be considered a moralistic defense of our being left in possession of certain goods, insofar as her argument cites other *moral obligations* we are under as the reason why we are entitled to keep those goods out of general circulation. On Herman's picture, in sum, I am allowed to hold on to those goods only because I need them for other moral purposes. Such an argument seems not, however, to capture the full intuitive force of the thought that we may legitimately object to putative demands of beneficence that are so extreme as to threaten to take over our lives. Apart from their impeding our fulfillment of other moral duties, such demands also seem objectionable in their own right and for a different sort of reason. The case for its being legitimate for us to insist on being left space within which we may pursue, attain, and cherish certain central human goods does not seem to depend solely on their efficacy in promoting specifically moral purposes.

Cullity. The argument that Garrett Cullity offers in Part II of his book may be able to transcend the limitations just noted. In his Part II he offers a highly sophisticated version of a "they can't take that away from me" argument, in response to the argument he made in Part I for virtually limitless demands of beneficence. His counter-argument exploits what he calls the presuppositions of beneficence: according to Cullity, the very things that beneficence demands entail limits on what beneficence can demand. These limits imply that there are certain things that you cannot be asked to give up in the name of beneficence, and therefore that the Extreme Demand is false. In sum, "beneficence requires us to accept that other people's interests give us compelling moral reasons for acting in their favour; but in accepting this we are making presuppositions from which it follows that acting out of partiality towards our own interests is not wrong" (128).[10]

An argument of this style is particularly attractive because it does not involve our occupying—and defending—some independent normative perspective from which we reject the Extreme Demand as "demanding too much." The modesty of the theoretical presuppositions of Cullity's response mirrors the modesty in the theoretical presuppositions of Cullity's affirmative argument *for* strong demands which made it so dangerous. In particular, Cullity's response does not grow out of any specific (and therefore contentious) moral theory, as did Herman's. Rather, it is supposed to grow organically out of the very convictions about beneficence with which we began.

In order to discover what beneficence *cannot* require, Cullity's idea is that we should begin by probing our conception of what beneficence *does* require. In particular, we should probe *which interests* support or create obligations of beneficence. For while it seems uncontroversial that beneficence calls on us to advance others' interests, it is perhaps less noticed that beneficence does not in fact call on us to advance just *any* interest another person might have. (We shall return to this point shortly.) Cullity's suggestion is that it will be fruitful to focus our attention on the range of interests which satisfy the following condition:[11]

Condition C

Let us say that *x is in the range of Condition C* or *satisfies Condition C* when a person *S*'s interest in obtaining or retaining *x* can generate a requirement of beneficence on others to aid *S* in obtaining or retaining *x*, at least when those others can do so at minimal cost to themselves.

[10] His argument, like Herman's, would thus be better labeled "*beneficence* can't take that away from me," since it leaves open whether some *other* source *could* indeed ask you to give up the things in question.

[11] This condition is nameless in Cullity, but for ease of subsequent reference I have dubbed it "C" (for "condition").

What interests fall within the range of Condition C? We saw one such interest in the examples of rescue with which we started. Our moral judgments in Singer's pond case and, more generally, in Cullity's cases of "direct rescue" show that we think a person's interest in *remaining alive* is one which can indeed make it obligatory for us to aid her when we can do so at minimal cost to ourselves. Cullity emphasizes, however, that if we consult our convictions about beneficence we will discover far more things than a person's interest in life itself which can ground requirements of beneficence on others. Suppose for example that by expending minimal effort you could reunite a family, or make it possible for a talented student to receive a musical education (136; he returns often to these two examples). Surely those are also interests which could make it obligatory for you to aid at minimal cost to yourself: to refuse even to lift a finger to make such a difference in someone else's life would clearly also constitute a failure of beneficence on your part.

Even if the range of interests satisfying Condition C is wide, it is not universal. For Cullity holds that some interests do not generate any requirements of beneficence at all on the rest of us, even if the cost to us of advancing those interests would be very slight. He gives two main types of example. On the one hand, he holds that extremely trivial interests do not generate requirements of beneficence on the rest of us. Suppose for instance that Susan would very much like a chilled martini right now. Even if we grant that Susan genuinely has an interest in a chilled martini, this fact does not seem to place any obligation of beneficence on us—even if we could easily help Susan obtain one (151). The second type of case concerns not trivial interests, but what we might term immoral interests. According to Cullity, a gangster's interest in getting his gun unjammed, for instance, generates no requirement of beneficence on the rest of us to aid him in this endeavor (138, 140–1). Cullity proposes the following general principle[12] to capture such cases.

Principle P

If *x* satisfies Condition C—that is, if a person *S*'s interest in obtaining or retaining *x* can ground requirements of beneficence on others to aid her in obtaining

[12] As before, Cullity gives no name to his principle; I have given it a toy name ("P" for "principle") for ease of subsequent reference.

or retaining *x*—then *x* must be something it is permissible for *S* to seek or to hold on to.[13]

The contrapositive of Principle P explains our verdict in the gangster case. The gangster's interest in getting his gun unjammed places no obligations of beneficence on the rest of us because he has that interest only as a part of a larger activity which is morally impermissible, namely seeking the death of his target. Principle P also sounds plausible when described in general terms. As Cullity puts it, "your interest in obtaining what it is *wrong* to have cannot be a good reason for my being morally required to help you" (139); "only interests in obtaining goods that it is not *wrong* to have can properly ground claims on our help" (145).

Having motivated Principle P, Cullity will use it to draw conclusions that go well beyond the gangster case. His idea is to use Principle P to deduce limits to the demands of beneficence: limits that rule out, in particular, the Extreme Demand. Recall that Principle P says that all the things *x* that meet Condition C are things it is morally permissible for a person to seek or to hold on to. In supposing that *S*'s interest in getting or holding on to some thing *x* genuinely calls for our beneficent help, therefore, we are presupposing that *x* is something it is permissible for *S* to seek or hold on to. But if *x* is something it is permissible to seek or hold on to, *x* must be something it is permissible for *me* in particular to seek or hold on to. This importantly limits morality's demands: we can now legitimately claim, for any such *x*, that morality cannot take *x* away from me. There can be, in short, no genuine obligation of beneficence which requires me to relinquish *x*, for any *x* which is in the range of Condition C. If I am required by beneficence to give *x* up, then it is *not* permissible for me to hold on to *x*. But in thinking that *x* grounds requirements of beneficence we are committed (at least according to Principle P) to the idea that it *is* permissible to seek or to hold on to *x*.

According to this argument, there would be a kind of internal incoherence in holding that beneficence can require me to give up the kinds

[13] This is actually a slight emendation of Cullity's principle, since he speaks of things it is permissible for *S* to *have*. I am not sure I understand what it means to say that it is permissible for a person to *have* something, since his having it seems to be a state of affairs rather than an action, and I take deontic terminology like "permissible" to apply to actions. We can preserve the spirit of Cullity's principle by changing the verb "have" to ones which more clearly indicate *actions* (that can then be assessed as permissible or not).

of goods which I am required by beneficence to aid others in obtaining or retaining. The very convictions about beneficence with which we started—and which, when extended by "the life-saving analogy," threatened to entail extreme demands—in fact rule out at least certain kinds of extreme demands. According to Cullity, this argument rules out in particular our being required by beneficence to lead the sort of altruistically focused life that the Extreme Demand requires. For the interests that we think ground requirements of beneficence include people's interests in their cherished projects and relationships, even when these are *not* altruistically focused. (We saw this in the examples of family reunification and pursuing a musical education that we noted earlier.) Indeed, Cullity's argument implies not just that beneficence cannot require us to lead an altruistically focused life, but that it cannot require us to give up goods without which we would be significantly worse off (where "significantly" means sufficiently worse off that others would be required to help us get those goods back if they could do so at minimal cost to themselves). So by exploiting Principle P we can deduce significant *a priori* limits to the reach of morality's demands, and to the demands of beneficence in particular.

While I find Cullity's strategy of argument incredibly ingenious, I want to raise three worries about whether his argument ultimately goes through.

First Cullity's argument draws conclusions about the moral entitlements of a potential *benefactor* from moral claims about the potential *beneficiary*. His point is in effect that what is sauce for the goose is sauce for the gander: if other people's pursuit of *x* can ground requirements on me to help them, then it must be permissible for me to pursue *x* as well. But I am not sure we are entitled to turn the tables in this way. Perhaps it is permissible for *S* (my needy beneficiary) to pursue the goods of a non-altruistically focused life, since she is poor and cannot do much to help others anyway. But it does not seem to follow that it is permissible for *me* to pursue those same goods, given that I am affluent and could do a lot to help others if I were willing to lead an altruistically focused life. Cullity's argument-form goes through only when we can pass seamlessly from claims about the moral status of the potential beneficiaries' interest in something to the corresponding claims about the moral status of *our* interest in that same thing. But it is not at all clear that we can do this.

Second Cullity believes that noting these limits to what beneficence can demand does not refute, but rather supports an alternative interpretation of, the argument he gave in Part I. In particular, it is supposed to support an *aggregative* rather than an *iterative* interpretation of the costs I can be asked to bear to aid others, and thus of principle (AA). As we saw, on the iterative interpretation, when the cost to me of aiding *this one (more) time* would be trivial, I am deemed not to have a sufficient countervailing reason to refuse to aid, given how powerful a reason I have to aid. Cullity says the disparity between what is at stake for me and what is at stake for the potential beneficiary makes it wrong for me to refuse to aid (71). And he specifically denies that "I have already done my fair share" or "I have already aided a lot of other people" count as legitimate countervailing considerations (77, 86): in the face of *this person's* claim on my assistance, they are simply beside the point.

But on the aggregative interpretation which Cullity favours, the *overall* cost I have incurred through aiding *is* supposed to be a sufficient countervailing reason that justifies me in refusing to go any further, provided I am following a reasonable policy of helping others that "allows me to retain a defensible engagement with my own projects, relationships, and other life-enhancing goods, while recognizing the claims that other people's interests make on me" (191). It is not obvious, however, how this fact can be a sufficient defeater of helping in one more case, if the cost to me of doing so would indeed be trivial. This is especially so if we have already rejected as missing the point the other possible defeaters already mentioned, for it is far from clear how this reply is less beside the point than the ones Cullity expressly rejects. So in general, I am worried that Cullity's description of the moral structure of the situation in any single case of potential life-saving makes it near impossible to identify any sufficiently cogent countervailing reason which could release us from saving anywhere short of the Extreme Demand.

Finally, it is not clear that Condition C and Principle P—which are crucial to Cullity's argument—have been adequately motivated, or clearly enough defined. To see this, let us ask whether the notions of permission and requirement which play a key role in those principles are to be understood as *pro tanto* or, alternatively, as all things considered.[14] Does a person's interest in a thing *x* count as satisfying Condition C if that interest generates a *pro tanto* obligation on others? Or does it so qualify only if it generates an all-things-considered obligation under the circumstances, at least when the cost to the benefactor is minimal?

Suppose it is the former. Then the permission concerning *x* which is established by Principle P must also, for parallelism, be *pro tanto* rather than all-things-considered. In that case, though, Principle P cannot be used to deduce the kinds of conclusions for which we hoped. Principle P may yield the conclusion that we are *pro tanto* permitted to pursue non-altruistic sources of personal fulfillment, but the adherent of the Extreme Demand was not disputing *that*. She might say with perfect propriety:

> I am not saying there is anything in itself immoral about pursuing such sources of fulfillment. All else being equal, it is fine to pursue them. All I am saying is that in the present circumstances it is wrong *all things considered* to devote yourself to such pursuits when you could instead be doing so much more to save lives.

On the *pro tanto* interpretation, then, Cullity's principles do not reach to the main point at issue, which is whether our non-altruistic pursuits are *all-things-considered* permissible in the present circumstances.

Suppose, alternatively, that only interests which generate all-things-considered obligations of beneficence (assuming minimal cost to the benefactor) count as satisfying Condition C. In that case Principle P will also yield all-things-considered permissions, and we can be confident not only that we are *pro tanto* permitted to seek or hold on to *x* but that we are so permitted all things considered, a much more robust result. There is, however, a serious drawback to interpreting the argument in this way: it will tend to shrink the range of Condition C, and thus the range of *x*s for which we will be able to draw the conclusion that morality cannot take *x* away from us. Cullity stated confidently that a talented student's interest in receiving a musical education satisfies Condition C. But consider the following exchange between people discussing whether this is so:

[14] Many thanks to Tom Hurka for suggesting that my worry could fruitfully be pressed in this way.

A: “Do I think I ought to help a musically talented person receive a musical education, if I can do so at minimal cost to myself? Sure I do.”

B: “But it’s wartime—this person ought to be serving in the armed forces defending her country, not studying the flute.”

A: “Hmm, you’re right. Under the circumstances, then, I don’t think I should encourage her musical studies at this time.”

A clearly thinks that an interest in receiving a musical education is sufficient to ground a *pro tanto* obligation of beneficence, at least on those who could aid at minimal cost to themselves. But when B brings up further morally relevant elements of the circumstances, A agrees that there is no all-things-considered obligation to provide this student with a musical education at this time, even at minimal cost to oneself. On the present interpretation, though, to say that is to say that this student’s interest in a musical education does *not* after all satisfy Condition C. In other words, now that we are insisting on all-in obligations, the range of Condition C has effectively shrunk.

Suppose we become moved by the thought that considerations of *beneficence*, and not just wartime, can make it impermissible (or at least morally problematic) for a person to pursue interests that, while morally innocent in themselves, should not be prioritized under present circumstances. Then an exchange like the one above seems equally plausible, and we may well conclude that our interests in non-altruistic sources of fulfillment do not after all satisfy Condition C. In that case, though, we cannot use Principle P to defend our pursuit of those interests in the face of putative moral demands, and the refutation of the Extreme Demand again fails.[15]

9.4 THE LIMITS OF “THEY CAN’T TAKE THAT AWAY FROM ME” ARGUMENTS

I have suggested—to my own disappointment, I confess—that none of these “they can’t take that away from me” arguments appears to be fully

[15] One might think it *good* news if the range of interests which give rise to obligations of beneficence shrinks. For wouldn’t that reduce the demands of beneficence? The difficulty is that as long as we grant that an interest in *remaining alive* grounds obligations of beneficence, those of us in a position to save lives by donating to aid agencies will immediately be subject to iterated instances of (AA) which will collectively add up to an extremely strenuous demand. We were counting on a *rich* range of interests satisfying Condition C in order to block that conclusion.

satisfactory. Let me now raise a further depressing point. Even if any of these "they can't take that away from me" arguments worked, it is striking how little they seem to prove. First, there is the plain fact that saying "you can't take that away from me" registers as a downright invitation to take everything *but* that away from me. You can build a fence around the farmhouse, but the cows *will* come right up to the fence; you can take refuge in your impregnable citadel, but this is cold comfort if you remain under siege, eating nothing but turnips and surrounded by your unfinished cheap furniture.[16]

Less rhetorically, we can illustrate the limited scope of this kind of argument by returning to the issue of career and lifestyle choice which we briefly mentioned in connection with Singer. Sarah Buss raises the following arresting challenge. "I might have chosen a career that permitted me to do more to help people in need," she writes.

> I might have chosen to make more money so that I could spend more on helpful projects; I might have chosen to work at a less time-consuming job so that I could devote more energy to such projects, or to local, national, and international policies that would have more beneficial effects; I might have chosen to make a profession out of helping others. As far as I can tell, these choices are still open to me. *I believe I could make any one of them without sacrificing my happiness or settling for a diminished life.* (Buss 2006, 373; emphasis added)

Suppose Buss is right, and it would indeed *not* ruin her life to change careers in one of the ways mooted above. Buss thinks this makes it a very open question whether she is justified in continuing as a philosopher (and as a devoted mother, and so on), when she could instead make one of those career changes. Buss is perfectly willing to grant that (in Barbara Herman's words) "whatever morality requires of us, it should not make our lives unlivable, or too severe" (Herman 2001, 228); the issue she wants to raise arises posterior to accepting that general constraint, and cannot be resolved by citing it. The question she is asking would arise

[16] The example of cheap furniture is deliberately chosen: although, as we have seen, Cullity believes his argument has refuted the Extreme Demand, he still thinks that "buying expensive clothes or furniture, a new car (or, often, any car at all), or books for a private library is usually morally wrong, as the world now stands" (183). After many years in which the furniture in my house was mostly purchased at Erney's Unfinished Furniture, down Rt. 1 from Princeton towards Trenton, I confess I find the idea that it was wrong of me finally to buy a few handcrafted pieces in cherry quite depressing.

even if a successful "they can't that away from me" argument had managed to shield a certain core level of well-being from the voracious demands of beneficence.[17] For if Buss is right about the resilience of her own happiness, the career changes she describes would not threaten that core level of well-being that we had taken such pains to protect. "I cannot convince myself," she says, "that I would cease to be happy if I were to change my career" (Buss 2006, 375). Indeed, she thinks she could still be a flourishing agent even if she had to flee her country and start all over again flipping burgers (Buss 2006, 389).[18]

Consider the possibility that it is wrong—unjustifiable—for you to be, and to keep being, a philosopher, rather than quitting and becoming an investment banker so that you would earn much more money that you could give to charity. I find that genuine consideration of this idea produces the same vertiginous sensation of the moral floor giving way under one as did the arguments we considered at the beginning of this essay. The proposition I just floated seems highly threatening, at least to me, so if Buss is right that this possibility remains on the table even after all the elaborate argumentation we have already reviewed, I find that an unsettling state of affairs indeed. Why did we go to all that trouble, only to be told that we might still have to quit philosophy for investment banking?

Buss herself thinks we can get only very limited mileage, in this context, out of what she calls "deep" facts about reasons, values, or the nature of rational agency. According to Buss, in order to settle the question she is asking we will have to go down into the muck where substantive reasons duke it out. We will have to end by directly considering the relative strength of reasons, unaided by any sweeping "transcendental" argument to the effect that "they can't take that away from me." This way of seeing the dialectic would seem to be supported by something we noted earlier, namely that our view of whether a given moral demand is acceptable varies with the *reason* offered in support of the demand: the reason for which we are being asked to do that thing. Can you be morally required to go into exile and start a new life flipping burgers?—No, if the only reason proffered for why you must do this is that you could do

[17] A referee for Oxford University Press rightly points out that since Williams' argument concerns ground projects rather than well-being, it might evade the above verdict. Unfortunately we could not endorse Williams' argument in its most ambitious form, in which it sought to insulate our ground projects (or at least those that are not morally objectionable in themselves) from the reach of morality's demands.

[18] If this seems unthinkable, consider that many of our forebears did just that.

more good for others that way.—Yes, if you can only evade this fate by betraying your friend, confidante, and collaborator to a brutal regime.

9.5 ANOTHER WAY OUT?

Responses of the "they can't take that away from me" type are in a certain sense concessive: they implicitly grant that the original argument *for* extreme demands was a good one as far as it went. That argument fails to prevail, on these views, not because it contains a fatal internal flaw, but because its scope is properly limited by a different set of considerations which place restrictions on what morality can ask us to give up, even on the basis of a good argument. These "they can't take that away from me" responses in effect let the argument for extreme demands gather momentum and then attempt to place a boulder in its path to stop it before it tumbles over the cliff into the ravine.

Since the results obtained using this technique have been on the whole disappointing so far, I want to explore in this last section a less concessive way of responding to the original case for extreme moral demands. My hunch is that we will have to object to the arguments which appear to generate extreme demands *at an earlier stage* than the one at which we might object that morality can't take that away from me. Without purporting to offer a general recipe for how to do this, I do want to make an observation about Cullity's Part I argument which might offer another way out.

As I noted in passing earlier, Cullity devotes pretty much all his firepower in Part I to defending "the life-saving analogy." (Indeed, he has a tendency to refer to the whole Part I argument as "the life-saving analogy.") By contrast, I find myself worried about his base case: the supposedly uncontentious judgment concerning cases of "direct rescue" with which his argument begins. I got worried about Cullity's base case when I noticed that his argument is subject to a scope ambiguity which I don't think Cullity ever clearly resolves. Recall that we started with (BC), which can be summarized as:

> It is wrong not to save a life directly when you could easily do so.[19]

[19] "Directly" in this summary is intended to point to cases like that of the child in the shallow pond, in which someone's life is threatened right in front of you; "easily" is meant to indicate that you could save without incurring any cost that you would consider at all significant.

In the next phase of the argument we wheeled in "the life-saving analogy" and argued that the issue of giving to aid agencies has the same moral profile as cases of direct rescue. The conclusion of the argument was (AA), which can be summarized as:

> It is wrong not to give to aid agencies when you could easily do so.

Notice, though, that both the premise and conclusion of this argument are subject to a scope ambiguity.[20] To look first at the premise, (BC): statements like "failing to save a life directly is wrong" (32) and "it is wrong not to save a life when you could easily have done so" (11) are ambiguous, because they contain both an existential and a negation. Such sentences are always subject to two possible interpretations, depending on the relative positions of those two elements. So, let us ask: *what* precisely does Cullity's premise say is wrong? Does principle (BC) condemn us only if we *never* save—if we *always* walk away—or does it condemn us if we *ever fail* to save—if we *ever* walk away? On the first interpretation, I have acted wrongly if (with variables ranging over lives)

$$\exists\, x \text{ (I had the opportunity to save } x \text{ easily and directly)} \wedge \neg\exists\, x \text{ (I saved } x\text{)}.$$

On the second interpretation, by contrast, I have acted wrongly if

$$\exists\, x \text{ (I had the opportunity to save } x \text{ easily and directly} \wedge \neg \text{ (I saved } x\text{))}.$$

The difference between *never*—that is, *not ever*—and *ever not* is clearly important, and it behooves one to specify which one means.

Cullity does not comment directly on which he means—he does not seem to have noticed the potential ambiguity. But close study of his language suggests that he intends the stronger interpretation of the base case, the one according to which it is wrong if I *ever* walk away. (Rigorous informal polling suggests that most interlocutors believe that

[20] Jack Woods has suggested to me that the phrase "scope ambiguity" does not fit this case because one of the two possible interpretations is not a natural reading. Even if it would be too much to charge Cullity's English sentences with ambiguity, the fact remains that there are two different logical structures that could be expressed using those words.

(BC) is true on this stronger interpretation.) On this interpretation, any time you have the opportunity to save someone's life easily and directly, you must do so.[21] So assuming that I have had *n* opportunities to save someone's life easily and directly, I have acted wrongly if the number of lives I have saved is less than *n*, since it is wrong to pass up *any* opportunity for direct, easy rescue.

Let us now move on to consider the conclusion of Cullity's Part I argument, (AA), which is of course subject to the same scope ambiguity as the premise, in virtue of containing an existential and a negation. Does (AA) allege that it is wrong if I *never* donate to aid agencies? Or, rather, that it is wrong if I pass up even a single opportunity to do so? Assuming I have had *n* opportunities to donate to aid agencies at minimal cost to myself, does (AA) say that I have acted wrongly if the number of times I have donated is less than *n*, or only that I have acted wrongly if the number of times I have donated is zero? When it comes to the conclusion of his Part I argument, Cullity's language usually suggests the weaker interpretation, according to which it is wrong for an affluent person not to give at all. As Cullity puts it, Part I of his book "argues that affluent individuals are acting morally wrongly if they do not . . . contribute . . . to . . . aid agencies" (8); or, more briefly, "not giving to aid agencies is wrong" (11). Statements like these are most naturally read as expressing the weaker construction.

Since this announced conclusion sounds unthreatening and unobjectionable—if you put $1 a year in Santa's hat, it poses no threat to you—I think we are not made as suspicious of Cullity's Part I argument as we should be. For if his argument is any good, it actually supports the stronger conclusion that I act wrongly if I *ever* pass up an opportunity to donate. That claim is hard to believe; if we knew that his argument led to that conclusion, I think we would scrutinize the argument more skeptically. In particular, once our attention has been drawn to the scope ambiguity, I think it will be worth our while to reconsider the base case, and to question whether we really do endorse (BC) on the strong interpretation.

Let me emphasize the difference between the two interpretations. On one way of resolving the scope ambiguity—with the existential inside the scope of the negation—Cullity's principles will condemn us only

[21] As a referee for Oxford University Press pointed out, this is effectively to posit a *perfect* duty to save when you can do so easily and directly.

when the ratio of our rescues (or donations) to our opportunities to rescue (or donate) is zero. On the other way of resolving the ambiguity—with the negation inside the scope of the existential—those principles will condemn us whenever that ratio does not equal one: whenever the numerator of that fraction, *m*, is less than *n*, the number of opportunities we have had. These two interpretations are, obviously, very different. But *they will differ little when n is small.* They agree completely when $n = 0$ or $n = 1$; they only start disagreeing at all when $n = 2$; but by the time $n = 1000$ the two will give different verdicts about almost every case.

I mentioned above that most interlocutors accept (BC) on the strong interpretation, according to which it is wrong *ever* to let someone die right in front of you whom you could easily save. Should they, though? I want to suggest that this statement *sounds* to us unobjectionable because we are influenced by our knowledge that in real life, *n* for this kind of case will be extremely low. (So far, it is zero for me.) It therefore *seems* to us more or less costless to assent to the stronger version of the claim: we are sure there is no way this will end up swallowing up our whole life. But when we see where that innocent-sounding admission leads, if Cullity's argument is any good, perhaps we should be more careful.

So far, I have just put up some caution signs about Cullity's Part I argument—I have not actually given you any reason to doubt his premise. But I think we do have grounds to question the strong version of (BC). Suppose yours is the only house bordering a public park which features a shallow pond. Suppose, moreover, that this particular pond appears to exert a magnetic attraction on stumbling strangers, who fall into it with astonishing regularity. It seems that every time you leave your house you are called upon to rescue one of these unfortunates, and sometimes a whole stream of them—no sooner have you pulled one out than another falls in, and another. It is to the point that you can't ever get your kids to school on time, you are perennially late for important meetings, and the constant interruptions for rescue duty have effectively sabotaged your conduct of your own life. Is it as clear as it seemed for $n = 0$ or 1 that you *must* rescue every single one of these strangers—that you must allow your life to be taken over in this way by the duty to rescue that unfortunately falls to you? It is *not*, I think, so clear, and indeed I am inclined to deny outright that you are so obligated.[22]

To take a less fanciful example, consider those who spent day after anguished day searching for possible survivors who could have been trapped under the rubble in the aftermath of the earthquake in Haiti a few years ago. Given the magnitude of the disaster, however many people these brave rescuers managed to find and save from death, there must have been still others who remained trapped beneath the debris and could have been saved had the rescuers only continued sifting through the wreckage a little longer. Are we prepared to say that those rescuers acted wrongly when they eventually put down their tools? This seems highly implausible, suggesting that we do not after all expect the ratio of people's rescues to their opportunities to be one when the denominator is large. (I repeat that for most of us the denominator is extremely small, which tends to occlude the difference between the strong and weak interpretations.)

If I am right that we have reason to question even Cullity's base case, this would offer an alternative route to resisting the argument for extreme demands which we found the most dangerous. On the view I am suggesting, we would not attempt to contain the kudzu by a liberal but *ex post* application of the principle that there are certain things which morality cannot take away from us. We would instead rely on *ex ante* vigilance and firmness to prevent it from taking root in our garden without our considered consent.

REFERENCES

Buss, Sarah (2006). "Needs (Someone Else's), Projects (My Own), and Reasons." *Journal of Philosophy* 103: 373–402.

Cullity, Garrett (2004). *The Moral Demands of Affluence*. Oxford: Clarendon Press.

Funder, Anna (2004). *Stasiland: Stories from Behind the Berlin Wall.* London: Granta Books.

[22] The response "you ought to move," which I have heard, seems to concede that you cannot be obligated to rescue all of those strangers. After all, if you move away there may be no one to rescue them and they will all die, whereas if you simply eliminate rescues between midnight and 9 am, say, many of them will survive (thanks to your rescue activities during the other hours of the day). It seems to me that someone who thinks it would be all right to move away, thus consigning all the strangers to a watery grave, cannot genuinely think it would be wrong to refuse to rescue between midnight and 9 am.

Herman, Barbara (2001). "The Scope of Moral Requirement." *Philosophy & Public Affairs* 30: 227–56.

Murphy, Liam B. (2000). *Moral Demands in Nonideal Theory*. New York: Oxford University Press.

Scheffler, Samuel (1982). *The Rejection of Consequentialism*. New York: Oxford University Press.

Singer, Peter (1972). "Famine, Affluence, and Morality." *Philosophy & Public Affairs* 1: 229–43.

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

Williams, Bernard (1976). "Persons, Character and Morality." In Bernard Williams, *Moral Luck* (Cambridge: Cambridge University Press, 1981).

10

What We Know and What We Owe

VANESSA CARBONELL

What to study, where to live, what career to pursue—for those of us with the luxury of making them, these decisions both give our lives meaning and determine what sorts of things we *know*. Adding that minor in Spanish will render you passably fluent, allowing you to communicate with people you otherwise could not have. Choosing to live in an economically diverse neighborhood might give you knowledge of the cares and concerns of the working poor. Pursuing a career in medicine will probably make you competent to perform certain basic procedures, even if you decide to specialize in psychiatry. But how does what we *know* bear on what we *owe*? In one sense, the answer is obvious: knowledge is necessary for certain moral obligations. In learning something new, we sometimes acquire the last necessary element in a set of elements that are jointly sufficient for making us morally required to act. Knowledge thus triggers new obligations, so the choices that shape what we know can also shape what is morally demanded of us. But, as I will show, the existence of these knowledge-based obligations is in substantial tension with the thought that we are free not only to choose the course of our own lives, but also to change our minds and quit, no matter how much knowledge we have already acquired. To resolve this tension, I argue that we need to adopt a relatively demanding understanding of the conditions under which it is permissible to swap one career or life project for another. The result is a compromise that reconciles the moral force of knowledge-based obligations with a basic freedom to choose less-than-morally-optimal life paths.

The essay is divided into four parts. In Section 10.1 I sketch a puzzle: reasonable principles about freedom are seemingly incompatible with the claim that knowledge can trigger new moral obligations. In Section 10.2 I argue that knowledge-based obligations are uncontroversial and robust, which suggests that we cannot solve the puzzle by abandoning them. Section 10.3 introduces a solution to the puzzle: we must narrow the scope of our freedom to "quit at any time." Finally, in Section 10.4 I examine the *timing* of knowledge-based obligations, arguing that they do not entail an earlier obligation to embark on educational endeavors, nor do they take full effect until certain formal or informal thresholds have been passed.

10.1 THE MOTIVATING PUZZLE

Moral philosophers have already made great progress in understanding the relationship between knowledge and obligation. There is much excellent work, for example, on cases where people *do not* know something, and whether this gives them an *excuse.*[1] I shall focus on the related question of whether what people *do* know generates *additional* obligations. And while the existing literature tends to focus on *particular* beliefs and *particular* actions,[2] I shall focus instead on the relationship between large stocks of knowledge or know-how (and large-scale educational endeavors) and patterns of action or behavior. My project is thus concerned not only with the narrow moral question of *what to do in a particular circumstance*, but the broader question of *how to live*.

We can see how this broader question arises by examining an apparent tension between three principles that are, I believe, fairly widely (though not universally) accepted by philosophers and non-philosophers alike. The first is a principle about how we choose our careers or life paths (to the extent that we are able to do so). Call it *Life Path Freedom* (LPF):

(LPF) We are not morally required to pursue the morally best career or life path available to us.

[1] To give just two examples: In "Culpable Ignorance," Holly Smith considers the case of a doctor who did not know, but should have known, that high concentrations of oxygen can cause blindness in premature babies (1983, p. 543). In "Justified Wrongdoing," Sarah Buss considers the case of an abused teenager who does not realize, and perhaps cannot be expected to realize, that not all adults are "potentially dangerous enemies" (1997, p. 354).

[2] Accordingly, Smith (1983) considers the moral status of the doctor's particular action—administering excess oxygen—in light of the doctor's particular belief about proper treatments for respiratory distress; Buss (1997) considers the moral status of the teenager's particular action—attacking an adult perceived to be a threat—in light of his particular belief about the dangerousness of adults.

The LPF principle is both weak and broad, by design. It is weak insofar as it only says we are not *required* to pursue the morally best path; it leaves open the possibility that we have very good reason to do so, and that our chosen path can be subject to negative moral evaluation. Further, it says only that we are not required to pursue the *best* path; it leaves open the possibility that we are required to choose some minimally decent path. It is broad insofar as it concerns both careers and "life paths," where the latter would include many long-term, engrossing, non-remunerative projects, such as having children, pursing a PhD, emigrating from one's country of birth, and pursuing hobbies or volunteer work. LPF is also neutral about what paths count as "morally best." Different moral theories will give different answers to this question. I use "morally best" rather than, say, "most helpful," in order not to give the impression that the only moral consideration that bears on what we ought to do with our lives is helpfulness (and the absence of harmfulness).[3]

One final note about LPF: it is only a principle about what we are (or are not) *morally* required to do. It says nothing about what it would be good to do according to some other normative standard, such as prudence, well-roundedness, or aesthetic value. Nor does it make any claim about how we ought to adjudicate between the requirements of morality and the requirements or recommendations of these other normative standards. It says nothing about whether the moral standard is the ultimate, trumping standard; it makes no claims about whether there is some all-things-considered perspective outside of these standards. In light of all this, we should be modest about how much insight the LPF principle brings to the question of how we ought to live. There is more to life than the narrow dictates of morality. But we have to start somewhere, so in this essay I focus on the question of what *morality* requires of us.[4]

[3] In this respect I am drawing the question more broadly than Sarah Buss does in "Needs (My Own), Projects (Someone Else's), and Reasons" (2007). Buss asks whether she can justify pursuing a minimally helpful career instead of a maximally helpful career. But I am sure Buss would agree that there are other moral considerations that bear on our choice of career or life path besides helpfulness. That said, the way I am framing the question might complicate her argument, since it leaves open the conceptual possibility that an unhelpful career or life path can be the morally best one. In any case, helpfulness is clearly the most uncontroversial way that an action or life path can be morally good, and so most of my examples involve *helping* in one way or another.

[4] In this respect I am focusing on a much narrower question than Susan Wolf in "Moral Saints" (1982). Wolf, too, is interested in how we ought to live, and in what a morally excellent life would look like. However, I take it that Wolf's primary concern is to show that the morally best life would be a life that we find unattractive from an independent, non-moral

I will not offer a formal argument for LPF. I have drawn it so weakly that I doubt it is especially controversial.[5] It ought to be accepted by anyone who rejects a kind of "maximizing rationality" in ethics—that is, anyone who rejects the idea that, with respect to a given choice, the morally *best* option is morally required. It ought to be accepted by anyone who thinks that the demandingness of a moral option can render it non-obligatory.[6] It ought to be accepted by anyone who thinks that one cannot be morally required to do something that interferes with the kind of non-moral personal projects that are sometimes integral to one's very identity.[7] And it ought to be accepted by anyone who thinks that, even if there is a highly demanding moral theory in the background, our *careers* and *life paths* in particular get a kind of special dispensation—a "pass" from the relentless demand to do the morally best action (if not a pass from the demand to avoid morally bad actions).[8]

In short, I suspect that LPF would be controversial only if we accepted something like a strict maximizing consequentialism, with no special exceptions for "personal projects," no "agent-centered prerogatives," and no distinction between obligation and supererogation.[9] There is not room here to assess whether such a form of consequentialism is

perspective, which she calls the "point of view of individual perfection" (p. 437). Her view is often thought of us as therefore critical of moral theory. Nevertheless, she would not necessarily have to endorse LPF, which says that *morality* does not always require us to do the morally best thing. I read Wolf as arguing that, *if* we lived the morally best life, it would be a life that looks awfully different from what we think the best life, period, should look like. This view has even more force if we *are* morally required to live the morally best life—that is, if LPF is false.

[5] Buss (2007) seems to think that a stronger principle—that we are not even morally required to move from a minimally helpful career to a *more* helpful career—is widely accepted. If she is correct, then I would imagine that the LPF principle—given that it is significantly weaker—would also enjoy wide support. That said, Buss makes a convincing case that her principle has not been adequately defended by philosophers, and the purpose of her paper is to raise doubt about its truth, or at least its having a "deep justification" (p. 398). Her arguments cannot be ignored, but they presumably apply only in weaker form to this weaker principle.

[6] The idea that a particular moral option is not obligatory because it demands too much is different, of course, from the view that an entire moral theory is implausible because it is too demanding. I think the former alone would be sufficient to entail LPF: if some career or life path is morally best, but not required because it demands too much, then LPF is true.

[7] This would be the kind of "integrity" objection voiced famously by Bernard Williams in "A Critique of Consequentialism" (1973).

[8] The idea would be something like Scheffler's (1982) "agent-centered prerogative," applied especially to careers or life-paths.

[9] A view like this is defended by Shelly Kagan in *The Limits of Morality* (1989).

plausible. So I shall remain agnostic on this question, while admitting that I think the LPF principle is probably true.[10]

LPF captures an extremely weak version of the thought that we are free to shape the course of our own lives. But does this principle, weak as it is, extend to decisions about whether to *discontinue* life paths? Allegiance to LPF might lead us to accept a somewhat more controversial principle, which we can call *Quit at Any Time* (QAT).

(QAT) Other things being equal, when pursuing a career or life path, one is morally permitted to quit and change course at any time.

This principle is not strictly entailed by LPF. At first it might seem to be: LPF says we do not have to pursue the morally best career, so if continuing in the morally best career is not required, it follows that

[10] LPF might also be controversial to those who think that, at least in some cases, people with rare or exceptional talents (such as Michelangelo) are morally required to choose careers or life paths that use their talents. (Thanks to Andrew Schroeder for suggesting this point.) If this is correct, and if those paths or careers are morally best, then LPF is false as stated. Would this be a way to reject LPF without endorsing strict consequentialism? I suspect that consequentialist reasoning underlies many of these judgments. For instance, we might think the reason Michelangelo had an obligation to produce art is that, in doing so, he brought more value to the world than he could have by doing anything else. Nevertheless, there may be non-consequentialist ways to view this case. Some might think that Michelangelo's obligation, if in fact he has one, is grounded not in the value he produces, but in the idea that the perfection of one's talents is the mark of a virtuous person, and the morally best life-path is the one that best exemplifies virtue. Alternatively, one might see his obligation as grounded in an interesting sort of interpersonal relation: talents are unequally and unfairly distributed, and those who are lucky enough to have them owe it to the rest of us not to let them go to waste. We see this line of thinking in the film *Good Will Hunting* (1997). Ben Affleck's character Chuckie pleads with his best friend, Matt Damon's Will, a self-taught mathematics genius, to use his talents to take a fancy job and move away from the old neighborhood. Chuckie says, "Listen, you got somethin' that none of us have." Irritated, Will says, "Why is it always this? I owe it to myself? What if I don't want to?" Chuckie replies, "Fuck you. You owe it to *me*. Tomorrow I'm gonna wake up and I'll be fifty and I'll still be [working construction]. And that's all right 'cause I'm gonna make a run at it. But you, you're sittin' on a winning lottery ticket and you're too much of a pussy to cash it in. And that's bullshit 'cause I'd do anything to have what you got! And so would any of these guys. It'd be a fuckin' insult to us if you're still here in twenty years." Chuckie's appeal to the idea of an *insult* suggests that he is concerned not simply with consequences but with a kind of interpersonal *respect*. That said, it is clear in the film that this job is not the *morally best life path* for Will, so Chuckie's stance is compatible with LPF. Moreover, the respect in question seems to trade just as much on Will's special relationship with Chuckie as it does on Will's talent. To make a general non-consequentialist case for the principle that those with exceptional talents have obligations to use them would require a way of understanding the "insult" or disrespect without appealing to a particular relationship. (Thanks to Ethan Katz for pointing me to this scene.)

discontinuing the career is permitted. But this would hold only for the few people already pursuing the morally best career, and even so it is not obvious that a permission to discontinue is a permission to quit at *any* time. So QAT is not a genuine corollary of LPF. Nevertheless, QAT is appealing for the same reason that LPF is appealing: it carves out a sphere of life choices and protects them from at least the most stringent type of moral scrutiny. After all, LPF would be a rather pathetic sort of freedom if, once you were established in a given life path, you were *stuck* there, on pain of violating a moral requirement.

To see why QAT is plausible, we need to appreciate how weak it is, which requires specifying what is covered by the "other things being equal" qualification. The main purpose of this qualification is to set aside certain special or circumstantial obligations that arise in the course of a career or life path, some of which may be common to all careers or life paths, so as to focus on whether our freedom to choose the course of our own lives includes a freedom to quit and change course *irrespective of the comparative moral status of the old and new paths*. This is best illustrated by way of an example. Suppose that Jimmy is a certified nurse's aid taking care of people in a nursing home. One day he decides to quit and make a living playing online poker from his basement. Is this change of course morally permissible? Several considerations immediately come to mind. Consider these five:

(1) Jimmy has dependent children. Will poker provide a consistent income to put food on the table?
(2) Unpredictable financial windfalls could tempt Jimmy to make shortsighted decisions, such as purchasing a speedboat instead of flood insurance. Will switching careers leave the family worse off overall?
(3) The legal status of profiting from online poker is ambiguous. Is this new career morally forbidden because it might be illegal?
(4) Jimmy makes his decision in the middle of a shift, or even the middle of a procedure. Can he quit if it means the patient in front of him will die?
(5) Even if poker is not *wrong*, nursing is just a *morally better* career, for a variety of reasons that vary with your moral theory. Is Jimmy permitted to leave the morally better career for the morally worse one?

The purpose of the "other things being equal" clause is to set aside considerations of type (1)–(4), among many others. Why should we set these aside? Regarding (1), we probably all agree that, if Jimmy has an existing special obligation to put food on the table for his children, and if online poker will not allow him to do so, then he is morally forbidden from quitting nursing and switching to poker. But the fact that Jimmy lacks permission to quit in such a case has nothing to do with the relative moral status of nursing-home work compared to online poker; rather, it is grounded in his special obligation to his children. Similarly, regarding (2), if Jimmy knows that poker will cause him to mismanage his finances in a way that ultimately harms the family, then he is probably morally forbidden from switching to poker. But if so, it is because he is forbidden from making choices that he knows will harm his family. This is a general moral restriction that governs all of Jimmy's choices, and is external to his career or life path. The same goes for considerations of type (3): if we are morally forbidden from doing illegal things, and if poker is illegal, then Jimmy is not morally permitted to switch from nursing to poker. But it is not due to a moral prohibition on quitting; it is due to a moral prohibition on illegal careers. We want the QAT principle to allow us to hold equal things such as the relative legality of the old and new careers, so as to avoid a proliferation of trivial counter-examples, such as the fact that you are not morally permitted to quit your nursing job at any time *in order to become a gangster*. The same goes for (4): without an adequate fleshing out of the "other things being equal" clause, the permission to quit at *any* time would allow Jimmy to quit, say, after he has disconnected a patient's ventilator tube to quickly clean it, but before he has put it back. It is true that Jimmy cannot quit in the middle of the procedure, but it is because other things are not equal: he has a pre-existing obligation to finish the procedure or find someone who can, independent of any obligation not to quit his job. Perhaps it is a special obligation to this patient, or a contractual obligation to his employer, or just his general obligation as a human being not to harm another human being. Whichever one it is, it is already covered, so we do not need to deny the freedom to quit at any time in order to cover it.

What I am interested in capturing with the QAT principle is the freedom to change one's mind about what to do with one's life without having the decision guided exclusively by moral reasons. QAT says that, assuming other things like legality and special obligations are equal, Jimmy is permitted to switch from nursing to poker, *even if poker is morally worse*. In other words, QAT directs us to ignore considerations of type (1), (2), (3), and (4), in order to focus on (5). But then it says that considerations of type (5) are not decisive. There is a general moral permission to quit doing what you are doing, and to change to something else, even if doing so is a moral downgrade, so to speak.

Of course, QAT does *not* say that Jimmy's career change is immune from moral criticism. It does *not* say that he should *ignore* moral considerations in making his decision, or that the relative moral status of the careers provides him with *no reasons whatsoever*. It simply says that, in quitting, he will not have violated a moral requirement—he has a general permission to quit. This permission is presumably grounded in the same basic freedom to shape the course of our own lives that underlies LPF.[11]

To recap, I have suggested that the following two related principles are widely accepted:

Life Path Freedom (LPF). We are not morally required to pursue the morally best career or life path available to us.

Quit at Any Time (QAT). Other things being equal, when pursuing a career or life path, one is morally permitted to quit and change course at any time.

I have also claimed that while QAT is not logically entailed by LPF, the two principles are based on the same underlying ideas about freedom, and those who believe QAT likely do so for the same reasons that they believe LPF. Now, while one may certainly *disagree* with either of these principles, there does not seem to be anything especially *puzzling*

[11] I think we not only accept the permission to quit at any time, but in fact *glorify* it, both in real life and in fiction. For instance, when a JetBlue flight attendant had a particularly stressful flight, he quit his job by grabbing a beer from the beverage cart and sliding down the emergency slide onto the tarmac. People treated him like a hero. (Lauren Frayer, "JetBlue Flight Attendant Who Flipped is a Folk Hero," AOL News, 10 August 2010. <http://www.aolnews.com/2010/08/10/flight-attendant-who-flipped-becomes-folk-hero/>) The movie *Office Space* (1999) also glorifies quitting one's job, even though the protagonist's new stress-free lifestyle is made possible only by an illegal scheme. Of course, it makes sense to cheer when people quit jobs or projects because they are overworked, oppressed, or simply want to live a less stressful life. But this "Good for you!" attitude sometimes carries over even to cases where people quit non-oppressive and indeed morally admirable projects in order to do nothing, or do something significantly less admirable.

about them, alone or taken together. The puzzle comes when we combine QAT with a view about what happens to us, morally speaking, when we acquire certain kinds of knowledge (or certain kinds of skills or experience that are constitutive of knowledge). This view is captured by a principle that I will call the *Burden of Expertise*.

Burden of Expertise (BEX). One has certain moral obligations in virtue of possessing certain knowledge, skills, or experience.

I will postpone for the moment discussing what exactly is meant by "in virtue of" in this principle. For now we can understand it as expressing, at a minimum, that if one did not have the knowledge, skills, or experience, one would not have the moral obligations.

The best way to illustrate the BEX principle is by looking at some examples. First consider Peggy.

Peggy. Peggy wrote her doctoral dissertation in anthropology on the power relations between rival tribes in Waziristan, the volatile region between Pakistan and Afghanistan. Now happily employed as a professor at a small liberal arts college and writing about other things, Peggy gets approached by officials at the State Department. They ask her to take a year-long sabbatical and come to advise them on peace-making and terrorism-prevention strategies in the area. No one in the world has exactly the same expertise Peggy has, though there are others with other kinds of relevant knowledge. Though Peggy would prefer teaching, she feels obligated to accept the offer and agrees to the one-year post.

It seems that Peggy's obligation to take the post—if she is indeed so obligated—is conditioned on her expertise and the fact very few people have comparable expertise. If she did not have this particular expertise, or if there were a long line of comparable experts ready and willing to take the post, she would not be so obligated. Surely other moral factors—such as beneficence and civic duty—also contribute to the obligation, but a good explanation of why she ought to accept it would refer to her technical knowledge.

Now consider the case of Roz.

Roz. Roz is a pediatric cardiac surgeon who specializes in fixing a rare congenital heart defect in young children. She has trained for years to do this work: college, medical school, residency, and fellowship all prepared her to be the surgeon

> she is today. At age 39 she is the most talented surgeon in her geographical area, which means that she gets sent the most critical cases. It is fair to say that without her the care of these children would be compromised, and some might die while being transported to distant hospitals to be treated by the few other experts. Roz likes her work and is not burnt out, but she also has another passion: French literature. She is considering leaving medicine, at least for several years, to pursue a PhD in French literature and write a book.

Is Roz morally required to stay at her job rather than pursue her passion for French literature? Suppose she cannot do both, and suppose that, since she is not burnt out or depressed, abandoning her plan to study French literature would not render her so disappointed as to compromise her abilities as a surgeon. I propose that Roz *is* morally required to stay in medicine, in virtue of her expertise. If she left her job, it would be fitting for those in her community not only to lament the state of affairs—to say "what a shame" or "think of those poor sick kids"—but also to *blame* her, to be *angry* with her, to feel *wronged*. These are reactive attitudes characteristically associated with the failure to meet a moral obligation. We can imagine her coworkers shaking their heads to each other, saying that her decision was "selfish" or "frivolous," even when they would not normally accuse a scholar of French literature—one who had pursued that path since college—of being selfish or frivolous. And if Roz follows through with her move we can imagine her feeling guilty while sitting in French class, thinking of the children she could have saved. This would not be the same guilt that *anyone* might feel at not having donated more money to UNICEF to prevent childhood diseases in the developing world. Nor would it be a guilt associated with *breaking promises* to particular children already under her care, since she is thinking of hypothetical, future patients. Rather, it would be guilt about a more personal misgiving: guilt about not using the technical knowledge and experience that she possesses. It would be the guilt of not carrying the burden of her expertise.

Judgments about these cases may differ, and I am not offering the cases as an *argument* for the BEX principle. Nevertheless, I think the cases can make vivid how the principle captures a strain in our commonsense thinking. Granted, we may have mixed feelings. Perhaps we go back and forth between thinking that Roz would be blameworthy for making the switch and thinking, as Oprah might put it, that Roz is free to live her *best life*, and that she does not have to carry all those sick children on her shoulders. But even ambivalence on this matter is enough to raise interesting questions. To the extent that we are even *tempted* to believe both that Roz has a moral burden associated with her technical knowledge and that we are generally free to quit and change course at any time, we are faced with a puzzle.

The puzzle is this: if it is true that (other things being equal) we can quit at any time (QAT), then how can it be the case that, once far enough along in a career or life path to have acquired some knowledge (or skills or experience), we can be *morally obligated* to take certain actions *in virtue of* that knowledge (BEX)? After all, QAT says we can quit at any time—presumably including now! Instead of fulfilling this new moral obligation, this "burden" of our expertise, *why not just quit*?

While the puzzle can be generated by QAT and BEX alone, LPF is lurking in the background. I have argued that a commitment to QAT is grounded in the same views about freedom that motivate LPF. And without QAT, LPF loses some of its force. After all, without the freedom to quit at any time, the permission to choose a less-than-morally-best path only goes so far: if you have already chosen the morally best career, you are stuck! So in order to resolve the puzzle, we need to either weaken our commitment to knowledge-based obligations, or weaken our commitment to quitting at any time, which would mean scaling back our freedom to shape the course of our lives.

Of course, the Peggy and Roz examples might show that BEX is present in our commonsense moral thinking, but they do not show that BEX is true. So in the next section I try to spell out exactly why BEX is true. Then, in Section 10.3 I argue for a compromise that preserves BEX at the expense of weakening QAT. We may need to admit that you cannot quit at just *any* time.

10.2 KNOWLEDGE-BASED OBLIGATIONS

How can the mere possession or acquisition of knowledge create moral obligations? After all, knowledge may generally be necessary for moral obligation, but it is never by itself sufficient. We may agree that, in most or all cases, if you do *not* know how to φ, you are *not* obligated to φ. To think this is just to be committed to *obligated implies knows how*, a species of the *obligated implies can* principle,[12] itself a weaker and even less controversial cousin of *ought implies can*. And we *should* be committed to *obligated implies knows how*, at least for most cases, because to say that someone is *morally obligated* to do something is to say quite a lot.[13] It is to say that she has decisive reason to do it, that we can legitimately demand it of her, that we can hold her accountable if she fails to do it, that she could demand the same of us if we were similarly situated, and that it would be fitting for her—and for us—to feel and express morally-laden reactive attitudes if she failed to do it. So it would seem just downright unfair, not to mention pointless, for someone to be morally obligated to do something she simply *cannot do* because she does not know how.

But it is obviously not the case that, in general, if you *do* know how to φ, then you *are* obligated to φ. (Know how to steal a car? It does not follow that you must. *Knows how implies obligated* is a non-starter.) Instead, new knowledge creates new obligations *only* in circumstances where it adds to existing obligation-grounding reasons. So the principle we are seeking is something like this: *knows how, and has otherwise nearly decisive moral reason to, implies obligated*. Another way to put it is this: knowledge "triggers" obligation when it provides one of the necessary and jointly sufficient conditions for obligation—specifically, when it is either the last (temporally) of the conditions to obtain, or when it is the condition that sets a given agent apart from some comparison class. I will hereafter use "create" and "trigger" interchangeably; both are really just metaphors for a set of facts or reasons coming to obtain.

[12] Not knowing how is not *the same thing* as not being able; it is one *species* of not being able. We are interested in cases where it is precisely the lack of *knowledge* that entails the lack of ability. I cannot lift an elephant because my muscles are not up to the task. I cannot make a round square because it is metaphysically impossible. But I cannot speak Spanish because *I do not know how*. This distinction is important if knowledge plays a distinctive role in our lives or identities.

[13] *Obligated implies knows how* may not always be true. Perhaps it is possible to be able to do something despite not knowing how. I think this question turns on tricky conceptual issues about knowledge and ability that I cannot explore here. An additional question is whether the relevant factor is what a person knows or rather what it would be reasonable for her to believe in light of her evidence. In "The Moral Clout of Reasonable Belief" (2011), Holly Smith makes a case for ignorance being an excuse even when the ignorance itself was inexcusable. In other words, she argues that it is what we *actually* believe, not what it would be reasonable for us to believe, that is relevant to how we should be morally assessed for whatever actions we take in light of our beliefs.

Recall the BEX principle. One has certain moral obligations in virtue of possessing certain knowledge, skills, or experience. For BEX to be true, there simply need to be cases where knowledge is added to existing moral reasons that are otherwise nearly decisive, or where knowledge sets one agent apart from others. Surely such cases abound. For example, suppose a man collapses on the railway platform and is dying while waiting for the paramedics. As the sole bystander I would be obligated to save his life, but I do not know how. (Fortunately, the paramedics arrive just in time.) Coincidentally, a CPR course is offered at my workplace that day, and I take it. On my return commute, shockingly, another man collapses on the railway platform. No one else on the platform has the relevant knowledge, but now I do. A knowledge-based obligation has been triggered. My training triggers the obligation both by being the last necessary condition to obtain and by being the necessary condition that distinguishes me from the other bystanders.

Now, at this point the existence of knowledge-based obligations may seem so uncontroversial as to be not worth discussing: *of course* the person who knows CPR now bears a new moral burden that other bystanders do not. And yet these knowledge-based obligations are noteworthy for several reasons. First, they illustrate that what is morally demanded of a given agent is remarkably plastic; the set of moral obligations she faces can change rapidly in the face of contingent, non-moral considerations such as whether she happens to have just seen a documentary on global warming or dog-fighting.[14] Also, what is morally obligatory can differ dramatically from one agent to the next. While this may seem obvious in light of the wide diversity of circumstances in which people find themselves, it takes on a provocative tone when we focus specifically on differences in knowledge: we do not always have control over how much or how little we know, and yet those who know more, owe more.

Another noteworthy feature of knowledge-based obligations is their relation to an agent's identity. Of course, knowledge is not the

[14] Perhaps this should come as no surprise. After all, moral obligation need not be anything magical, eternal, or unchanging. Rather, it is arguably a contingent and contextual social phenomenon—a system of demands we are justified in making upon one another in light of an underlying authority we have as mutually responsible agents in a community. This authority might be "second-personal" in nature, as Stephen Darwall argues in *The Second-Person Standpoint* (2006). Whatever the nature of the authority and the corresponding obligation at the conceptual level, I take it that the *content* of the obligations could take a variety of forms, including consequentialist.

only thing that can trigger obligations. But we are often *responsible* for our knowledge and skills—responsible for choosing to pursue them, for succeeding in acquiring them, and for tirelessly working to perfect them. In this way they are part of us, and sometimes reflective of our character, in a way that other obligation-triggering abilities may not be. For example, if someone has fallen into a hole and needs help, a bystander may be obligated to help in virtue of being the only bystander with long enough (but not exceptionally long) arms. This is a case where the fact that the bystander *can* help in this particular way triggers an obligation that is defeated for other bystanders with shorter arms. But being able to help make peace in Waziristan because of your scholarly expertise and being able to lift someone out of a hole because of the length of your arms are two very different things. Knowledge-based obligations can be tied to our identities, and triggered by features about ourselves for which we can take credit, in ways that set them apart.[15]

Now, the CPR example was simple and narrowly circumscribed. What are even more interesting are large-scale bodies of knowledge, long-term educational endeavors, and major decisions about careers and life paths—decisions like those faced by Peggy and Roz. What makes these cases more interesting is that they involve a kind of knowledge or expertise that takes years to acquire. Generally speaking, the more time and effort you devote to a knowledge-producing project (whether it be formal schooling or getting to know the needs of a community), the more likely it is that your knowledge is substantial, highly specialized (that is, narrow), and rare. The more substantial, specialized, and rare your knowledge, the greater the number and urgency of the demands it may trigger.[16] Of course, the longer it takes you to acquire the knowledge or skill, and the more substantial, specialized, or rare it is, the more likely it

[15] Of course, knowledge may not be the *only* ability that is special in this way. If saving the person who fell in the hole requires not *long* arms but *strong* arms, and I have strong arms because I am an Olympic athlete, then perhaps my strength is tied to my identity in interesting ways despite not being a kind of *knowledge*. Even this case, though, would count as knowledge if brute strength is considered a skill and skill is considered a kind of knowledge.

[16] There are exceptions, of course. Insubstantial, unspecialized, or relatively common knowledge can trigger obligations in moments of extreme demand, like natural disasters. And rarity is always relative to context; language skills, for example, can be highly valuable in one community and relatively useless in another. It is also worth emphasizing that many careers or life paths involve a type of knowledge that is substantial, specialized, and rare (relative to context), without requiring advanced degrees or being traditionally accorded a high status

is to be tied up with your identity and your values. So you are not simply faced with a bunch of bland, bureaucratic prescriptions: because I know x, I must do y. Rather, you are faced with the broad moral implications of the *kind of agent* you have chosen to mould yourself into.[17] This brings us back to the puzzle of the previous section. Does this new moral landscape, filled with knowledge-based moral obligations, conflict with the freedom to QAT?

Quit at Any Time (QAT) Other things being equal, when pursuing a career or life path, one is morally permitted to quit and change course at any time.

Burden of Expertise (BEX) One has certain moral obligations in virtue of possessing certain knowledge, skills, or experience.

If my argument has succeeded in demonstrating the existence of knowledge-based moral obligations, then the BEX principle is true. If BEX is true, then it is hard to see how QAT could also be true. Let us assume that Peggy's knowledge of Waziristan triggers an obligation to work for the State Department, and that Roz's skill in performing surgery on children's hearts triggers an obligation to stay at her job.[18] If Peggy and Roz are *morally obligated* to do these things, how could it be true that they are *morally permitted* to quit or change course and pursue some other life

in society. After Hurricane Sandy hit New York there was a shortage of qualified licensed electricians who were needed to inspect and repair the wiring in homes whose basements had flooded. One electrician interviewed by National Public Radio noted his sixteen-hour workdays and claimed that, though he misses his wife, turning down requests would mean leaving people without heat at the beginning of winter. It seems as if he felt bound by a knowledge-based obligation. (Joel Rose, "N.Y. Electrician Shortage Hampers Sandy Recovery" *National Public Radio*, 29 November 2012.)

[17] Of course, knowledge is just one part of what makes an agent who she is. I do not mean to suggest that it is any more morally relevant than, say, the emotional constitution one has developed over time or the web of relationships one has built. Indeed, I mean to construe knowledge as broadly as possible, so as to potentially include certain emotional capacities and relationships.

[18] Must Peggy and Roz be *uniquely qualified* for the jobs in order for the obligation to be triggered? I think not. It is enough that they are *qualified*, and that no one else has stepped up (that is, they are *uniquely willing*). Perhaps others are also obligated, but are shirking their duties. This does not seem to weaken the force of the obligations for Peggy and Roz. This is a tricky issue though, and I cannot do justice to it here. Suffice it to say that issues of unique qualification and non-compliance need to be sorted out for obligation *in general*. Knowledge-based obligations do not raise *special* problems here.

path? Quitting would seem to be an all-too-convenient way to escape their moral obligations.

10.3 QUITTING WITHOUT BLAME

The most straightforward solution to this puzzle is to modify the QAT principle. Although freedom to quit is an important component of our general freedom to shape the course of our own lives, and although it is very tempting to think that we can (other things being equal) quit at *any* time and for *any* reason, we may be too quick to let ourselves off the moral hook. Once a knowledge-based moral obligation is triggered, as with *any* moral obligation, we must fulfill it on pain of being blameworthy.

Of course, knowledge-based moral obligations are, like all moral obligations, defeasible. A knowledge-based obligation might be defeated if one cannot discharge it while reasonably maintaining one's mental health, financial security, or familial relationships. For example, suppose Peggy's work with the State Department exposes her to repeated death threats that, though unlikely to amount to anything, will cause her significant psychological distress, to the point where she cannot sleep at night. This would surely be a defeater if we thought it would compromise Peggy's ability to do a good job; in such a case there would be a violation of *obligated implies can*. But it may even be a defeater if we thought that Peggy could continue to do a good job, albeit while suffering a kind of psychological torture. In this case the obligation would be defeated simply because it asks too much. This can all take place from within the "moral point of view;" we need not reach any conclusions about how to weigh Peggy's moral obligations against, for example, the counsels of prudence.

A knowledge-based obligation might also be defeated if it conflicts with some other, stronger obligation, even if the knowledge-based obligation were triggered first. For example, suppose Eve is an astronaut with special training in how to fix part of the International Space Station. She might have a knowledge-based obligation to follow through with a planned mission, even if she decides before departure that she would rather be an ice skater. But if her only child then becomes gravely ill, her obligation as a mother to be at her child's bedside could surely override and thus defeat her obligation to go on the mission. She cannot do both, and she *must* be with her child. So she cannot go on the mission. So the obligation to go on the mission would violate *obligated implies can*. And so the obligation is defeated.[19]

After accounting for defeaters, many knowledge-based obligations remain. So we need a modification of the QAT principle—one that makes it consistent with the existence and strength of these obligations. I propose the following principle, which we can call *Quitting Without Blame* (QWB):

> (QWB) Once knowledge-based obligations associated with one's career or life path have been triggered, one is morally permitted to quit or change course *only* if one fulfills one's existing or immediate particular obligations and then either (1) switches to a life path that will allow one to make an approximately equally significant moral impact *or* (2) ensures that the knowledge-based obligations are taken over by a comparable substitute.

This may initially seem like a rather high bar for quitting. In a case like Roz's, where her obligation to continue performing surgery seems not to be defeated in any of the usual ways, she would need to meet one of the two conditions in QWB in order to change course blamelessly. After performing the surgeries in her existing queue, she would need to either find a suitable permanent replacement for herself, or switch to a life path that allowed her to make an approximately equal moral impact. Would writing a book on French literature that a few dozen people might read count as making an equal moral impact? Probably not, though a wide variety of other things might.

Of course, this is asking a lot of Roz. If she had studied French literature from the beginning and had never gone to medical school, we would not be bothering her; after all, we believe in LPF. But it would also be asking a lot to say that Roz is morally required to remain as a surgeon when there are other paths she would like to pursue. What the

[19] Rather than calling these circumstances "defeaters" that "nullify" the obligation, some will insist that the obligation exists, but the agent is not blameworthy for failing to fulfill it, because she has an *excuse*. But if *obligations* are demands that we legitimately make of one another, I do not see how it could be *legitimate* for us to demand *both* that Eve the astronaut fix the space station *and* that she stay home with her sick child. Nevertheless, it does not matter for my argument which framework we use. In the language of excuses, I would simply say that: needing to be at her child's bedside excuses Eve from her obligation to fix the Space Station, but wanting to be an ice skater does not.

QWB principle allows her to do is to *transfer* the burden of her expertise to some other life-path.

The two conditions of QWB are intended to preserve what is most compelling about QAT while doing justice to the moral force of knowledge-based obligations. What is most compelling about QAT? First, the *quitting* part: it allows us not just to change course, but even to slow down, stop, or downgrade what might be a morally admirable life path, without having violated some general obligation to "stay the course."[20] QWB preserves this by allowing even those with significant knowledge-based obligations to make a moral downgrade, provided that they satisfy the second condition and find a substitute. In this sense, those who bear a "burden of expertise" are just as free to quit as anyone else, albeit in less dramatic fashion: rather than walking out on a whim, they must transfer their burden responsibly. The second compelling element of QAT is the *any time* part. The idea here is that part of our freedom to shape our lives is essentially *temporal*. If you have to stick around for three years to train your replacement, your life path is held hostage to an old decision—you are less free. In QWB, those with knowledge-based obligations can quit at *any time*—possibly the same day!—if they discharge their existing particular obligations and then satisfy the first condition by switching to a morally comparable career.

Of course, QWB still imposes a thicker level of moral bureaucracy, as it were, than QAT did. And QWB does not allow an agent *both* to quit immediately without finding a replacement *and* to make a moral downgrade. So the agent in a sense has to choose just one of the two compelling kinds of freedom that the original principle provided. This compromise is necessary in order to preserve the BEX.[21] Indeed, QWB reminds us that agents need to take their knowledge-based obligations at least as seriously as they take their other moral obligations, including the obligations set aside by the "other things being equal" clause of

[20] Recall, however, the "other things being equal" clause in QAT: we may have pre-existing, special, contractual, or other ancillary obligations that render quitting impermissible.

[21] QWB is offered as a *proposal*, and unfortunately there is not space to defend it against various alternative proposals. Specifically, one might legitimately worry that QWB favours preserving the strength of BEX when one could instead propose a principle that weakens BEX in order to preserve the intuitions behind QAT. While I have tried to show that BEX is uncontroversial, it does depend on a fairly rigid conception of the strength and scope of our moral obligations, and it does privilege the notion of *obligation* over weaker forms of moral recommendation. I intend to explore these issues further in future work.

QAT. We said earlier that you cannot quit nursing in order to become a poker player if doing so means violating a special obligation to take care of your children. What we can see now is that you also cannot quit nursing to become a poker player if your nursing expertise has triggered a robust set of knowledge-based obligations; just like the obligations to your children, you must discharge or transfer these responsibly before it is permissible to quit.

10.4 TIMIMG, THRESHOLDS, AND TRIGGERING

I have argued that there are serious moral implications to the knowledge we acquire in the course of our lives. Not only does knowledge frequently trigger new obligations, but these obligations are especially stubborn. Perhaps without even realizing it, we may find ourselves in a situation where what we have learned actually constrains our *freedom*, or at least our freedom to shape our lives without the weight of moral criticism and blame. This moral landscape might seem rather oppressive, but just how oppressive it is depends on answers to certain questions about the timing of knowledge-based obligations. In this section I address two such questions: (1) Do prospective knowledge-based obligations carry forward into the present, generating an obligation to learn? (2) At exactly what point during the knowledge-acquisition process do knowledge-based obligations become triggered?

10.4.1 Prospective knowledge-based obligations

If knowledge triggers obligations, and if I am interested in being a good person, it would seem that I should acquire as much knowledge as possible! But even if I am only aiming to be a minimally decent person, we might wonder whether it is permissible for me to forgo various opportunities for formal or informal education, especially if I know that such education would bring moral burdens.[22] Of course, there are various moral and prudential arguments for why education is good, whether for its own stake or as an instrument to other valuable ends. But the question here is whether, specifically, the existence of hypothetical future

[22] Elsewhere, I have argued that one particular way of increasing our obligations is by learning about real-life moral saints. They show us what it is like to take on great sacrifices, alleviating a kind of ignorance we otherwise had, which had been acting as a defeater of certain obligations (Carbonell, 2012).

knowledge-based obligations renders a person *obligated* to take certain educational paths. This could be seen as a version of the question of when ignorance is culpable, but on a much grander scale. It is one thing to look backwards at a moral failure resulting from ignorance, and ask: should the agent have known better? It is quite another to look into the uncertain future, where various types of knowledge might trigger obligations, but only in combination with the right contingent moral and non-moral background conditions, and ask: what must one learn?

Suppose, for example, that Esther, a law student in Minnesota, is contemplating taking classes in a Hmong dialect.[23] There are tens of thousands of Hmong refugees living nearby, and Esther believes that if she gains some fluency in the language, she will be obligated on account of her rare expertise to volunteer in her law school's legal clinic helping Hmong clients who have nowhere else to turn. Given this prospective knowledge-based obligation, is she obligated to sign up for the class? I think we should be skeptical that she is, for at least two reasons. First, the background conditions grounding the prospective obligation could change during the year it takes to study the language. Apple or Google might produce an artificially intelligent translation tool so effective that it renders human interpreters unnecessary. Or Esther might acquire other, stronger moral obligations that displace this new one. These may be remote possibilities. But the point is that there is a high bar for obligation: we may think Esther *ought* to learn the dialect, but this cannot be the decisive ought of *obligation*, given how much uncertainty is involved.

Second, there are an infinite number of hypothetical knowledge-based obligations. There are simply too many of them for it to make

[23] Since my examples involve doctors, lawyers, teachers, and so on, one might naturally question whether these knowledge-based obligations are simply "role obligations" in disguise. A role obligation is "a moral requirement, which attaches to an institutional role, whose content is fixed by the function of the role, and whose normative force flows from the role" (Hardimon 1994, 334). Certainly the two are related. Often one comes to occupy a certain role in virtue of having certain knowledge. Nevertheless, not every role obligation reduces to or presupposes a knowledge-based obligation—think of the role obligations associated with being a daughter, which may exist regardless of any particular knowledge. And not every knowledge-based obligation reduces to or presupposes a role obligation—think of a knowledgeable retired teacher stranded in an airport, who finds herself obligated on account of her knowledge to teach the fellow passengers a lesson (in survival skills, or critical thinking, or whatever) despite not occupying the role of "teacher" with respect to them. There is more to say here, of course, but to do the issue justice will require another paper.

any sense that they could generate decisive reasons for action in the present without being unreasonably demanding. Learning the Hmong dialect is just one of countless things Esther could learn, and helping the Hmong in the law clinic is just one of countless morally worthy things she could do if she had the right knowledge, skill, or expertise. Even if we limited the set of potential activities to just things Esther has some antecedent interest in learning, or to skills that are desperately needed in her area, there is a further worry: choices such as whether to learn a new language or take on a long-term volunteer endeavor are precisely the sorts of decision that are meant to be covered by LPF. Recall that LPF said that we are not morally required to choose the morally best life path. Now, it is compatible with LPF that Esther has *good reason* (albeit not an obligation) to work at the law clinic. But this reason presumably is grounded in the needs of the refugees, and not anything having to do with Esther's knowledge or skill, which she does not yet possess. It seems that not only is there not an obligation here, but what moral reasons there may be are grounded in considerations independent of the prospective knowledge-based obligation. The knowledge-based obligation *itself*—or rather, the fact that it would obtain in the hypothetical situation in which Esther had the knowledge—cannot entail an earlier obligation to learn Hmong.

10.4.2 Thresholds

If knowledge-based obligations do not take effect *before* the knowledge is acquired, then when exactly *do* they take effect? Often there is no simple fact of the matter about *whether* and *when* a given agent *knows p* or *knows how to* φ. Learning is typically a process, not an event. When we say that an agent knows how to speak Spanish, or knows how to fix the Space Station, or can perform a certain surgery, or is an expert on Waziristan, we are making a claim that is inherently vague. Must Roz's surgeries *all* be successful? Must Eve know how to fix *every* part of the Space Station? If not, then we need to know at what point in their training Roz and Eve became *good enough* to trigger knowledge-based obligations. This question is important, because once these obligations are triggered, I have argued that you need to meet certain conditions in order to quit or change course without blame (QWB). But what if you want to quit or change course *during* your training? Can you then invoke QAT?

What is at stake here is moral obligation, and I have been understanding moral obligation to be in part a social phenomenon in which we hold each other, as equals, mutually accountable for behaving in accordance with reasons that are shared and public.[24] Accordingly, I think it makes most sense to see knowledge-based obligations as being triggered by certain formal and publicly accepted *thresholds*. These thresholds include milestones, degrees, tests, or credentials within a field that are considered to be either indicative of, or in fact constitutive of, mastery of a given body of knowledge or skill. Paradigm examples of such thresholds would be completing the CPR course, passing the Bar Exam, graduating from medical school, becoming a certified dog trainer, and so on. Because we all know the purpose of a CPR course, and we trust those who design it to ensure that passing it is a reliable indication of competence, it is reasonable for us to expect that the one bystander among us who has passed the course take on responsibility for intervening in an emergency.

However, while these formal thresholds are sufficient for the triggering of knowledge-based obligations, it is worth asking whether they are always necessary. After all, there is a degree of arbitrariness in setting the thresholds, and they are often intended to define the scope of *professional* (or even legal) roles and responsibilities, not *moral* responsibilities. Someone who fails the Bar Exam by one point cannot practice law, but she may know just as much about how to research the case of a wrongfully accused prisoner as someone who passed with flying colors. Since knowledge-based obligations are supposed to be triggered by the acquisition of the relevant knowledge itself, we should admit that in some cases this knowledge will be acquired before, or in the absence of, the passing of a public threshold. Thus *partially-trained* agents can incur knowledge-based obligations.

The problem, though, is that as soon as knowledge-based obligations are triggered, a person's freedom to shape her own life is constrained, because QWB requires her to find a substitute or a morally comparable alternative path. If you learn enough in the first semester of nursing school to be bound by QWB, what good is LPF? Our choices about what paths to follow would be ominous indeed if we

[24] See fn. 14. For a more detailed discussion of why questions of moral obligation require appeal to mutually intelligible values or reasons, see Carbonell (2012).

were stuck after just one semester. Given that knowledge is so crucial to developing projects of moral value, it makes sense to protect the knowledge-acquisition process itself from the reach of those obligations, lest no one ever achieve real expertise. In order to avoid a system of obligations that is self-defeating in the long term, it makes sense to say that partially trained agents can *only* incur knowledge-based obligations if (1) no fully-trained agents are available, and (2) discharging these obligations does not interfere with the completion of the training itself. For example, a third-year medical student may do a month-long internship in a rural area and find that her nascent skills are in such dire need that she bears a "burden of partial expertise." She may incur a knowledge-based obligation to work extra hours while she is there, or to go beyond her minimal job duties, or return on her next vacation. But we would not want to say that this burden requires her to quit medical school and move to the rural area immediately, as this would undermine her acquisition of a level of expertise that might enable her to perform even more urgently needed moral duties in the future.

Of course, many important kinds of knowledge are acquired informally, without thresholds of any kind, arbitrary or otherwise. For example, one who lives in a struggling neighborhood may, over time, acquire knowledge about the needs and concerns of its residents. This knowledge could ultimately trigger obligations—to vote a certain way in a city council election, to participate in a rally, and so on. In the absence of any well-defined threshold, we will have to say that these obligations get triggered individually, that is, whenever it becomes the case that there is something of moral value that one can do, and that one has otherwise nearly decisive moral reason to do, but that one could not do until one knew how—and now one does (more or less) know how. If these individual actions culminate in a large project—say, playing a leadership role in the neighborhood association—then one might have to meet the conditions of QWB if one wants to quit or change course. Such judgments are highly contextual, of course. Nevertheless, the general framework I have provided can apply not only to well-defined careers and formal courses of education, but also to the vast body of equally important non-codified knowledge that enables people to take on projects of moral value.

10.5 CONCLUSION: FREEDOM AND MORAL DEMANDS

The view I have put forward in this essay tries to reconcile our freedom to choose our own career or life path without moral blame (LPF), our corresponding freedom to quit and change course (QAT/QWB), and the fact that many of our life choices bring with them burdensome moral obligations (BEX). The compromise I have articulated is, admittedly, quite demanding. One might worry, though, that the view is worse than unreasonably demanding: that it is unsatisfactory on its own terms. In replacing QAT with the more demanding QWB, have we not violated LPF? QWB says that you must trade your life path for one that is approximately morally equal; you cannot downgrade without finding a moral understudy for yourself, which is likely to be more difficult as your level of expertise increases. In this way, QWB might be thought to punish excellence by constraining the freedom of precisely those agents who are already bearing large moral burdens.

In fact, QWB does not conflict with LPF. LPF only says that you are free to choose something other than the morally best path. QWB does not require you to choose the morally best path—though if you are that rare person who is *already* living the morally *best* life, QWB says that you can only change to one that is tied for best. To be sure, QWB significantly constrains one's freedom; but LPF is compatible with this reduced freedom. LPF permits you to choose a less-than-best path, but it does not say you can choose any path whatsoever. Perhaps the real worry here is that LPF was too weak to begin with: maybe we *do* think you should be able to choose any path whatsoever. But the more freedom we build into LPF, the more it will conflict with the BEX principle, which I have shown is straightforwardly true.

Some people simply will not be able to discharge all of their knowledge-based obligations while maintaining complete freedom to choose the course and shape of their lives, especially after they cross thresholds of competence. The problem is not that QWB is too restrictive, but that knowledge—combined with the right background conditions—makes possible such burdensome moral obligations in the first place. Thus the view leaves us with a version of the familiar problem of morality's demandingness. I cannot solve this problem. But if gaining expert knowledge puts you in a position to contribute to morally beneficial projects, to fight injustice, or to lessen suffering, then the simple fact that you ought to continue doing so seems to be not such a bad problem to have.[25]

REFERENCES

Sarah Buss (1997). "Justified Wrongdoing," *Nous* 31 (3): 337–69.

—— (2007). "Needs (Someone Else's), Projects (My Own), and Reasons," *Journal of Philosophy* 103 (8): 373–402.

Vanessa Carbonell (2012). "The Ratcheting-Up Effect," *Pacific Philosophical Quarterly* 93 (2): 228–54.

Stephen Darwall (2006). The Second-Person Standpoint. Cambridge, MA: Harvard University Press.

Lauren Frayer (2010). "JetBlue Flight Attendant Who Flipped is a Folk Hero," AOL News, 10 August 2010. <http://www.aolnews.com/2010/08/10/flight-attendant-who-flipped-becomes-folk-hero/>; accessed 1 December 2012.

Michael O. Hardimon (1994). "Role Obligations," *Journal of Philosophy* 91 (7): 333–63.

Mike Judge (Director) (1999). *Office Space*. DVD. 20th Century Fox.

Shelly Kagan (1989). *The Limits of Morality*. New York: Oxford University Press.

Joel Rose. *"N.Y. Electrician Shortage Hampers Sandy Recovery,"* National Public Radio, 29 November 2012. <http://www.npr.org/2012/11/29/166139923/n-y-electrician-shortage-hampers-sandy-recovery>; accessed 14 December 2012.

Samuel Scheffler (1982) *The Rejection of Consequentialism*. New York: Oxford University Press.

Holly Smith (1983). "Culpable Ignorance," *The Philosophical Review* 92 (4): 543–71.

—— "The Moral Clout of Reasonable Belief," in M. Timmons, ed., *Oxford Studies in Normative Ethics*, *Vol*1. New York: Oxford University Press.

Gus Van Sant (Director) (1997). *Good Will Hunting*. DVD. Miramax Studios.

Bernard Williams (1973). "A Critique of Utilitarianism," in Smart and Williams, *Utilitarianism: For and Against*. Cambridge: Cambridge University Press.

Susan Wolf (1982). "Moral Saints," *Journal of Philosophy* 79 (8): 419–39.

[25] I am grateful to Marie Jayasekera, Amanda Roth, and Dustin Locke for their generous comments on early fragments of this essay. I would also like to thank the audience at the Third Annual Arizona Workshop on Normative Ethics for helpful discussion. With apologies for those whose names I have forgotten, let me especially mention Teresa Blankmeyer Burke, Cheshire Calhoun, Don Fallis, Daniel Groll, Thomas Hurka, Howard Nye, Douglas Portmore, and Holly Smith. Finally, I would like to thank two referees for Oxford University Press, one of whom revealed himself to be Andrew Schroeder, for valuable feedback which improved the essay considerably. Work on this essay was supported by a Center Fellowship at the University of Cincinnati's Charles Phelps Taft Research Center.

11

Objective Double Effect and the Avoidance of Narcissism

HOWARD NYE

The Doctrine of Double Effect [DDE] states roughly that it is harder to justify causing or allowing harm as a means to an end than it is to justify conduct that results in harm as a side effect. The DDE is typically interpreted as maintaining that there are stronger moral reasons against causing or allowing harms with the intention of doing so than there are against causing or allowing harms that we foresee but do not intend. Let us call any such reading of the DDE, according to which our moral reasons and permissions depend upon our intentions, a *subjective reading*.

Over the past several decades, various authors have argued convincingly that our moral reasons cannot depend upon our intentions in the way that subjective readings allege.[1] It is, however, possible to interpret the DDE as maintaining that it is harder to justify causing or allowing harm as a means because of something about the objective explanatory relationship between the effects our conduct has on those it harms and the effects it has on those it benefits. Such *objective readings* have, however, been largely dismissed as absurd.[2]

In this essay I defend an objective reading of the DDE. I argue that a theory of deontological constraints on harming needs something like the DDE in order to avoid the charge that it reflects a narcissistic obsession with our personal moral purity instead of an appropriate concern for the welfare of others. But, I contend, the central problem with subjective readings of the DDE is that, by making our own intentions more

[1] See especially Thomson (1991, 292–6) and Scanlon (2008). See also Ross (1930, 4–6) and Bennett (1981, 96–8; 1995, 194–6).

[2] See Frey (1975, 279–83), Bennett (1995, 198–9), and Norcross (1999, 115–17).

important than the welfare of others, they embody exactly this kind of implausible narcissism.

I believe that objective readings of the DDE have been dismissed primarily because they have been construed as claiming that the explanatory relationship between an act's harmful and beneficial effects is a reason against performing it. But I suggest that we should instead interpret the DDE as a denial of the Machiavellian dictum that the ends justify the means. On this reading, the DDE holds that the benefits of our conduct for some individuals do not count (as strongly) in its favour if they come at the expense of others. This, I argue, makes best sense of the original motivations for the DDE and provides a plausible foundation for deontological constraints.

11.1 WHY DOUBLE EFFECT?

The DDE is often invoked as a justification of two kinds of intuitions about particular cases. The first are intuitions about the *permissibility of collateral damage*. Many intuit that it could be permissible to engage in tactical bombing that one knows will destroy enemy military installations and kill civilians as a side effect, yet wrong to target civilians in a terror raid that one knows will result in identical benefits and harms. Many also intuit that it would be permissible to divert a trolley from a main track on which it will hit five people onto a side track where it will hit one, but wrong to stop a trolley from hitting five people by pushing a fat man into its path. Since the terror bomber and fat man pusher seem to cause harm as a means while the strategic bomber and trolley diverter seem to cause harm as a side effect, the DDE appears to provide a natural justification of these intuitions.[3]

The other kind of intuitions the DDE is invoked to justify concern certain kinds of *wrongful omissions*. It seems permissible to withhold a supply of life-saving drugs from one individual who needs the entire supply in order to give it to five others who each need only 1/5.[4] But it seems wrong to withhold life-saving drugs from an individual infected

[3] See Foot (1967, 23–4), Costa (1986), Shaw (2006), Bennett (1981, 95), and Quinn (1989b, 336).

[4] Or at least to flip a coin to decide what to do (Taurek 1977). In what follows, Taurekians can replace my talk of the permissibility of benefiting a greater number with talk of the permissibility of flipping a coin.

with an old strain of a disease in order to observe its fatal progression and learn how to cure five people infected with a newly mutated strain. A plausible justification of these intuitions is that failing to give the drugs to the one allows harm to her as a means of saving the five in the second case (where they would not be saved unless her disease progressed), but as a side effect of saving them in the first (where her affliction plays no role in saving them).[5]

I am, however, quite skeptical of our intuitions about permissible collateral damage. If the salvation of five individuals can justify inflicting harm on one as a "side effect," we would seem justified in driving over one trapped on a road if that was the only way to save five from drowning, but this seems about as abhorrent as pushing the fat man into the trolley's path. To many it also seems wrong to save five by performing a surgery that will release lethal fumes into a room in which one is trapped, or by destroying a trolley headed towards them with an explosion that you know will envelop a bystander.[6]

Moreover, I do not believe that we should in any event treat our intuitions about what it is wrong or permissible to do in particular cases as a kind of data that our moral theories must fit. Our moral judgments are subject to a host of distortionary factors, such as indoctrination, the asymmetric salience of different considerations, and confusions in understanding and reasoning. The only way in practice to determine whether our convictions reflect these biases is to determine whether they can be supported by general ideas and principles that are directly plausible, or seem true independent of inference.[7] Our reasons to accept general principles have more to do with whether their plausibility can survive critical clarification and integration with other plausible ideas than whether they match our pre-theoretical intuitions about cases.

[5] See Foot (1967, 24–5) and Quinn (1989b, 336).

[6] See Foot (1984, 179, and 1967, 29). Principles have been suggested that would permit "diverting threats" without permitting these acts (see Thomson 1976, 216–7—who has (2008) renounced her proposal—and Kamm (2007, 147). But these principles attribute intrinsic moral relevance to such factors as the identity of the material with which you harm someone (was it the same as that which would have harmed the five?), how close it was to her (was it "in her context?"), and how "directly" it harmed her, which—like your victim's skin color—seem *obviously* devoid of intrinsic moral relevance. It thus seems to me highly unlikely that our intuitions about permissible collateral damage reflect defensible moral ideas. I suspect that they reflect instead the greater salience of the good we are doing in relation to the harm we are inflicting, and the fact that in real life acts that risk harming as a side-effect are more likely to secure goods and less likely to inflict harms than acts of inflicting harm as a means.

[7] See Singer (1974, 515–17).

What is most important about the DDE is, I believe, that it is essential to a defensible theory of deontological constraints on harming.[8] If there are any deontological constraints on harming, I take it that they must explain why it is permissible to do things such as saving five drowning swimmers rather than one, but wrong to push a fat man in front of a trolley to save five. Perhaps the most intuitively obvious account of this is the Doctrine of Doing and Allowing [DDA], according to which there are stronger moral reasons against inflicting harm than there are against failing to prevent harm.[9] To many of us, this seems directly plausible.

But there is a major threat to the plausibility of the DDA. As Kai Nielsen (1972, 330) suggested, it can seem to be telling us simply to retain our own "moral purity" by avoiding "dirty hands." Consider:

Alastair and the Fat Man. A trolley is headed toward five people. You see a utilitarian named Alastair sneaking up on an oblivious fat man, ready to push him onto the tracks in order to stop the trolley. You can keep quiet while Alastair pushes the fat man, or call out to warn him of Alastair's approach.

By itself, the DDA seems to entail that you are forbidden to push the fat man yourself, but allowed, out of concern for the five, to keep quiet and let Alastair do the dirty work. It seems to permit you to benefit the five at the fat man's expense by staying out of Alastair's way, just not by taking Alastair's place. As such, the DDA threatens to embody a narcissistic obsession with your personal purity instead of an appropriate concern with how you treat others. Call this *the dirty hands objection*.

To avoid the dirty hands objection it seems that a theory of deontological constraints on harming must explain why it is wrong, not only to push the fat man, but to let Alastair do it. You might think that this is because Alastair's pushing would be wrong, and you would be complicit in this wrong if you fail to prevent it. But this seems incorrect. First, suppose that you could save 1,000 people from being killed by a murderer or 1,001 from being killed by a natural disaster. If failing to prevent wrong

[8] That is, a theory according to which certain ways of causing or allowing harm are in themselves or intrinsically harder to justify than others, in a way that does not simply reflect which acts or policies would bring about the best states of affairs.

[9] Where the reasons against inflicting harm are *sufficiently stronger* that they decisively outweigh the reasons to save five. When I speak of 'the DDA' and 'the DDE' I have in mind versions of this kind.

was so much worse than simply failing to prevent harm, you would be required to save the 1,000, but surely you are permitted to save the 1,001.[10] Second, it would not make a moral difference if Alastair were replaced by a humanoid robot, a giant Roomba, or even a moving steel rod. The claim that it is permissible to let these things do the dirty work of pushing the fat man onto the tracks but wrong to do it yourself is equally vulnerable to the dirty hands objection.

It appears that a plausible account of the why it is wrong to let something push the fat man must make reference to something like the fact that you would be passively using his death as a means of saving the five. It must, in other words, invoke the basic idea of the DDE.[11]

There is, of course, a famous problem with the idea that in saving the five by allowing the fat man to be pushed, you use his death as a means of saving them. For what saves the five is simply the collision of the trolley with his body; the fact that this injures him, and certainly the fact that he suffers the harm of death some moments later, plays no role in their salvation. As several authors have pointed out, harm itself is almost never a means in cases to which the DDE is applied; what is a means is only something intuitively "close to" harm.[12] The *problem of closeness* for proponents of the DDE is to give a plausible and principled clarification of the

[10] See Quinn (1989b, 347) and McMahan (2009, 358). To explain why it is *wrong* to let Alastair push the fat man, reasons against allowing wrongful death would have to be so important as to make it *wrong* to prevent four more non-wrongful deaths instead (and thus certainly wrong to prevent only *one* more non-wrongful death instead).

[11] In his attempt to justify intuitions about wrongful omission without the DDE, Scanlon (2008) considers exactly one very special kind of case, where to save five you must allow a victim to die, after which others will harvest his organs. Scanlon simply *assumes* we have an obligation not to take (or, evidently, allow the taking of) a *living* person's organs to save five, and notes that it would be crazy to think "the advantages of our being relieved of this obligation by his dying... justify an exception to the principle requiring us not to kill that person, or to save [his] life when we can easily do so" (33–5). But the question is *why*, if we could save five others by not saving the single person, there *is* any principle "requiring us... to save [his] life when we can easily do so." There is no principle that requires this in the case of saving five drowning swimmers rather than one, so why would there be such a principle in the case of saving five by not saving the one whose organs will be used to save them? The DDE (unlike Scanlon) offers a principled answer to this question, which applies to other cases of wrongful omission, like allowing someone's disease to progress or allowing her to be pushed in front of a trolley, where Scanlon's remarks about "allowing someone to die to relieve ourselves of an obligation we have while she is alive" are obviously inapplicable.

[12] See Foot (1967, 21–22), Bennett (1981, 98–116; 1995, 201–13), and Quinn (1989b, 336–44).

Doctrine according to which it is difficult to justify using as a means not only harm but the relevant sort of thing "close to" harm.[13]

There are several proposals about how to solve the problem of closeness, a detailed discussion of which is beyond the scope of this essay. For now I wish only to indicate that (1) anyone who wishes to give a plausible response to the dirty hands objection is in the business of solving the problem of closeness, and (2) the problem is not utterly hopeless. To answer the dirty hands objection we must explain why it is wrong to allow something to push the fat man into the path of the trolley to save five. But to explain why it is wrong to do this it seems that we must appeal to the idea that this would involve something like harmfully using the fat man as a means, saving the five at the fat man's expense, or sacrificing the fat man to save the five.

The best solutions to the problem of closeness identify the DDE's root idea with the claim that it is particularly difficult to justify treating individuals in one of the foregoing ways, and to seek to make the idea precise. Warren Quinn's (1989b) "harmful involvement" solution is an excellent example. Quinn argued that we should interpret the DDE as claiming that it is particularly difficult to justify harmfully using someone as a means. His precise suggestion was that we harmfully use someone just in case we involve her in our plans—or use as a means her instantiation of some property—which involvement or instantiation in fact causes her to be harmed. Thus, when you save five swimmers rather than one, the one's instantiation of properties that harm him (*being in the water*, *slipping beneath the waves*) are completely immaterial to your salvation of the five. But when you save five from the trolley by allowing the fat man to be pushed in front of them, you use as a means the trolley's striking him, and this does in fact harm him.

[13] Some authors (Fitzpatrick 2006, Shaw 2006) seek to solve the problem of closeness by maintaining that in the relevant cases events that constitute harm would be means, so harm itself is a means. It is, however, preposterous to maintain that the event of the fat man's dying is identical to or constituted by the event of his being impacted with the trolley. The former could take place several minutes or hours after the latter and at a completely different location (if he were rushed to hospital). And it is preposterous to claim that the moral barriers to using as a means a "harm" like mere violent impacting, quite independent of death, are anything like the moral barriers to using death as a means. If you could quickly anesthetize the fat man before pushing him and, after the trolley had hit him, quickly reassemble his body so he awoke a few minutes later without noticing a thing, I venture that everyone should agree that you would be *required* to do so.

An alternative solution can actually be derived from what Quinn (1989a) misleadingly characterized as a version of the DDA. On this view, it is particularly difficult to justify not only actions that produce harm, but deliberate failures to prevent "actions of objects or forces over which we have control" that produce harm.[14] Quinn's idea seemed to be that it is difficult to justify benefitting some by deliberately producing or allowing events that produce harm to others.[15] Thus, when you save five swimmers rather than one, the events that produce the one's death (his remaining in the water, his slipping beneath the waves) play no role in saving the five, so you need not intend them. But saving the five on the track by allowing the fat man to be pushed requires events that produce his death (his moving in front of the trolley, his being struck by it), which you must intend as means. While Quinn presents this as a version of the DDA, it does not seem that in allowing the fat man to be pushed you inflict harm on him in any familiar sense. Quinn (1989a, 300) notes that in deliberately allowing an event that produces harm, your agency seems implicated in the harm, but this seems more like a Double Effect idea than a Doing/Allowing idea. Indeed, I think it is plausible that our intuitive distinction between:

(i) benefitting some at the expense of (or by sacrificing) others, and
(ii) simply benefitting some instead of others

can be made precise by something like Quinn's distinction between

(i') saving some in virtue of ensuring the existence of events that produce harm to others, and
(ii') saving some simply in virtue of failing to prevent events that produce harm to others.[16]

[14] Ned Hall (2004) suggests that there are at least two concepts of causation: counterfactual dependence and "production." He characterizes "production," in the actual world, as obeying transitivity, locality, and intrinsicness; perhaps it is something like conserved-quantity transfer (Dowe 1995) or trope-persistence (Ehring 1997). I use "production" to refer to what Hall would call "production *and* dependence."

[15] More precisely, it is difficult to justify "most proximate contributions" or effects of one's conduct on the whole that have these effects (Quinn 1989a, 301–2).

[16] There are, however, reasons to broaden (i') to include saving some in virtue of ensuring certain events that result in harm through what Hall (2004) calls "double-prevention." Whether certain acts that double-prevent harm are morally akin to inflicting harm is debated (see McMahan 1993 and Hanser 1999). A satisfactory specification of (i') would require a resolution of these difficult issues.

Since a plausible root idea of the DDE is that it is harder to justify (i) than (ii), precisifying this as a moral distinction between (i') and (ii') seems like a promising solution to the problem of closeness.

While these Quinnian solutions to the problem of closeness are certainly controversial, they are, I believe, promising enough to provide hope that the problem is tractable.[17]

11.2 AGAINST SUBJECTIVE READINGS

I have thus argued that, in order to avoid the dirty hands objection, proponents of deontological constraints must explain why it is wrong to save five by allowing something to push a fat man in front of a trolley, that to explain this we must appeal to something like the ideas of harmfully using or sacrificing the fat man as a means, and that these are best interpreted as "root ideas" of the DDE, precisifications of which will constitute solutions to the DDE's problem of closeness. I believe, however, that there are at least two sound arguments against the most common understandings of the DDE, which read it subjectively as the claim that there are stronger moral reasons against causing or allowing harmful effects with the intention of doing so than there are against causing or allowing these effects with the foresight but without the intention of doing so.

The first, which I will call *the volitional argument*, runs roughly as follows:

(P1) What there is stronger or weaker moral reason to do must be something we can voluntarily choose to do.

(P2) We cannot voluntarily choose to have certain intentions. So we cannot voluntarily choose to perform acts with certain intentions.

(C1) Therefore, acts performed with certain intentions cannot be what there is stronger or weaker moral reason to do.

[17] For excellent criticism of Quinn's (1989b) proposal, see Bennett (1995, 218–21). Using Quinn's (1989a) proposal for the DDA as an understanding the DDE would violate the assumption that there is an intrinsic moral difference between terror raids and tactical raids that are known to have the same consequences (as well as the assumption that it is permissible to divert the trolley—though Quinn mistakenly thought otherwise). But as I indicated in fn. 6, I suspect that no credible principles can support these assumptions.

(C2) Therefore, it cannot be the case that there are stronger moral reasons against causing or allowing harmful effects with the intention of doing so than there are against causing or allowing these effects without this intention (subjective readings of the DDE are false).[18]

(P1) seems to follow simply from the practical nature of deontic[19] assessment: what we are asking about in trying to determine what there is moral reason to do is, well, what to *do*: the sort of thing we can choose or will. (P2) also seems obvious, and is illustrated by Kavka's (1983) toxin puzzle, in which an eccentric billionaire will pay you $1 million if at midnight tonight his completely reliable brain-scanner detects that you have an intention to drink a toxin tomorrow morning which will make you sick for a day. Try as you might, you will *not* be able to form the intention simply in response to the consideration that having it will get you $1 million. But if intentions were, like movements of our limbs, under our voluntary control, we could form them simply in response to the fact that they will make us rich in the same way that we can extend our arms (to catch $1 million) in response to the fact that it will make us rich.[20]

While it is important, I think the volitional argument fails to get to the heart of what is wrong with subjective readings of the DDE. First, the fact that we cannot choose our intentions is probably a contingent fact. We can choose to move our limbs because the neurons responsible for moving them are wired in the right way to the neural correlates of

[18] See Ross (1930, pp. 4–6), Bennett (1981, pp. 96–8; 1995, pp. 194–6), and Scanlon (2008, chapter 2).

[19] *Deontic assessments* of moral reasons for and against doing things, and how they stack up to make acts wrong or permissible, are forward-looking, action-guiding judgments about what to do in a situation. They stand in contrast to *aretaic assessments* that look back upon the quality of an agent's reasoning and motivation in acting, and assign esteem and blame accordingly (see Frankena 1963).

[20] I state the volitional argument in terms of the relative strength of moral reasons because that is, I believe, what the DDE is fundamentally a thesis about. But most responses to the volitional argument (McMahan 2009, Wedgwood 2011) have been to versions couched in terms of moral permissibility and impermissibility. By 'permissibility' and 'impermissibility' I mean assessments inextricably linked to moral reasons:

(P3) Something is morally permissible iff it is not decisively opposed by moral reasons, and morally impermissible or wrong iff it is so opposed. So something can be morally permissible or impermissible only if it is something that there can be stronger or weaker moral reason to do.

(P1)–(P3) entail the conclusion directly challenged by critics of the volitional argument:

(C3) Therefore, causing or allowing harmful effects with or without the intention of causing or allowing them cannot be what is itself permissible or impermissible.

voluntary choice. It seems that we can conceive of the neural correlates of voluntary choice being wired to the neural correlates of intention in such a way that we could, on the basis of the good consequences of forming an intention (that it will make us rich in Kavka's puzzle) form it in the same way we can move our arms on the basis of such considerations. But it does not seem that this change in our neural wiring would alter whether a subjective reading of the DDE were true.

Second, the volitional argument leaves open the possibility that something very much like a subjective reading of the DDE is true, namely, a

Modified Subjective Reading of the DDE: There are stronger intrinsic moral reasons against *letting yourself* cause or allow harmful effects with the intention of doing so than there are against *letting yourself* cause or allow harmful effects that you merely foresee.[21]

Even if we cannot make moral decisions about what to intend, if we foresee that we will do something with certain intentions, we can decide to take action to alter those intentions, or decide not to perform the act at all if that is the only way to avoid performing it with problematic intentions. Suppose you are about to save five swimmers rather than your rival, and you suspect that you will do so not only out of concern for the five but in part out of a desire that your rival die. Suppose you then learn that you are being monitored by the Purity Police—a group of demented mind-readers who you know will kill *six* others if you let your rival die with any intention of his dying. Clearly you have very strong reasons in this case to make sure that you do not let your rival die with any intention of his dying; before saving the five you should try to talk yourself into thinking that your rival does not deserve death, or take any mind-altering substances that might remove the intention that he die. If none of this will work, there is a strong case to be made that you should save your rival to minimize the number of deaths.

Modified subjective readings claim that there are powerful *intrinsic* reasons against acting with the intention of harm coming to someone, so we do not need the Purity Police to provide instrumental reasons. On these views it is wrong not to save the fat man because saving him is the only way to avoid allowing him to be pushed with the intention

[21] Bennett (1995, 195–6) makes exactly this point.

of his being harmfully impacted.[22] While this is a coherent position, it seems obviously false. It is plausible that not acting with the intention of someone dying is more important than saving five *if it is the only way to save six*. But how could it be so *intrinsically* important to avoid letting someone die with the intention of her dying that avoiding it *per se* is more important than saving five lives? Suppose you faced the prospect of saving five swimmers or your rival, without any method of purging your propensity to act with some intention of your rival dying if you do not save him, but also without any Purity Police to kill six if you do this. Far from being morally required, saving your rival instead of the five on the grounds that you would otherwise be intending his death looks morally dubious. Even if Taurek is right that you are permitted to save one rather than five, doing so because you would otherwise be intending his death looks like morally bad decision-making. You would be settling a life-or-death question on the basis of a narcissistic obsession with your personal purity rather than an appropriate concern for the welfare of others.

But upon reflection, *un*modified subjective readings of the DDE seem to face an identical problem. Unmodified readings differ only in that, instead of telling us not to *let ourselves* cause or allow harmful effects with the intention of doing so, they tell us simply not to *cause or allow* harmful effects with this intention.[23] Assuming for a minute that we *can* voluntarily control our intentions, why should it be so morally important that, in doing something that results in harm, we choose to do it without rather than with the intention of a harmful effect occurring? Suppose, for example, that in the foregoing case I saved the five rather than my rival, but I forgot to choose to do it solely out of an intention to save them and ended up choosing to do it in part with the intention of my rival dying. Is this really such a big deal? Is it really wrong in anything like the way failing to prevent something from pushing a fat man into the path of a trolley is wrong?

[22] You might also allow this pushing out of sheer indifference, but presumably proponents of modified subjective readings would hold that, so long as you can avoid acting with the problematic intention of someone's being harmed, it is unacceptable not to save someone at trivial cost to yourself.

[23] Wedgwood (2011, 468–9) makes essentially this point. I believe Thomson and Scanlon put their "looking inward" arguments in terms of modified versions because they also accept the volitional argument.

If it is wrong to allow the fat man to be pushed into the path of the trolley to save five (four more than your alternative), it must be wrong to allow him to be pushed in front of it to save two (one more than your alternative). So if what is wrong with allowing him to be pushed is your acting with the intention of a lethal effect on him, it cannot be permissible to choose to act with the intention of a lethal effect on someone to save an additional individual. But consider:

The Impurity Police. You are about to face the situation of saving the five swimmers or saving your rival, and this time you know that your mental states are being monitored, not by the Purity Police, but by the Impurity Police. The Impurity Police credibly promise that they will recue an additional child from being killed by the Purity Police if and only if you choose to save the five rather than your rival *in part with the intention of your rival dying*.

If it is wrong to choose to act with the intention of a lethal effect on someone in order to save an additional individual, then it must be wrong to choose the option of saving the five with the intention of your rival dying over the option of saving the five without this intention. But surely it is *not* wrong to choose to act with the intention that your rival die in order to save the child. *Who cares* if you have this intention? Surely not your rival, who is going to be allowed to drown either way.

Assuming as we have been that you can choose your intentions, *how could you refuse* to choose to save the five with the intention of your rival dying? How could you explain this to the child's parents? You would have to say: "I'm sorry, but the only way for me to save your child would have been for me to allow someone to die with the intention of his dying. It is true that I allowed him to die anyway. But you see this way I was able to choose to allow him to die without intending it." That would be absolutely monstrous. You would betray the fact that you cared more about the purity of your own intentions than you did about their child's very life.

Of course, you might *try* saying: "It is not that I care about my intentions considered in isolation, but you see if I chose to allow my rival to die with the intention of doing so I would have *disrespected* him, and that is, you know, even worse than just allowing him to come to harm." While this might not be selfish, it is, I think, no less monstrously narcissistic to think that your intentions *per se* have this kind of importance. At this point we should bring in your rival's parents, who should tell you: "Rubbish. Our son is dead, and you were going to let him die anyway. He never even knew what was in your heart and would hardly have cared if he did know. It had no tangible effect on him whatsoever. Considered in themselves, insults and disrespect are nothing compared to someone's life. If you could have saved this other child by screaming racial slurs at our son or mocking him as you saved the five it would have been wrong not to do so. Instead, you refused to save this child on the grounds that you would have had this inner state, which our son never even knew about, that was so disrespectful to him that it was more important not to have it than to save this other child? You are seriously sick in the head!"

It seems, then, that both modified and unmodified subjective readings of the DDE are subject to the same central objection: by understanding deontological constraints as moral reasons to be concerned about the intentions with which we act (whether by allowing or simply choosing these intentions), they embody narcissistic obsessions with our personal purity of heart rather than an appropriate concern about what we do to others. I shall call this the *dirty heart objection*. I think the dirty heart objection is particularly damning because, as its name recalls, one of the main attractions of the DDE is its promise to save a theory of deontological constraints from the dirty hands objection that, in forbidding us to push a fat man but allowing us to let other things push him, it embodies a similarly narcissistic obsession with our personal purity.[24]

11.3 RESISTANCE TO OBJECTIVE READINGS

I have thus argued that a theory of deontological constraints on harming is implausibly narcissistic without the DDE, but that it is also implausibly narcissistic with the DDE given its most common (subjective) reading. One reasonable conclusion would be that there are no deontological constraints on harming. But I do not believe that the DDE's bid to save deontology from narcissism has yet been given a fair trial. For I believe that there is a better way to understand the Doctrine.

[24] The dirty heart objection is, I believe, what gives force to Thomson's (1991, 291–2) observation that (modified) subjective readings of the DDE implausibly tell us to "look inward" and decide what to do on the basis of the intention with which we would be acting. (*Un*modified subjective readings tell us to look inward only to make sure we act with the right intention—but given the dirty heart objection this does not seem much more plausible).

Return to the "root idea" behind the DDE that it is particularly difficult to justify benefitting some at the expense of others. Whether the benefits that some individuals derive from your conduct come at the expense of others is actually an objective, intention-independent matter, about which you could be misled. Suppose you thought that by purchasing Soylent Green rather than cheaper food, you were simply benefitting yourself instead of the children you could have saved by donating the price difference to Oxfam. Conveying such trivial benefits on yourself instead of vitally needed benefits on others is opposed by weighty moral reasons, and I believe that there are conclusive arguments to the effect that it is wrong.[25] But suppose you were to learn that (1) Oxfam has been destroyed, and (2) Soylent Green is manufactured by killing children and processing their bodies into the stuff.[26] Somehow this seems to constitute a discovery of an even weightier moral case against purchasing it. That such consumption benefits you, not just instead of, but at the expense of children would seem to make your past consumption even more unjustifiable and ceasing consumption even more urgent than you had thought. If Soylent Green were the only food available, it would seem permissible to purchase it if it were not manufactured at the expense of the children. But given that it is so manufactured, there is a strong case to be made that you may not purchase it even if you will thereby starve.

In such a case it would seem crazy to think, as subjective readings of the DDE suggest, that it is unfortunate that you have obtained your new evidence, in the absence of which you could have consumed Soylent Green in peace without intending a harmful effect on the children. If there is something distinctively problematic that you are doing to the children given your evidence, it seems that you were doing it to

[25] While Singer's (1972, 231) talk of "comparable moral significance" and "moral insignificance" can sound unpersuasive or obscure, I believe that his weakened principle is best interpreted as saying something like: "All else held equal, if you can prevent someone from suffering serious harm by incurring only costs that are absolutely trivial in comparison to what she would suffer, it is morally wrong not to do so." This principle has an *enormous* amount of direct plausibility, and as Unger (1996) has discussed at length, the intuitions it contradicts are extremely dubious.

[26] Assume that these children were not going to die at the same time without your help anyway. (Otherwise their being killed might not harm them, or make them worse off than they would have been. See Williams' "Jim" who can prevent the execution of twenty innocents by killing one of them himself—but see also McMahan's (2009, 249–52) "altruistic killer").

them even in your ignorance. Given the unjustifiability of what you were doing, you should be glad to have obtained your evidence, so you can prioritize stopping it at once.

These considerations support an understanding of the DDE according to which it is harder to justify acts which, as a matter of objective fact, benefit some at the expense of others than it is to justify acts that simply benefit some rather than others. Some authors have considered such objective readings of the DDE, but have tended to dismiss them as absurd. There are at least three reasons for this—two of which are not compelling, but one of which is very important.

The first reason some authors dismiss objective readings is that they simply conflate the criteria of objective wrongness and reasons in which these readings are framed with aretaic criteria and criteria of subjective wrongness. Alastair Norcross (1999, 116–17) considers a case such as:

> *Misleading Evidence*. On Friday you have excellent evidence that to your right is one drowning swimmer and to your left are five. So you omit to save the one swimmer and instead pull from the water what you take to be five drowning swimmers. On Saturday there is a party for you at which the Pope gives you a Seal of Approval for gallant action permitted by the DDE. But, just after the Pope awards you the Seal, a hospital official informs you: "You know, it was funny. Those five things you pulled from the water turned out to be convincing inanimate robots. But all was not lost: the organs of the swimmer who drowned turned out to be a unique match for five patients who were dying from organ failure, so we used them to save the five." Hearing this, the Pope angrily snatches the Seal from you and remarks: "Ah ha! So the benefits your act generated actually depended on the harm it allowed to the one! You should have saved the one instead! I denounce you, and will have no further part in this celebration!"

Norcross's suggestion is that objective readings of the DDE would entail that the Pope's obviously inappropriate reaction would be appropriate. But this is false. By way of comparison, consider a case in which you have excellent evidence that you are helping many when you are in fact harming many. Bad criticisms of actual-consequence formulations of consequentialism similarly charge that the theory must be wrong because it entails that in such a case you should have practically reasoned your way to doing otherwise and that others are justified in blaming you. But as has been pointed out repeatedly, these falsehoods are in no way entailed by consequentialism so understood. Such versions of consequentialism give us a criterion of *objective rightness*, which we are to try to approximate by using our evidence to determine the likelihood that our acts will be supported by the considerations that the theory identifies as reasons.

As we lack omniscience, these subjective assessments of wrongness and reasons in light of our evidence are all we can use in practical reasoning. Whether an agent behaved rightly in the objective but not the subjective sense is completely irrelevant to the quality of her reasoning and aretaic assessments of her blameworthiness or estimability, which are tied to rewards and punishments like denunciations and snatchings of Seals. Objective readings of the DDE, like actual-consequence formulations of consequentialism, are theories of objective moral reasons. It is understood that they will of course be implemented as theories of subjective reasons through reasonable expectations of which courses of action will benefit some at the expense of others. Since in Misleading Evidence you did what your evidence told you the objective DDE permitted, it will entail that you blamelessly did right in the subjective sense.

A second reason some authors dismiss objective readings is that they consider versions of the Doctrine that are dubious in ways that are independent of its being read objectively. Frey's (1975, 279–83) early criticism of an objective reading focused on all the implausible features associated with the Catholic tradition: that the DDE prohibits using harmful effects on ourselves as well as others, that its force cannot be attenuated by the wrongdoing or culpability of those who are harmed, and that masturbation is intrinsically immoral. More importantly, I think there can be interactions between the plausibility of objective readings and the ways we assume the problem of closeness should be solved.

It is, I believe, distinctly plausible to say of my Misleading Evidence case that because the benefits of not saving the one swimmer came at her expense, the facts of your case did not justify your failure to save her, although you reasonably thought they did. But suppose your evidence told you that the only way to save five innocents was to throw a grenade into a room containing both a sixth innocent and a weapon that would otherwise be used to kill the five. In fact, the weapon is not in the room and the grenade saves the five by killing the one and consequently demoralizing those who would otherwise have killed them. Here it does not seem to me plausible to say that because the benefit to the five came at the one's expense, the facts of your case did not justify chucking the grenade, though you reasonably thought they did. This, however, is because it seems to me that in saving the five by chucking the grenade, you would have saved the five at the one's expense *even if your evidence had been accurate*.[27]

A third reason why authors have dismissed objective readings of the DDE is, however, that they have accurately perceived the absurdity of the most simple-minded understanding of what these readings are saying. On this understanding, objective readings are telling us to make sure that our acts do not have beneficial effects that depend upon their harmful effects. That is, they are saying that it is particularly difficult to justify acts that benefit some at the expense of others because the fact that an act benefits some in virtue of causing or allowing harmful effects on others *is itself a weighty reason not to perform the act*. But as Bennett (1995, 199) points out, "There is no evident reason why morality should forbid the [benefit-on-harm dependence] structure" itself.

That is actually an understatement: as Norcross (2008, 76) observes, the view really amounts to the ridiculous claim that it is worse to cause or allow a harmful effect if it does any good. Consider:

The Other Three. You can save two drowning swimmers to your left or one drowning swimmer to your right. You also know that there are three totally different people in hospital dying from organ failure, for whom the organs of the one swimmer are a unique match. So if (but only if) you save the two rather than the one, the one will drown and his organs will be used to save the other three as well.

[27] Importantly, Norcross (1999, 116–17) and McMahan (2009, 368–9) consider cases that are more like this than Misleading Evidence. Of course, counting both kinds of grenade chucking as benefiting some at the expense of others makes trouble for the distinction some try to draw between tactical and terror bombing. But as I mentioned in fn. 6, I think harmful tactical bombing *is* difficult to justify, and is typically more easily justified by our evidence than terror bombing only because it is much less likely to inflict harm and much more likely to produce benefits.

On the simple-minded understanding, objective readings of the DDE hold that it is wrong to allow something to push a fat man into the path of a trolley in order to save five because

(i) the fact that your act will save some in virtue of allowing a lethal effect on others is a reason against performing it, that
(ii) decisively outweighs the fact that your act will save five.

But this means that although it is permissible to allow one swimmer to drown rather than two where this is all that happens, it is wrong, given the presence of the other three, to allow the one swimmer to drown rather than the two because your act would (i) save some in virtue of allowing a lethal effect on others, which (ii) counts decisively against your act despite the fact that it would save five. It would, in this context, be permissible to let the one drown so long as it does not do any good. This, of course, is preposterous.

11.4 HOW "THE ENDS DO NOT JUSTIFY THE MEANS"

To avoid the dirty hands objection, a theory of deontological constraints must, I have been saying, explain not only why it is wrong to push a fat man in front of a trolley to save five but wrong to let something push him as well. Subjective readings of the DDE say it is wrong to let the fat man be pushed because you would be letting him die with the intention of his dying. But this faces the dirty heart objection that it amounts to a narcissistic obsession with the purity of your intentions. Simple-minded objective readings of the DDE say it is wrong to let the fat man be pushed because your conduct would benefit some individuals in virtue of allowing harm to others. But this seems to entail, preposterously, that it can be permissible to allow harmful effects as long as they do no good.

So why *is* it wrong to let the fat man be pushed into the path of the trolley in order to save five, assuming that it is? It would seem that *the reason it is wrong*, the feature that *makes it wrong*, and your *reason not to do it* is actually the same reason you should not let the fat man be pushed into the trolley's path when the five are not present: if you allow him to be pushed he will die, and you can easily prevent this. That, I believe, is the most natural and unforced explanation. It appeals to nothing more than his welfare, and the fact that you could easily promote it.

But how could the fact that the fat man will die if you don't save him make it wrong to fail to save him, when the fact that one swimmer will die if you don't save him *doesn't* make it wrong to fail to save *him*? In both cases our reason to save the one seems opposed by the same reason not to save him: that by doing so we can save five others. So how can that reason be a sufficient justification for failing to save the one swimmer but not for failing to save the fat man? The assumption common to both subjective and simple-minded objective readings of the DDE is that this explanation must cite some *additional reason*, beyond just the effect on the fat man's welfare, *against* pushing him into the trolley's path. But this, I fear, is exactly where theories of deontological constraints go wrong.

It is often suggested that in addition to failing to save the fat man, you disrespect him, or somehow offend against his autonomy in a way you do not disrespect or impose upon the swimmer.[28] But because your effect on how things are for the fat man is identical to your effect on how things are for the swimmer, it is very difficult to believe that there is any such difference in treatment that could be more important than your reasons to save the five. As we saw with the Impurity Police, it is preposterous that the fat man or his representatives should care significantly about any secret "disrespect" allegedly embodied in your intentions towards him in acting. In the same way, it seems absurd to think that you have offended against the autonomy of the fat man in any way in which you have not offended against the autonomy of the swimmer. In both cases you allow effects that interfere in identical ways with their ability to live their own lives as they see fit.

Of course, if it is *for independent reasons* wrong to allow the fat man to die but not wrong to allow the swimmer to die, then the former or his representatives might justifiably complain of the deprivation of a good where the latter might not, for you owed the good in the one case but not the other. They might even put their complaint in terms of your failing to "respect" the fat man by giving him what you owed him. But, obviously, it is then the antecedently greater difficulty of justifying the failure to save the fat man that explains the disrespect, not the disrespect that explains the greater difficulty of justifying your failure to save him.

[28] See Quinn (1989a, 1989b).

How, then, could it be wrong to fail to save the fat man but not the drowning swimmer if there are no moral reasons that count against doing the former that do not equally count against doing the latter? Consider the Machiavellian dictum that 'The ends justify the means.' In context, the idea seems to be that if the only way to promote a beneficial end is to use harmful means, the benefit counts as a perfectly good reason to use them. It is, however, plausible to understand deontological constraints on harming as, fundamentally, a rejection of this idea. It is not that we have some special kind of reason *against* using harmful means. It is rather that when an act will benefit some only by having harmful effects on others, the benefits simply *do not count* in the same way as reasons *to* perform it.

Jonathan Dancy has emphasized the distinction between considerations that favour and oppose acts on the one hand and considerations that strengthen or weaken the force of other reasons on the other. For instance, while the fact that you promised to go to the store is a reason to go to the store, the fact that the promise was given under duress weakens this reason without itself counting against going. If you had no other reasons for or against going to the store, the fact that the promise to go was given under duress would move your situation in the direction of both options being permissible; it would not by itself tend to make going to the store something you positively should not do.[29]

This is, I believe, exactly how we implicitly think about the fact that our conduct will have beneficial effects on some in virtue of its harmful effects on others: it weakens the status of the benefits as reasons to engage in the conduct without counting positively against the conduct. Consider the following pair of cases suggested by McMahan (1994):

Accident Victim 1. An accident victim will die if you do not help him, but your risk of contracting a fatal disease is so great that it is supererogatory to help.

Accident Victim 2. The same as before, except you know that if you fail to help and the victim dies, his organs will be used to save five people in hospital.

[29] Dancy (2004, especially 38–52). See also Kagan's (1988) distinction between features that "additively" make an independent positive or negative contribution to an act's deontic status and features that "multiplicatively" affect the contributions of other features.

The fact that in the second case the five will benefit at the expense of the accident victim is not a new reason *to help* him. The presence of the five in the second case cannot make it obligatory to help the accident victim where it was supererogatory to do so before. But neither, deontologists should say, does the presence of five in the second case (do much to) add to the case *against* saving the accident victim. What justifies (that is, permits) your not saving the accident victim in the second case is, as in the first, simply the risk to yourself, not the benefit to the five in the hospital. The fact that not saving the accident victim will benefit the five simply does not count (very strongly) in favour of not saving him, because these benefits would come at the expense of the victim.

Along these lines, I suggest that we read the DDE objectively, but as a claim about the *weakening* or undermining of reasons *to* cause or allow harm rather than some new set of reasons *against* doing so.

The Preferred Objective Reading of the DDE: All else held equal,[30] the fact that an act or omission will result in benefits for some individuals at the expense of other individuals weakens the extent to which those benefits count in favour of the act or omission.[31]

The preferred objective reading is not saying that there is anything particularly objectionable about the fact that an act has benefits in virtue of having harmful effects; it is not saying we should try to make sure that our acts do not have this property. According to this reading, there is absolutely nothing wrong with saving two swimmers rather than one

[30] The all-else-held-equal clause is required because there are plausible factors that attenuate or undermine the applicability of the DDE (these will be weakeners of weakeners). Such factors include consent to be harmed, culpability, and a duty to bear the harm. Candidate attenuators must not be *ad hoc*, but it is, I believe, directly plausible that these considerations undermine the DDE's applicability.

[31] I speak of the *weakening* of the strength of the reasons constituted by the benefits rather than the *total disabling* of their status reasons in order to allow a non-absolutist formulation of the DDE. Since the DDE is intended to apply to sub-lethal upshots, an absolutist formulation would be intolerable (it would entail that it is wrong to simply push someone down to save someone else's life). Of course, the weakening must be substantial if it is to explain why it is wrong to push or allow the pushing of the fat man, and non-absolutists need a plausible, non-arbitrary account of its degree. For serious harms, I think we should start with the vague idea that the weakening is "massive" or, given the weakening, only "an absolutely ridiculously crazily greater" amount of good could justify the benefit.

when three others will benefit from the one's death by receiving his organs. Here the omission to save the one is fully justified by the fact that it is the only way to save the two. As in a case when the other three are not present, your act simply benefits the two swimmers rather than the one; it does not benefit the two at the one's expense. All the preferred objective reading insists is that because they come at the one swimmer's expense, the benefits to the five in hospital do not count very strongly in favour of allowing the swimmer to drown.

According to the preferred objective reading, allowing the fat man to be pushed in front of the trolley to save the five is wrong, not because its beneficial effects depend upon its harmful effects, but—just like allowing him to be pushed when it does no one any good—because it allows him to die when you could easily prevent it. This is the *reason not to do it* and the fact that *makes it wrong*. The fact that the five would benefit at the fat man's expense merely explains why the benefits to them do not count very strongly in favour of allowing his pushing and why they fail to make this omission permissible. This is why allowing the fat man to be pushed differs from saving five swimmers rather than one. Since saving the five swimmers simply benefits them instead of the one and does not benefit them at the one's expense, there is nothing to prevent the benefits to the five from counting fully in favour of not saving the one and rendering that option permissible.

The assumption that relationships of dependence between beneficial and harmful effects would have to be reasons against acting or wrong-makers is, I believe, the primary reason why people have thought it absurd that they could matter morally. By way of analogy, suppose you made a promise to go to the store that would, under ordinary circumstances, oblige you to go there rather than stay where you are and provide costless help to someone who needs it. But suppose that the promise was made under duress, and that absent these reasons to go it is wrong not to stay and provide help. It is, in particular, wrong to go to the store rather than stay. If someone asked you why it was wrong to go to the store rather than stay, you would not say, "In going to the store, I would be doing what I promised to do under duress." That would be crazy! How could there be anything morally objectionable about doing what you promised under duress to do? Surely the view is absurd on its face! This is not, however, because the fact that a promise was given under duress is morally irrelevant. It is simply because, in answer to the question of why an act is objectionable, you have cited a consideration that explains why something that could have made it unobjectionable did not, rather than what made it objectionable in the first place.[32]

I believe that the preferred objective reading of the DDE gives us a plausible way to rescue a theory of deontological constraints from the charge of narcissistic obsession with the cleanliness of our hands, without running into the charge of narcissistic obsession with the purity of our hearts. Whether you push the fat man yourself or allow him to be pushed, the benefits to the five come at his expense, so according to the preferred objective reading, they fail to count very strongly as reasons to do *or* allow the pushing. Consequently, pushing the fat man yourself or letting something else do the dirty work are wrong for the same reason: they result in harm to the fat man (which you could avoid at trivial cost to yourself). This is a powerfully important fact about the effect of your conduct on the welfare of another individual rather than a dubiously relevant fact about the beauty or ugliness of your internal states.

The preferred objective reading's explanation of why this consideration makes it wrong to let the fat man be pushed, and how this case differs from that of the six swimmers, also looks appropriately focused on how your conduct affects others. On this view, the benefits your conduct would generate for some lose their force as reasons to engage in that conduct, not because of anything about your internal states, but because these benefits would be generated at the expense of others.

[32] The fact that the relationship between our conduct's harmful and beneficial effects works as a weakener of the status of its benefits as reasons rather than a reason not to act should help clarify (if it really needs clarification) how this relationship matters morally without a similar relationship between natural events' harmful and beneficial effects mattering axiologically. Clearly, we should not care whether an avalanche kills one rather than five because it simply lands on the one rather than the five or because the one shields the five. Neither outcome is better nor worse than the other (Tadros 2011, 219–20). Of course, this kind of axiological evaluation seems irrelevant to proposals about deontological constraints (as opposed to strange consequentialist views according to which we should save the fat man because it is intrinsically bad for the world to contain instances of some benefiting from the misfortunes of others). But the preferred objective reading helps clarify why this is so: it is a basic, agent-relative deontic fact that the beneficial effects of your conduct do not count (as strongly) as reasons to engage in it if they come at the expense of others. Apart from the harms to those others, no reason against engaging in that conduct—like the alleged intrinsic badness of some benefiting from the misery of others—is needed to explain why it is wrong.

It is important to be clear about how the preferred objective reading makes the reason-giving force of benefits to some dependent upon facts about others. For each individual, we seem to have standing reasons to treat her in general kinds of ways, the strength of which are affected only by facts about her (such as the extent to which the treatment will benefit her) and facts about us (such as whether we are specially related to her as a family member or friend). The preferred objective reading does not maintain that the strength of these general reasons to benefit someone depends on facts about others; what it does is place constraints on the extent to which these general reasons to benefit her can support specific courses of beneficial action or omission. The fact that an act or omission will benefit an individual at the expense of others does not affect the strength of our reasons to pursue the general end of helping her; it merely makes it difficult for this end to justify its pursuit by means of that act or omission. This, I believe, is a directly plausible constraint on our reasons to pursue morally important goals in particular ways, which does not make the moral importance of the general goal of helping an individual implausibly dependent upon facts about others.[33]

Because of its appropriate focus on how our conduct affects others, the preferred objective reading enables proponents of agent-centered deontological constraints to give an adequately non-narcissistic justification of their refusal in certain situations to bring about the most good. Suppose a deontologist saves the fat man from falling onto the track, and the parents of the five object: "How could you do this to our children, who otherwise would have been fine? Were their five lives less valuable than the fat man's one?" "Certainly not," the deontologist should reply, "it is just that, under the circumstances, my reasons against allowing the fat man to be pushed outweighed my reasons to bring about the most valuable outcome by means of allowing him to be pushed." The parents inquire: "And what reasons were those?"

At this point, most subjective readings of the DDE would tell our deontologist to make the unacceptably narcissistic reply: "My reasons not to allow the fat man to die with the intention of doing so." But the preferred objective reading allows her to say simply: "My reasons to save the fat man, at trivial cost to myself." "But how," the parents ask, "could those outweigh your reasons to save our *five* children, simply by minding

[33] I am grateful to a referee at Oxford University Press for raising this issue.

your own business?" "Because," our deontologist can explain, "under the circumstances the benefits to your children were not very good reasons against intervening." The parents demand to know "Why Not?!?" Our deontologist could not look them in the eye and say: "Because to save your children I would have had to allow harm with the intention of doing so." But she need feel no embarrassment in saying what the preferred objective reading entails: "Because the benefits to your children would have come at the fat man's expense; the only reason non-intervention would have saved your children is that it would have ensured the fat man's smashing."

11.5 CONCLUSION

It is quite plausible in the abstract that it is harder to justify benefitting some individuals at the expense of others than it is to justify simply benefitting some individuals instead of others. According to the preferred objective reading of the DDE, this is true because, if doing or allowing something will generate a benefit for some individuals at the expense of others, the benefit loses (much of) its status as a good reason to do or allow that thing. On reflection, I think that there is *a great deal* of direct plausibility to this idea. In fact, it seems to me no less clearly true or self-evident than the Principle of Beneficence that underlies impartial consequentialism, according to which there is intrinsic moral reason to promote the welfare of others. Unless said of situations in which harm to some is the only means of preventing *radically* greater harms to others, the Machiavellian dictum that beneficial ends are perfectly good reasons to use harmful means looks implausible on its face.

While the apparent plausibility of a general ethical principle does not guarantee its truth, we should accept the principle if its plausibility survives critical scrutiny and harmonious integration with other plausible ideas. The best arguments against deontological constraints acknowledge the initial plausibility of something like the DDA or DDE, and attempt to show that critical scrutiny undermines this plausibility. In this essay I have considered what I take to be one of the most powerful such arguments—that, on reflection, deontological constraints seem to embody an implausibly narcissistic obsession with the purity of our hands and hearts. I have argued that this argument fails to undermine an objective reading of the DDE according to which the benefits of our conduct do not count as strongly in its favour when they come at someone's expense. If I am right about this, and the plausibility of this view survives other forms of scrutiny,[34] I believe we should accept it.

Ethical justifications must give out somewhere, and ethical theories need to take certain principles as fundamentally axiomatic or constitutive of the deepest theoretical justifications there are. Like most impartial consequentialists, I think that the Principle of Beneficence is an axiom of this kind. Scrutiny will reveal, I believe, that nothing could be more clearly or basically true than the idea that there are moral reasons to promote the welfare of others. But the idea that the benefits of our conduct for some do not count as strongly in its favour when they come at the expense of others looks to me to be equally axiomatic. Like Beneficence, it does not seem to need any further justification. I think that deontological theories will be on far firmer ground if they acknowledge this idea, rather than anything about respect, rights, or autonomy, as the fundamental axiom underlying constraints on harming.[35]

REFERENCES

Bennett, Jonathan (1981). Morality and Consequences. *The Tanner Lectures on Human Values*. Vol. 2. Ed. Sterling McMurrin. Salt Lake City, UT: University of Utah Press, 46–116.

—— (1995). *The Act Itself*. Oxford: Clarendon Press.

Costa, Michael (1986). The Trolley Problem Revisited. *Southern Journal of Philosophy* 24: 437–49.

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

Dowe, Philip (1995). Causality and Conserved Quantities: A Reply to Salmon. *Philosophy of Science* 62: 321–33.

Ehring, Douglas (1997). *Causation and Persistence*. Oxford: Oxford University Press.

Fitzpatrick, William (2006). The Intend/Foresee Distinction and the Problem of "Closeness." *Philosophical Studies* 128: 585–617.

[34] Such as whether it can be combined with a convincing detailed solution to the problem of closeness.

[35] This essay has benefited from conversations with and feedback from too many people to remember or name. I am particularly indebted to John Ku for many valuable discussions. Special thanks are also due to Tom Beauchamp, Stephen Darwall, Allan Gibbard, Alastair Norcross, Drew Schroeder, Steve Sverdlik, Ralph Wedgwood, and an anonymous referee at Oxford University Press. I am also grateful for helpful feedback from audiences at the Third Annual Rocky Mountain Ethics Congress, the 2009 Eastern Division Meeting of the American Philosophical Association, and the Third Annual Arizona Workshop in Normative Ethics, as well as the students in my fall 2011 seminar on Deontological Ethics at the University of Alberta.

Foot, Philippa (1967). The Problem of Abortion and the Doctrine of Double Effect. *Oxford Review* 5: 5–15. Reprinted in P. Foot, *Virtues and Vices*. Oxford: Oxford University Press, 19–31.

—— (1984). Killing and Letting Die. In J. L. Garfield and P. Hennessey (eds.). *Abortion: Moral and Legal Perspectives*. Amherst, MA: University of Massachusetts Press, 177–85.

Frankena, William (1963). *Ethics*. Englewood Cliffs, NJ: Prentice Hall.

Frey, R. G. (1975). Some Aspects to the Doctrine of Double Effect. *Canadian Journal of Philosophy* 5: 259–83.

Hall, Ned (2004). Two Concepts of Causation. In J. Collins, N. Hall, and L. Paul, Causation and Counterfactuals. Cambridge, MA: MIT Press, 225–76.

Hanser, Matthew (1999). Killing, Letting Die, and Preventing People from Being Saved. *Utilitas* 11: 277–95.

Kagan, Shelly (1988. The Additive Fallacy. *Ethics* 99: 5–31.

Kamm, Frances (2007). *Intricate Ethics: Rights, Responsibilities, and Permissible Harm*. New York: Oxford University Press.

Kavka, Gregory (1983). The Toxin Puzzle. *Analysis* 43: 33–6.

McMahan, Jeff (1993). Killing, Letting Die, and Withdrawing Aid. *Ethics* 103: 250–79.

—— (1994). Revising the Doctrine of Double Effect. *Journal of Applied Philosophy* 11: 201–12.

—— (2009). Intention, Permissibility, Terrorism, and War. *Philosophical Perspectives* 23: 345–72.

Nielsen, Kai (1972). Against Moral Conservativism. *Ethics* 82: 219–31.

Norcross, Alastair (1999). Intending and Foreseeing Death: Potholes on the Road to Hell. *Southwest Philosophy Review* 15: 115–23.

—— (2008). Off her Trolley? Francis Kamm and the Metaphysics of Morality. *Utilitas* 20: 65–80.

Quinn, Warren (1989a). Actions, Intentions, and Consequences: The Doctrine of Doing and Allowing. *Philosophical Review* 98: 287–312.

—— (1989b). Actions, Intentions, and Consequences: The Doctrine of Double Effect. *Philosophy and Public Affairs* 18: 334–51.

Ross, W. D. (1930). *The Right and the Good*. Oxford: Clarendon Press.

Scanlon, T. M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, MA: Harvard University Press.

Shaw, Joseph (2006). Intentions and Trolleys. *Philosophical Quarterly* 56: 63–83.

Singer, Peter (1972). Famine, Affluence, and Morality. *Philosophy and Public Affairs* 1: 229–43

—— (1974). Sidgwick and Reflective Equilibrium. *Monist* 58: 490–517.

Tadros, Victor (2011). Wronging and Motivation. In R. A. Duff and S. Green (eds.), *Philosophical Foundations of Criminal Law*. Oxford: Oxford University Press, 207–27.

Taurek, John (1977). Should the Numbers Count? *Philosophy and Public Affairs* 6: 293–316.

Thomson, Judith (1976). Killing, Letting Die, and the Trolley Problem. *The Monist* 59: 204–17.

—— (1991). Self Defense. *Philosophy and Public Affairs* 20: 283–310.

—— (2008). Turning the Trolley. *Philosophy and Public Affairs* 36: 359–74.

Unger, Peter (1996). *Living High and Letting Die: Our Illusion of Innocence*. New York: Oxford University Press.

Wedgwood, Ralph (2011). Scanlon on Double Effect. *Philosophy and Phenomenological Research* 83: 464–72.

12

Some Intellectual Aspects of the Cardinal Virtues

PAUL BLOOMFIELD

12.1 INTRODUCTION

Nothing is more common in the burgeoning field of virtue epistemology than to find papers written about the moral aspects of the intellectual virtues. Hence, we have excellent work on intellectual courage by Robert Roberts and Jay Wood, as well by Jason Baehr; with regard to the role of temperance in intellectual virtue, Christopher Hookway has written insightfully on epistemic *akrasia*, while Heather Battaly has done the same for epistemic self-indulgence.[1] And as for epistemology and justice, Miranda Fricker has written a fine treatise on epistemic injustice, or the sort of injustice perpetrated when people's testimony is discredited for arbitrary reasons, such as race.[2] There is also a small literature on judicial virtue, or the moral virtues of a good judge in a legal system.[3] And while virtue epistemologists have been strangely silent about wisdom *per se*, John Kvanvig has at least approached the topic with his work on the difference between knowledge and understanding.[4] So, many virtue epistemologists

[1] Roberts and Wood, *Intellectual Virtues* (Oxford: Clarendon Press) 2007; Baehr, *The Inquiring Mind* (New York: Oxford University Press) 2011; Hookway, "Epistemic Akrasia and Epistemic Virtue," in *Virtue Epistemology*, A. Fairweather and L. Zagzebski (eds.) (New York: Oxford University Press) 2001; Battaly, "Epistemic Self-Indulgence," in *Virtue and Vice*, H. Battaly (ed.) (Oxford: Wiley–Blackwell) 2010.

[2] *Epistemic Injustice* (Oxford: Oxford University Press) 2007.

[3] For a good introduction, see Lawrence Solum's "Virtue Jurisprudence," in *Metaphilosophy* 34, nos. 1–2, 2003: 178–213. See also David Luban's "Justice Holmes and Judicial Virtue," Terry Pinkard's "Judicial Virtue and Democratic Politics," and Judith Shklar's "Justice without Virtue," all in *Virtue, Nomos*, *Vol.* 34 (New York: New York University Press) 1992.

[4] *The Value of Knowledge and the Pursuit of Understanding* (New York: Cambridge University Press) 2007.

have explored how the cardinal moral virtues can bear on issues in epistemology. Less has been said on the other side: namely, on the intellectual aspects of the cardinal moral virtues, which are courage (*andreia*), temperance (*sophrosune*), justice (*dikiaosune*), and wisdom (*phronesis*).

Wisdom is perhaps the exception, since *phronesis* is variously translated as practical intelligence, practical rationality, or practical wisdom, so its intellectual standing stands out. Indeed, two impressive book-length treatises on this topic have been recently produced by Julia Annas and Daniel Russell.[5] We will return to them and *phronesis* toward the end of the essay. The plan is first to explore the intellectual aspects of courage (*andreia*), temperance (*sophrosune*), and justice (*dikiaosune*). Then, with these results on the table, we can draw some conclusions about their relation to *phronesis*, the intellectual aspects of moral virtue, and the degree to which there is a unity among the virtues despite how they may differ. The issue at bottom is that, on one side, common sense (*eudoxia*) tells us that the virtues are, at least for the most part, unrelated. We should not be surprised by the idea that, for example, people may be courageous without being just. On the other side, there are theoretical considerations which have led philosophers to think that there is in fact only one virtue, *phronesis* or practical rationality (or practical wisdom), and that courage is the exercise of practical wisdom in dangerous circumstances, temperance is practical wisdom in tempting circumstances, and so on. This is the much contested "unity of virtues" thesis: that possessing practical wisdom is necessary and sufficient for possessing all the virtues. The solution which falls out of looking into the intellectual structure of the moral virtues is that *phronesis* is necessary for all the virtues but is not sufficient. Aside from the experience required to master individual virtues, there are intellectual aspects of each of the virtues which may not be derived by *phronesis* alone. So, in Section 12.2 a model of how the intellectual aspects of the cardinal virtues are related to each other is given, and in Section 12.3 there is a discussion of the intellectual aspects of each of the virtues taken on its own. Finally, in Section 12.4 the unity of virtues thesis is discussed.

One might wonder if much of what is to come is really necessary.[6] Is there really any debate about how "intellectual" the cardinal virtues are?

[5] Daniel Russell, *Practical Intelligence and the Virtues* (Oxford: Oxford University Press) 2009; Julia Annas, *Intelligent Virtue* (New York: Oxford University Press) 2011.

[6] I thank Drew Schroeder for wondering just this.

In fact, there is. Common sense often caricatures the virtues in a non-cognitive way: courage is about handling fearful feelings, temperance is about handling temptation, justice is about handling greedy desires, and wisdom is about avoiding foolishness. Undoubtedly, this is a start, but it is far from a proper account of the virtues. More to the point perhaps is the fact that theorists of the virtues, both moral and intellectual, debate over the nature of virtue itself. Many take an approach based on Greek eudaimonism, agreeing that the virtues are intimately related to (if not necessary and/or sufficient for) living a happy flourishing life. Many who take this approach follow Socrates and the Stoics in thinking that the virtues are skills, much like the prosaic skill of being a cobbler or an automobile mechanic.[7] Others follow Aristotle and acknowledge that the virtues are very similar to skills, but differ from them in some important ways (for example, that skills concern the making of products (*poesis*) while virtues concern practices (*praxis*).)[8] Alasdair MacIntyre has argued for a relativistic virtue theory in which what counts as a "virtue" is determined by cultural convention.[9] There are also sentimentalist theories of virtue, such as that of Michael Slote, wherein what counts as a virtue is determined by what is fit to be admired.[10] Consequentialist virtue theories, such as those of Thomas Hurka and Julia Driver, take virtues to be those character traits that lead to the best consequences, where these are defined independently of virtue itself.[11] And there are also "pluralist" views of virtue, such as those developed by Christine Swanton and Robert Adams, in which there is no univocal account of the nature of a virtue; virtues are simply character traits that respond well to those items in the "field" of the virtue.[12] So, the primary goal

[7] See for example, Julia Annas, "Virtue as a Skill," *International Journal of Philosophical Studies* 3 (2), 1995: 227–43, and *Intelligent Virtue* (New York: Oxford University Press) 2011; Matthew Stichter, "Ethical Expertise," in *Ethical Theory and Moral Practice* 10, 2007: 183–94, and "The Skill Model of Virtue" in *Philosophy in the Contemporary World* 14: 39–49 (2007); and my own "Virtue Epistemology and the Epistemology of Virtue," *Philosophy and Phenomenological Research* 60 (1), 2000: 23–43, and *Moral Reality* (New York: Oxford University Press), 2001, chapter 2.

[8] See, for example, Rosalind Hursthouse, *On Virtue Ethics* (Oxford: Oxford University Press) 1999, and Linda Zagzebski, *Virtues of the Mind* (Cambridge: Cambridge University Press) 1996.

[9] Alasdair MacIntyre, *After Virtue* (London: Duckworth Press) 1985.

[10] Michael Slote, *Moral Sentimentalism* (Oxford: Oxford University Press) 2010.

[11] Thomas Hurka, *Virtue, Vice, and Value* (Oxford: Oxford University Press) 2000; Julia Driver, *Uneasy Virtue* (Cambridge: Cambridge University Press) 2001.

[12] Christine Swanton, *Virtue Ethics: A Pluralistic View* (New York: Oxford University Press) 2005; Robert Adams*, A Theory of Virtue* (New York: Oxford University Press) 2006.

here is to vindicate the idea that each virtue has its own intellectual structure, a *logos*, which must be accounted for by any adequate theory of the virtues.

A terminological note to start. Very often when virtues are discussed by virtue ethicists, they end up looking far different than how they appear to common sense.[13] Still, there is some reason to think that virtue, especially wisdom, is something we should not expect to be wholly within the ken of the folk: the folk revere the wise because the wise have insight which the folk lack. And we should expect the same to be true of the other virtues.[14] We can imagine a "folk theories" of courage, temperance, justice, and wisdom, which could be formulated by psychologists gathering data from random subjects on what they think of these virtues. But we would not want this to be taken as the truth about the virtues any more than we take folk theories of physics give the truth about physics. To keep the truth about the virtues distinct from what the folk think about them, henceforth, "f-courage," "f-temperance," and so on, will be used to refer to what the folk think about the virtues and the words unadorned will be used to refer to how the virtues are modeled within a putatively true and complete theory of the virtues.[15] So, for example, f-temperance involves prudishness, tee-totaling abstinence, and perhaps even the nightmare of

[13] Thanks to Tim Elder for making me see the need for this terminological point.

[14] Of course, this opens the question about whether or not virtue ethics is elitist. Here I argue against Philippa Foot, Rosalind Hursthouse, Julia Annas, and perhaps Aquinas as well. In discussing the claim that it is "quite wrong to suggest that wisdom cannot be a virtue because virtue must be within the reach of anyone who really wants it," Foot responds: "Wisdom, insofar it consists of knowledge which anyone can gain in the course of an ordinary life, is available to anyone who really wants it. As Aquinas put it, it belongs 'to a power under the direction of the will'." My response to such a thought is that while I am perfectly, enthusiastically egalitarian about moral fallibility and about the fact that each of us can become more wise than we currently are, I understand this as being comparable to saying that we can each become better at mathematics than we currently are. Becoming truly wise, or becoming an exceptionally talented mathematician, is a feat which can only be accomplished with prodigious natural talent; all the best intentions and effort in the world are not sufficient. For the charge of elitism, see Julia Driver, *Uneasy Virtue* (Cambridge: Cambridge University Press) 2001. For Foot, *Virtues and Vices* (Berkeley: University of California Press) 1978, p. 6; Rosalind Hursthouse, "Practical Wisdom: A Mundane Account," *Proceedings of the Aristotelian Society* 106 (1), 2006: 285–309; Annas 2011. For the relevant notion of talent and achievement, see D. Lubinski and C. P. Benbow, "Study of Mathematically Precocious Youth After 35 Years: Uncovering Antecedents for the Development of Math-Science Expertise," *Perspectives on Psychological Science* 4 (1), 2006: 316–45.

[15] This notion of modeling is discussed in Russell (2009, p. 362), and will arise again at the end of this essay.

Carrie Nation wielding a hatchet, while temperance carries no such baggage. The thought here is that, with regard to practical wisdom, the folk have a rough and ready conception of it, along with being able to grasp simple articulations of it (consider the homilies in *Poor Richard's Almanack*), much like in folk physics. When matters become subtle, however, the judgment of the folk turns unreliable, while the judgment of the truly wise does not.

12.2 A PRACTICAL MODEL FOR VIRTUE

Let us begin with a rough and ready way of understanding the relation of courage and temperance to wisdom by saying that courage is the management of *phronesis* as applied to emotions and attitudes involving fear and confidence, while temperance is *phronesis* engaged with desire and revulsion.[16] More specifically, courage directs our behavior when we are faced with things from which we naturally shrink, like the prospect of pain or death, while temperance directs us when we are faced with what our passions and appetites crave, such as pleasure and satisfaction. Thus, both courage and temperance involve knowing how to "stand firm" in the face of what is repellant and attractive (respectively). How does justice fit into the picture? The answer is that it is justice at play when we must consider others, often in relation to ourselves. It involves knowing what people deserve, in terms of resources, rewards, and punishments, given both who they are and what they have done. And very often, this involves self-knowledge about who we are what we have done.[17] While we should not take the Aristotelian idiom of "virtue as a mean" too literally, we can caricature courage and temperance by noting that the former is a mean between cowardice and recklessness, while the latter is a mean between gluttonous over-indulgence and tee-totaling abstinence. If so, then, contra Aristotle, justice can be understood as a mean

[16] This idea is discussed in T. H. Irwin, "The parts of the soul and the cardinal virtues." in *Platon: Politeia*, O. Hoeffe (ed.) (Berlin: Akademie Verlag) 1997. I think I found even more helpful Irwin's "Do Virtues Conflict? Aquinas' Answer," in *Virtue Ethics, Old and New*, S. Gardiner (ed.) (Ithaca: Cornell University Press) 2005. I am indebted to Lionel Shapiro for discussion in which he pointed out the differences between revulsion and cowardice that I had not appreciated.

[17] For more on why this is so, see my "Justice as a Self-Regarding Virtue," *Philosophy and Phenomenological Research* 82 (1), 2011: 46–64.

between arrogance and servility.[18] On such an account, it is arrogance and not gluttony that is the cause of *pleonexia* or unjust greed.

As a model for understanding how the intellectual aspects of these virtues are related, consider how being a master carpenter, plumber, and electrician are related. The first works with wood, the second with water, and the last with electricity. But all require practical intelligence in design and construction, and all require the sorts of intellectual virtues that are involved in being methodical, careful, precise (to the degree required by the activity), creative, patient, and insightful.[19] Now, if we imagine that a master carpenter has leaky faucet or has to install a new electrical outlet in her own house, we should expect that this would not be too challenging: if one is capable of framing a house, one is most likely capable of fixing a leaky faucet. And if a master plumber wanted to build a tree house for a child, we should not expect this to be beyond her ken. The degree of practical intelligence required to be a master at any of these skills will be sufficient for successfully undertaking relatively undemanding construction jobs which are not within their specialty. Nevertheless, being a master at any of these involves large amounts of special knowledge. Carpenters must understand how different materials and different designs can support different amounts of weight, plumbers have to understand fluid dynamics, and electricians must have at least a rudimentary understanding of the physics of electricity. And being an expert at one of these certainly does not entail being an expert at any of the others (though, of course, there are those rare few who can do them all). Of course, a plumber will continue to be methodical, careful, precise, and so on, when trying to fix some electrical wiring and will not be sloppy about it or rushed or careless. If there are peculiar "vices" of construction, then our experts will avoid them even outside their areas of expertise. But being

[18] This argues against Bernard Williams' interpretation of Aristotle's idea that virtue as a mean. His thought is that one cannot be "too just" and so justice is an exception to the idea of virtue as a mean. But aside from justice as a mean between servility and arrogance, it can also be seen judicially in meting out just deserts, as a mean between being merciful and being draconian. Bernard Williams, "Justice as a Virtue," in *Essays on Aristotle's Ethics*, A. O. Rorty (ed.) (Berkeley: University of California Press) 1980; David Sachs, "Notes on Unfairly Gaining More: Pleonexia," in *Virtues and Reasons: Philippa Foot and Moral Theory*, R. Hursthouse, G. Lawrence, and W. Quinn (eds.) (Oxford: Oxford University Press) 1998.

[19] And we can even see further analogies to the virtues when we think of "character building."

an expert in one only guarantees a certain competence in the others—a far cry from any sort of full unity. The competence here does not imply expertise—only an ability to complete those tasks which can be "figured out" without any of the special knowledge that marks experts. (This is similar to a theoretical example, regarding *theoria* as opposed to *phronesis*, in which being an expert in metaphysics guarantees a certain competence in ethics or logic.)

12.3 THE CARDINAL VIRTUES

So, with this idea in mind, we may now turn to courage, temperance, and justice *per se*.

Starting with courage, it seems that clearly immoral people, from criminals to pirates to tyrants, can regularly behave in ways that at least appear to be courageous, and this seems to imply that one can be courageous and unjust. It is doubtful that anyone ever accused Stalin of being a coward. Common sense, as well as Bernard Williams, tells us that one person can both meet the "standard of the bright eye and gleaming coat" while still being "red in tooth and claw."[20] And even if everyone does want to draw a distinction between f-courage and foolish recklessness, still perhaps it seems possible that a person can be f-courageous and not too smart or at least without coming close to having the sagacity of a judge on the bench.[21] And there seems to be little superficial reason to think that f-courageous soldiers are necessarily going to be f-temperate as well.

[20] The first phrase here is Williams'; the second is from Tennyson. See *Ethics and the Limits of Philosophy* (Cambridge: Harvard University Press) 1985, p. 46, and *In Memoriam A. H. H.*, 1850, canto 56.

[21] See fn. 7. This is the point at which appears the fictional character of Forrest Gump, who seems to have an abnormally low IQ and yet is quite capable of brave-hearted courage. Nevertheless, given the intellectual aspect of courage discussed later, Gump's character strains credulity, as being someone not smart enough to discern the difference between recklessness and courage, and yet is always lucky enough to get away with his recklessness; foolish, lucky recklessness can make for f-courage, but does not true courage make, however superficially similar they may appear. At the very least, what Gump seems to lack is *euboulia* (good deliberation), which will be discussed more later. Two examples of this are given by Hursthouse 2006: (i) the person who sees a child drifting down a river and thinks to run down river ahead of the child before diving in, and (ii) the solider, upon finding out that the enemy is in his camp, thinks to grab his helmet and shield and not just run out with his sword. Gump runs into a burning forest to recklessly rescue comrades when only pure luck keeps him from burning with them. Again, f-courage may call this "courageous," but recklessness plus luck do not equal true courage.

If any of this is true, then there might be good reason to think that f-courage is not a moral virtue, given that the virtues are not supposed be compatible with unethical or immoral or vicious behavior: if f-courage really is a "virtue" which can be possessed in isolation from all the other moral virtues, and can, like money, be used in the pursuit of the basest ends imaginable, then perhaps it does not deserve the honorific of being "a virtue" at all. Of course, this is not how the ancient Greeks thought of courage. Plato, for example, tells us that courage is always noble (*kalon*).[22] And there is no doubt that Aristotle also thought of it as a virtue (*Nicomachean Ethics*, Book III, chapters 7–10). So, how can we reconcile contemporary thoughts about the possibility of immoral yet f-courageous people with the ancient understanding of the inherent nobility of courage?

Well, one quick way is to chalk up the issue to the idea that f-courage is understood wholly in relation to how people handle fear. And of course, managing fear is central to understanding courage and managing any emotion can require cognitive or even skillful consideration: we need, at some level, to train ourselves to handle fear well, to control, say, our breathing and how calm we are in the face of real danger so that we may reliably (and not just luckily) do the right thing. The psychological training of fearful response, brought about through practice and reflection upon handling what typically elicits fear, is the sort of thing that military boot camp is for and that real-life experience of danger deepens. The mistake of f-courage is in thinking that the content of courage is exhausted by handling one's fearful reactions correctly. For a start, consider how part of "fearing correctly" involves knowledge of the difference between what is truly fearful and what merely appears to be. As soon as one brings in an appearance/reality distinction, we have left the emotional world behind and have entered purely cognitive territory. Recognizing a phobia as such is not too difficult, but distinguishing real danger from what only seems dangerous may be far less easy. And we may remember that the Stoics thought that nothing is truly fearful at all: according to them, when we know what is truly of value and what is not, we will see that nothing is truly fearful—even torture on the

[22] *Laches*, 192C, and *Protagoras*, 349E. I found helpful Kenneth Seeskin, "Courage and Knowledge: A Perspective on the Socratic Paradox," in *Southern Journal of Philosophy* 14 (4), 1976: 511–21.

rack, or death. Of course, Aristotle disagreed with the Stoics, dismissing those who accept such thoughts as being in the grip of a theory. But remember that Aristotle himself was in the grip of *ta endoxa* with regard to *eudaimonia*, or what we might call the "folk theory of happiness," which includes the idea that certain "external goods" are necessary for a flourishing life, and we can see his account of courage and what is fearful as being based on his theory of what is of value in the world, what intellectuals call "axiology."[23] So, the full story about courage, on any sophisticated account, requires more than the ability to manage one's fears: at the barest minimum, it will require both an ability to discern real from apparent danger and knowledge of what is of value in life.

So, courage requires an ability to manage fear, a conative achievement, but it also requires an intellectual understanding of what is worth taking risks for. Still, this sort of axiology, while a cognitive inquiry, is oriented around value in general and not knowledge *per se*. Nevertheless, there is another even more purely epistemic layer to courage that becomes apparent upon considering the deliberative activity involved in risk assessment. One form of recklessness involves not fearing what truly merits fear, while another is taking inappropriate risks for the sake of trivial ends. It seems plain, even to common sense once it is pointed out, that true courage requires knowing what is worth dying for, that any fool can die for a cause, but truly courageous people are not fools. At this point we may note that "discretion is the better part of valor."[24] Now, even if Falstaff did use this thought in excusing his own cowardice, we might still insist that courage requires knowing what to risk for the sake of what. "Discretion" here can refer to the ability to discern when to charge forward and when to retreat, as well as the self-knowledge involved in comprehending one's own talents and abilities and the reliable application of this knowledge in the face of danger. One must be able to envision and evaluate different possible scenarios in order to deliberate upon them. But even assuming that "discretion" names a distinct epistemic virtue, there is nothing about it *per se* which

[23] For more on Aristotle on *endoxa*, see Julia Annas, *The Morality of Happiness* (New York: Oxford University Press) 1994.

[24] Falstaff says, upon faking his death on the battlefield, "The better part of valor is discretion, in the which better part I have sav'd my life." *Henry IV*, part 1, act 5, scene 4. A less cowardly take on the thought is found in Bob Marley's song *The Heathen*, in which he sings "Rise and take your stance again/ 'Cause he who fight and run away/live to fight another day," from *Exodus*, 1977.

yields this substantial practical knowledge. One must have experience in battle, or in the courtroom, to be able to see when continuing the charge forward is in fact self-defeating. One must be able to comprehend which missions are possible and which are impossible. And even keener sight is needed in order to be able to spot a trap when all appears safe. It is, at least in part, the intellectual assessment of risk which separates the courageous from the reckless, and yet this idea seems missing from f-courage. Courage cannot be had without experience of risk itself, and one's courage is developed in part by the development of the intellectual skill of assessing risk. In purely quantitative terms the science of risk assessment is nowadays known as the "actuarial sciences," about which the author admits to knowing practically nothing, other than that equations from physics describing Brownian motion are sometimes used in making predictions of success and failure in taking risks.[25] More philosophically, we are in the area of formal decision-making and planning, as these are discussed by philosophers such as Michael Bratman and Allan Gibbard, with perhaps Baysianesque epicycles appended.[26]

This is not to suggest that the consummately courageous person must be an actuary, much less a Baysian. On the contrary, what is more apt is to say that these formalized procedures of assessing risk are attempts to model what is known by those who are courageous The intellectual structure of full courage, encompassing both knowledge of value and the assessment of risk is epistemically far richer and deeper than merely managing an emotion like fear. And it is only when one begins to appreciate the intellectual aspects of courage that one sees that money or material goods are *not* worth the risks of theft or piracy, and that power over others is not worth the risks of being a tyrant. Even if these people escape punishment, they inevitably live plotting to keep it at bay. Evil

[25] My thanks to Auralia Perrica for discussion on this matter.

[26] Bratman, *Intention, Plans, and Practical Reason*, (Cambridge: Harvard University Press) 1987; Gibbard, *Thinking How To Live* (Cambridge: Harvard University Press) 2008. For general discussion of risk, see N. Rescher, *Risk* (Lanham, MD: University Press of America) 1982; S. O. Hansson, "Risk," *The Stanford Encyclopedia of Philosophy* (autumn 2011 edition), Edward N. Zalta (ed.), <http://plato.stanford.edu/archives/fall2011/entries/risk/>; and P. Weirich, "Causal Decision Theory," *The Stanford Encyclopedia of Philosophy* (winter 2012 edition), Edward N. Zalta (ed.), forthcoming: <http://plato.stanford.edu/archives/win2012/entries/decision-causal/>.

causes are not worth fighting and possibly dying for, and it is reasonable to think that all truly courageous people will agree that this is true. As noted, only reckless fools are willing to die for causes that are not worth it. Disagreement here only arises over the axiological question of determining "what is worth what?" and not on the epistemic claim that courageous people know what is worth dying for: people who are truly courageous know what is worth dying for and the wrong causes are not. And so, there is some reason to think that people who fight for evil causes are not truly courageous, even if the folk say something different, and, more to the point, even if f-courageous people are capable of acting in many circumstances as truly courageous people would. Being truly courageous entails comprehending the reasons for being courageous, and these reasons must be good reasons: but evil causes cannot be good reasons for action. (Of course, there are bad reasons for fighting for evil causes.) People who are not truly courageous may appear to be, and, as noted, the judgments of the folk, and the medals for bravery they distribute, are not authoritative. While putting off the "unity of virtues" thesis for a bit longer, we may in any case conclude that if discretion really is the better part of valor, if one must accurately assess the risks and values involved before one can justify courageously putting one's life on the line, then intellectual perspicacity is essential to courage.

Temperance is similar to courage in that its intellectual depth superficially appears to be exhausted by learning how to "stand firm" in the face of something non-cognitive: for courage it is fear and for temperance it is pleasure and satisfaction.[27] To give a sense of the current state of philosophical literature on what the Greeks called *sophrosune*, the *Philosopher's Index* database gives 187 hits for papers with *akrasia* or "weakness of will" in the title and only twenty hits for *sophrosune*, "temperance," and "will power." Moral philosophers seem to be almost ten times more interested in incontinence than temperance. Now, the psychology of moral philosophers aside, this approach is backward. How can one expect to explain what happens when things go wrong without first clearly understanding what it is like when they are as they ought to be?[28] This is something like

[27] My thinking about temperance in these regards is much influenced by conversations with Scott LaBarge and his paper "Socrates and the Recognition of Experts," in *Apeiron* 30 (4), 1997: 51–62.

[28] It seems to me that if *akrasia* is a failure of willpower, then again understanding willpower should precede the investigation of *akrasia*. For one philosophical article on willpower, R.

trying to fix a broken engine without understanding how engines are supposed to work. There is more to temperance than willpower, since if one is truly temperate, willpower is not needed at all: those who are temperate are not even tempted by what they ought truly not to be tempted by, and as such they do not even need willpower, since for them there is no temptation to resist.[29] Those who are temperate, *sōphrones*, are immune to improper temptation and, *a fortiori*, *akrasia*. Aristotle famously contrasted temperance to both continence and incontinence (*Nicomachean Ethics*, book VII). Willpower is only needed by those who are continent, which is better than being incontinent but not as good as being well-tempered. *Sōphrones* comprehend which pleasures are innocent and which are harmful, which salutary and which detrimental; they discriminate between pleasures. Of course, this does not inhibit their ability to be passionate about their chosen pleasures: while f-temperance may see itself as champion of abstinence and the enemy of passion, temperance is truly only the enemy of illicit passion and over-indulgence. And so, perhaps unsurprisingly, axiology again becomes relevant to the virtue. Consider: most probably none of us are even tempted by heroin, despite knowing the pleasure we could experience by taking it. This shows that we have at least some understanding of what is truly valuable in life.

The more purely epistemic aspects of temperance involve how questions of epistemology affect our moral decision-making. And I think these can be addressed by thinking about how we make judgments about who to trust and who to distrust; the epistemological jargon casts the debate in terms of "testimony." Fricker introduces the idea of an "anti-prejudicial virtue" by saying: "Let us call it (what else?) the virtue of *testimonial*

Roberts, "Will Power and the Virtues," in *Philosophical Review* 93 (2), 1984: 227–47. Recent data from psychology on willpower is relevant. Some have argued that willpower should be understood, literally, at least partly as a form of strength, given that glucose levels in the blood seem to affect one's ability to "stand firm" in the face of temptation. The hypothesis is contentious. See R. F. Baumeister, K. D. Vohs, and D. M. Tice, "The Strength Model of Self-Control," in *Current Directions in Psychological Science* 16 (6), 2007: 351–5, and M. T. Gaillot *et al*., "Self-Control Relies on Glucose as a limited Energy Source: Willpower is More Than a Metaphor," *Journal of Personality and Social Psychology*, 92 (2), 2007: 325–36. For a critique of this work, see V. Job, C. Dweck, and G. Walton, "Ego Depletion; Is It All in Your Head?: Implicit Theories About Willpower Affect Self-Regulation," in *Psychological Science*, published online 28 September 2010 at http://pss.sagepub.com/content/early/2010/09/28/0956797610384745.

[29] Continuing with the strength metaphor, "resisting" illicit temptation for people who are truly temperate is like a weight-lifter lifting a 1-pound weight; it does not even exercise one's capacities.

justice" (2007, p. 92, italics in original). In answer to the "what else?," an alternative, more classical way of thinking about how to combat prejudice is not by pitting it against justice and explaining its occurrence as a lack thereof, but rather by way of temperance and its absence.[30] Of course, people are done an injustice if their testimony is discounted for arbitrary reasons, but Plato's thought (in *Charmidies*) is that the relevant virtue which is supposed to manage issues of this sort is temperance, since prejudice is due to the undue influence of non-cognitive, appetitive elements of the mind, such as desire, insecurity, passion, and so on. The intellectual aspects of temperance appear in how and to what degree we let our appetites and desires cloud our judgment of who to trust and listen to; presumably, the more wanton we are, the more prone to prejudice we will be. In any case, it is common nowadays to note that people tend to listen to the "experts" with whom they already happen to agree: in America, Republicans tend to watch Fox News, and Democrats tend to listen to MSNBC. It is less common to see the influence of prejudice in these tendencies as failures of temperance, and yet, rightly understood, they are.[31]

There has been some discussion of these issues by contemporary epistemologists cast in terms of how we can discriminate between trustworthy and untrustworthy experts. Alvin Goldman tells us to (i) look at the how experts support their views, (ii) see if there is a consensus of experts on a question, (iii) look at "meta-expertise" or how credentials are earned, (iv) look for bias or a conflict of interest on the part of experts, and (v) look for success in the past.[32] But these helpful suggestions about

[30] The issue between Fricker and myself recapitulates somewhat a discussion between Hursthouse and Christine Swanton. Hursthouse points out that unjustly cheating fellow soldiers out of their rations out of a "pursuit of pleasure" is ultimately a failure of temperance not justice. Swanton, perhaps rightly, points out that not all unjust acts are the result of a lack of temperance; if injustice due to arrogance is a failure of temperance, it is not a normal failure. For the double self-deception involved in arrogance, see Robin Dillon, "Kant on Arrogance and Self-Respect," in *Setting the Moral Compass*, C. Calhoun (ed.) (New York: Oxford University Press) 2004. For Hursthouse, see *"A False Doctrine of the Mean,"* in Proceedings of the Aristotelian Society 81, 1980–81: 57–72, at 64. For Swanton, see *Virtue Ethics* (Oxford: Oxford University Press) 2003, at p. 21.

[31] Other surprising places to find a lack of temperance are in "rubbernecking" as people drive by car wrecks; when gossiping about others; or in watching melodramatic soap-operas or even "reality" TV shows. All these, in the end, amount to the same thing, though we should be careful to distinguish Schadenfreude from Nemesis. I thank Julia Annas for pointing out this final distinction to me.

[32] "Experts: Which Ones Should You Trust?" *Philosophy and Phenomenological Research* 63 (1), 2001, 85–110.

discerning between experts will not be of much help when we are trying to judge the trustworthiness of people in everyday life. Nor will they help when the problems in making good judgments about who to trust come from within us as opposed to being found in those who may pose as neutral experts but are not.[33] Some of the thorniest epistemological problems involving questions of who to trust are the result of our own biases of which we might be wholly unaware. And these are not problems regarding judging expertise *per se*, but are more general issues of our own propensities to trust writ large. Nothing is more human than for us to hear what we want to hear, to ask for advice and opinions from people who think like we do, and to think that those who disagree with us with regard to matters we care about must be wrong. True temperance, at the epistemic level, involves knowing if and how our personal psychological needs and passions are influencing our decision-making. Here, we approach a topic that has been discussed by virtue epistemologists; namely, the intellectual virtue of being "open-minded," but these discussions have not drawn any explicit connections to temperance.[34]

In the early modern period, Bishop Butler was one person who pursued these themes, though again not explicitly in terms of temperance, but he did employ the idea of "temperament" which is obviously germane. In his sermon "Upon Self-Deceit" he notes that there is nothing more common than for us to use our reflective abilities to justify to ourselves the way we favour the special objects of our desires or passions:

> But whereas, in common and ordinary wickedness, this unreasonableness, this partiality and selfishness, relates only, or chiefly, to the temper and passions, in the characters we are now considering, it reaches to the understanding, and influences the very judgment. And, besides that general want of distrust and

[33] Aristotle's remarks on *sunesis* (comprehension) (*NE* 1143a15), as an aspect of *phronesis* (practical wisdom), are related to the present point about temperance, for this is a virtue which concerns the evaluation of testimony. In particular, the way in which we can spot another's incorrect testimony or reading of a situation requires *sunesis* in a way which may be purely cognitive, in which there are no biases or prejudices causing the error which is spotted. For discussion, see Hursthouse (2006).

[34] See papers by Wayne Riggs and Jason Baehrs. While Riggs' discussion of open-mindedness does involve self-knowledge, Baehrs takes self-knowledge to be a precondition of open-mindedness and not part of it. Baehrs does use the word "tempted" once in relation to a tendency to fall back into a default cognitive position. Neither explicitly relate open-mindedness to temperance. Riggs, "Open-Mindedness," in Battaly 2010 Baehrs, chapter 8 2011.

> diffidence concerning our own character, there are, you see, two things, which may thus prejudice and darken the understanding itself: that overfondness for ourselves, which we are all so liable to; and also being under the power of any particular passion or appetite, or engaged in any particular pursuit.[35]

As Butler notes, as good as we may be at defending ourselves, we are perhaps even better at criticizing those with whom we disagree or do not like. The topic on which Butler writes, self-deceit, is one which obviously calls for self-knowledge and knowledge in general, and when we see that it is our passions and appetites that cause our self-deceit, then the epistemology of temperance comes to the fore.

We have to go back to Plato's *Charmides,* I think, to find a discussion of self-knowledge and recognition of experts that explicitly places the epistemological issues within the bailiwick of the moral virtue of temperance. Indeed, at one point in the dialogue, temperance is described as the virtue captured by the quotation at the entrance to the Oracle of Delphi, *Know Thyself*, the thought being that "Know Thyself" and "Be Temperate" are "the same thing:" temperate is what you are when you know yourself (164e). On this hypothesis, to say "know yourself" can be a piece of practical advice of special use when a person is about to give in to improper temptation—similar to saying "Nothing in excess."[36] A few paragraphs later it becomes clear that knowledge of the self has a less moral, purely epistemological character. There it is clear, as it is put, that the *sophrosune*, the temperate man (the gender being in the original):

> . . . alone will know himself and be able to examine what he in fact knows and what he does not, and he will be capable of looking at other people in the same way to see what any of them knows and thinks he knows, if he *does* know; and what, on the other hand, he thinks he knows but does not. (167a, italics in translation)

Of course, *Charmides* is all about the problematic nature of this sort of second-order "knowledge of knowledge," though (hopefully) we can

[35] *The works of Joseph Butler, D.C.L*, W.E. Gladstone (ed.) (Oxford: Clarendon Press) 1897, pp. 146–7.

[36] The translation is partly Donald Watt's. I substitute "Be Temperate" for "Be Self-Controlled" in the quoted phrase, but the identity asserted is clearly in the text. *Early Socratic Dialogues* (London: Penguin Books) 1987.

side-step these issues, taking it for granted that we have, in fact, gained some knowledge about knowledge over the centuries, and that we now call this knowledge "epistemology." And it barely needs to be mentioned that self-knowledge is a species of knowledge. Now, if *akrasia* exists, which is practically indubitable, we can conclude that temperance is not exhausted by knowledge, and thus we may side-step Socrates infamous claim that all vice is due to ignorance. And it seems equally indubitable that Aristotle's account of temperance as involving only bodily pleasures and appetites is also radically incomplete (*NE*, book 3, chapter 10). A complete account of temperance will require the investigation of matters that are both cognitive and non-cognitive.

Plato's conception of temperance as operating in the realm of the purely cognitive and Aristotle's conception of it as operating in the realm of the non-cognitive can be reconciled by the straightforward idea that, while temperance may be exercised in purely cognitive ways, such as in choosing who to trust, or in the knowing of when others know, very often (but not always) what makes these cognitive processes go awry, when they do, is something non-cognitive. Temperance is the virtue of not letting the exercises of one's judgment be clouded by emotions, desires, appetites, passions, and so on. One might say that while our non-cognitive lives are central to the human condition, they should only be taken as evidence or data to be considered during deliberation, which is best done (as Butler pointed out) in an emotionally cool state of mind. One might think that the "purely cognitive" sorts of knowledge involved in temperance might be purely descriptive or non-normative, but as Critias suggests (*Charmides*, 174b), axiology must once again be included, for there is perhaps nowhere our judgments are more likely to be non-cognitively influenced than in judgments about what is good and bad: humans naturally want to see what we want, or what we desire, as good.

A last word about temperance before moving onto justice. Epistemologists are most often concerned with knowledge in the broadest sense. And the judging of when people (in general) know and when they do not is often the focus. But, as noted above, there is the special case of knowledge which is peculiar to each of us considered as individuals: obviously, self-knowledge. There are aspects of this which are general that all self-knowledge shares, while other aspects are unique from person to person. Perhaps there are general lessons we can learn about how to strengthen our willpower so that it may face all but the very greatest of pleasures without a glimmer of temptation (see fn. 21 and 22). But from the personal point of view, at a basic level, the ways in which our own peculiar appetites and passions may affect, or infect, our judgments are something about which we must be autodidactic—no one can know any of us well enough to teach us this.[37] In the solitude of our own individual consciences, we must learn to discern the effects of the judgments we make about ourselves, our self-conceptions, upon the judgments we make about the rest of the world, and it will always be here that temperance will be hardest to master.

Now, perhaps the careful reader will have noticed that we have slipped into the cognitive idiom of "judgment" talk. This was no accident. There is a rich tradition of theorizing about what judgments are, what judging is; the *Oxford English Dictionary* lists twelve basic (though somewhat overlapping) usages. Descartes, Locke, and Hume, and I am sure many others, have had their own theories of judgment. Here, judgments may be considered evaluations or assessments of what is being judged, where an evaluation or assessment is the application of a standard to a case: one paradigm of judgment is in the application of rules practiced with the sensitivity to spot, grasp, or comprehend exceptions. The broadest sense of this idea of rule application is found in bare predication, in concept application, such that we judge an item to instantiate a property or think or assert that "the world is thus and not so." (We will return to this later.) We judge things as being of a kind, and we also judge them as being good or bad examples of their kind, better or worse, right or wrong, correct or incorrect, functioning or malfunctioning, or, colloquially, "up to snuff." And yet, despite all this implicit normativity, at least some such judgments have truth values: we can judge truly that the engine is broken, that it is not working "as it is supposed to work," or that it is not doing "what it ought to do," that something is "wrong with it." So too, truly, a heart may malfunction and not do what it ought to do. And so too, we may judge that some agent is acting properly, is not behaving as he or she ought to behave,

[37] I take this point from John Stuart Mill's *On Liberty*, chapter 4. Drew Schroeder insightfully points out that, for example, one's spouse might very well be able to spot the ways in which one's emotions are affecting one's judgments better than one can for oneself. The point is well taken, and is backed up by various writings on self-deception. The point about being an autodidact is perhaps best construed in terms of the privileged yet still all-too-fallible access which we have to our own minds.

that there is something wrong with that behavior. The logical structure of these judgments is identical, and this leads one to conclude that, at least *qua* judgment, they merit being treating in the same way. As such, a unified theory of judgment may require bridging the "truth/value" gap, may force a reconception or even dissolution of that gap. And interestingly, this opens up the possibility that normativity, as it is found in morality and epistemology respectively, should likewise not be conceived of as wholly non-factual evaluation. If the moral virtues and the intellectual virtues are so intertwined that they cannot be understood independent of each other, then a unification of metaethics and metaepistemology may be possible.[38] Obviously, these issues range far beyond the current topic. For now, it suffices to note that the content of judgments are normative, insofar as they mark when something "hits a standard," "makes the grade," or "counts as an *x*."[39] Notice that there is an analytic reason to think that "judgment" as a noun is predicated upon "judging" as a verb, that the verb is primary: without judging, there are no judgments. Judging is performative action, and so can itself be done well or poorly and is therefore itself up for normative assessment. We may judge our judgments, and we have just seen how this may involve temperance. In any case, judgments are normative, and judging is normative too.

In a purely epistemic mode, justice is the virtue of making good judgments.[40] The primary uses of "judgment," and "judge" (as both a noun and a verb) in the *Oxford English Dictionary* are the legal senses of these terms. So, it seems that etymologically we may say that judging is done by judges who make judgments, and that justice is the virtue of judges and therefore of judging. Notice that this is different from the typical discussion of the virtues of judges or jurisprudential virtue mentioned at the outset. While the primary sense of "judgment" in the *Oxford English Dictionary* refers to legal matters, the primary sense of "justice" pertains

[38] As Gary Watson says: "[O]ne cannot apprehend the Form of Justice without apprehending the Form of Human Well-Being. But to apprehend the Form of something is to know its essence. To know the essence of human well-being is to understand what it involves and what it is worth in itself—and hence in all contexts." From "Virtues in Excess," *Philosophical Studies* 46, 1984: 57–74, at 60.

[39] While I disagree with John Mackie about the nature of value, I agree with him that the question of whether a discourse is truth-apt is determined by the nature of the standards that are at play in that discourse. See Mackie (1977) on "standards of evaluation," p. 25–7.

[40] For more on this take on justice, see my "Justice as a Self-Regarding Virtue" 2011.

to morals.[41] Thus, the intellectual virtues of the judicial judge might depart from a strict moral sense of "justice." So, we are not surprised to learn that the issue of "judicial restraint" is a "hard case" for jurisprudential virtue.[42] There is some reason to think that judges on the bench ought to always err on the side of being conservative and of respecting *stare decisis*. But judging as cognitive phenomenon is certainly not all judicial, and there is no reason to think that all judging should be similarly conservative: sometimes it is the bold judgment which is most likely to be just and true.

This is not to say that any sort of judging can do without some version of "the rule of law." It is invidious, it is epistemically unjust, to apply a standard to one case and a different standard to a relevantly similar case. At one level, this can be seen as enshrined as a principle of supervenience, in which if all Xs are judged to be p in virtue of their being q, then any x which is q must be judged to be p, on pain of contradiction. (Any exceptions are to be justified based on differences in the exceptional cases which warrant them being treated differentially.) This may be one area in which the law of non-contradiction must be considered as sacrosanct. At another level, the very idea of a concept can be seen as a rule, where actually having the concept *cat* is what allows us to judge any cats that we come across as "cats." If someone sometimes judges dogs to be "cats," then we say that that person does not possess the concept of *cat* at all. The very concept of a *concept* has built into it the idea that a concept will only be applied correctly, it will only make sense if it is applied consistently, if it succeeds in picking out all the things in the world to which it applies and picks out no other things. The excellence to be found in epistemic judgment is captured by the idea that like cases be judged alike, and differences in judgment must be due to differences in the cases being judged. This can be seen as the combined exercise of the sort of comprehension (*sunesis*) involved in medical diagnosis taken together with the sort of discretion at play when

[41] From the *Oxford English Dictionary*: "(1303)...A public officer appointed to administer the law; one who has authority to hear and try causes in a court of justice.. "; "(*c*.1225)...To try, or pronounce sentence upon (a person) in a court of justice; to sit in judgment upon."; "justice, n. (1137–54). The quality of being (morally) just or righteous; the principle of just dealing; the exhibition of this quality or principle in action; just conduct; integrity, rectitude."

[42] See Luban, 1992.

one knows what counts as a "relevant difference." This is the epistemology of the virtue of justice.

If this seems novel, it bears noting that it is not terribly original, though the context in which these issues typically arise is not within moral philosophy but rather on the line between political theory and the law. In an early paper of Rawls, taken from a 1957 APA symposium paper called "Justice as Fairness," on the very first page he says that "justice is the elimination of arbitrary difference." And on the next page, while laying out the ur-version of his "first principle of justice," he writes:

> One can view this principle as containing the principle that similar cases be judged similarly, or if distinctions are made in the handling of cases, there must be some relevant difference between them (a principle which follows from the concept of a judgment of any kind). (p. 654)[43]

Nor is Rawls the only other philosopher to place treating like cases alike at the center of all cognitive judgment.[44] And thus, we may conclude, the epistemic virtues of being just in one's moral judgments will be the same as those involved in making non-moral judgments. So, with regard to the virtues of good judgment, we may find a perfect unity between moral and epistemic or intellectual virtue.

Let me briefly conclude the discussion of courage, temperance, and justice as follows. If justice is really so central to all judging, note that it may sometimes require the boldness of courage, and it certainly requires the "even-mindedness" or balance of the tranquil and temperate disposition, and most certainly requires sagacity and insight. The virtue of courage will require both justice's discretion of cool-headed judgment, especially under duress, as we understand having "grace under fire," as well as temperance's not letting passion or anger recklessly "get

[43] "Symposium: Justice as Fairness," *Journal of Philosophy* 54 (22), 1957: 653–62.

[44] See, for example, Isaiah Berlin, "Equality," *Proceedings of the Aristotelian Society* 56, 1955–56: 301–26; Richard Wasserstrom mentions the point in reference to rationality in "Rights, Human Rights and Racial Discrimination," *Journal of Philosophy* 61, 1964: 634–5; J. B. Schneewind (without reference) quotes Clarke on the issue, noting in a footnote that Cumberland also comments on it. See *Proceedings and Addresses of the American Philosophical Association* 70 (2), 1996: 25–41. Of course, there is much discussion of the centrality of treating like cases alike in jurisprudence. See, as a place to begin, H. L. A. Hart, *The Concept of Law* (Oxford: Clarendon Press) 1961, chapter 8.

the better of you"; the truly courageous person can "smell a trap," and there is certainly a sort of wisdom involved in this. Temperance, as a virtue, requires willpower and being able to endure stress and "stand firm," and these kinds of strength and endurance are often thought of as being the hallmark of courage. Temperance also requires good judgment and discrimination among the innocent and harmful pleasures in the world. And when all the interconnections between thinking justly, courageously, and temperately begin to look obvious, the unity of virtues thesis starts to look promising.

And this brings us, finally, to *phronesis*. One "mundane" way to conceive of it can be seen as having already been supported by our discussion of the other cardinal virtues. Rosalind Hursthouse (2006) gives an Aristotelian account of *phronesis* where we see its content as being constituted by the intellectual virtues of *euboulia* (good deliberation), (*eu*)*sunesis* ((good) comprehension), and *gnome* (correct discernment), and that possessing these inevitably leads to *eupraxia*, or good practice. Then one might think that, barring the stuff of tragedies, *eudaimonia*, or a good life, inevitably follows. Russell (2009) adopts a similar account of *phronesis* but expands its constituents to include *nous,* or intelligence, and cleverness as well. Perhaps cleverness requires creativity, but if it does not then it seems natural to include creativity on the list. Thinking back on courage, temperance, and justice, it is not hard at all to spot the roles of deliberation, comprehension, discernment, intelligence, cleverness, and creativity in each of them. (We can only hope that some day virtue epistemologists will take up these items directly.)

But while there is some reason to think the Hursthouse/Russell account incorporates elements which are all necessary for a full account of *phronesis*, it nevertheless seems incomplete. This is because it includes no mention of knowledge of the difference between good and bad, no mention of axiology. Of course, one might say this is built in, given that we do not just want *boulia*, or deliberation, but *euboulia*, or good deliberation, and so the *phronimoi* must already have some sense of the difference between good and bad. Of course, this goes for *eusunesis* and *gnome* as well. This is not enough, however. Recall that courageous people need to know what is worth what in order to know how much risk to take, that the temperate needed to know which pleasures are salutary and which are harmful, and that the just make correct or accurate judgments about the qualities of what they judge. Any account of the abilities of the *phronimoi* will have to include the ability to know good from bad and right from wrong, and so axiology must be included in the account.[45]

12.4 THE UNITY OF VIRTUES THESIS

Given all this, what can we say about how the virtues are related to each other? What falls out of this by way of the "unity of virtues" thesis? Well, first remember the model mentioned at the outset, based on the relationship between being a carpenter, a plumber, and an electrician. True, there are some features of the underlying intellectual structures of these which are identical to each other: no master plumber is going to be a fool when it comes to using a hammer or a wire cutter. But this does not entail that all master plumbers are also master carpenters or electricians either. So too, with the virtues, we see that the truly courageous will have to have the intellectual resources that would prevent them from being fools with regard to judging which pleasures to indulge or who deserves to be rewarded or punished. This does not, however, entail that the truly courageous are automatically consummate experts in temperance and justice.

We can triangulate on this position on the unity of virtues thesis by comparing it to others. There are, of course, virtue theorists who think the thesis is simply false.[46] Three views similar to the present one

[45] No metaethical questions are being begged here, though the implication of axiology here does require not just an account of normative value, what is good and what is not, but also an account of value itself. *Contra* Blackburn, Rorty, and Dworkin, metaethics is a distinct part of axiology. The brief argument for this is as follows: consider the Rawlsian distinction between how "the rules of the game" are employed on the field and how the rules for changing "the rules of the game" are not employed on the field. If arguing over issues in normative ethics is governed by "the rules the game," then metaethics is the distinct office in which "the rules of the game" are themselves argued over. Should ethical argumentation over a debated normative issue proceed by rational argument or sentimental suasion or some combination of these? Answering questions such as this is the job of metaethics. For the Rawlsian point, see his "Two Concepts of Rules," *The Philosophical Review* 64, 1955: 3–32. For more on my take on these issues, see "Archimedeanism and Why Metaethics Matters," in *Oxford Studies in Metaethics, Vol.* 4, R. Shafer-Landau (ed.) (Oxford: Oxford University Press) 2009. For "metaethical minimalism," see Simon Blackburn, *Ruling Passions* (Oxford: Clarendon Press) 1998; Richard Rorty, *Consquences of Pragmatism* (Minneapolis, MN: University of Minnesota Press) 1982; and Ronald Dworkin, "Objectivity and Truth: You'd Better Believe It," *Philosophy and Public Affairs* 25 (2), 1996: 87–139. For the most sophisticated version of minimalism I have found, see Matthew Kramer, *Moral Realism as a Moral Doctrine* (Oxford: Wiley–Blackwell) 2009.

[46] See, for example, Swanton 2003, and Robert Adams, *A Theory of Virtue* (Oxford: Clarendon Press) 2006.

should serve to demonstrate its balance. What we end up with is a limited unity of virtues thesis, though not one identical to that defended by Neera Badhwar in an important paper on the topic.[47] She defends three theses: (i) that the existence of a virtue in one domain in one's life does not imply its existence in other domains, such that one might be courageous on the battlefield but a coward in love; (ii) the existence of virtue in one domain implies the absence of vice there and "ignorance in most other domains" (p. 308); and (iii) that within a domain, having one of the virtues implies having the rest. From the current perspective we can agree on (ii), but must take issue with (i) and (iii). With regard to (i), the idea that one can be courageous on the battlefield but not in love implies that the person does not truly understand the reasons behind being courageous. Part of being courageous involves thinking like someone who is courageous, not merely acting like such a person (the difference between f-courage and courage), and the idea that one can think like a courageous person in one situation but not in another suggests something like knowing that 2 + 2 = 4 sometimes but not all the time, or a plumber who can install a sink but not a toilet. When we fully appreciate the intellectual aspects of the virtues we find that having a virtue in one domain implies that one must have it to some significant degree in others. Lacking nerve in love tells against the courageousness of apparently brave action on the battlefield. This does not entail that the expert soldier must also be an expert lover, but rather that if one is a coward in love, or, even worse, has a phobia about harmless spiders, then one is not fully courageous. To take a different example, one may be able exercise temperance in one's professional life, so as not to let one's emotions dictate one's choices, and yet succumb in all sorts of way to influence of emotions when one's children are concerned.[48] The conclusion to draw from these examples is not that one can be courageous or temperate in one domain and not in others. Rather, it is to remember that, at the very least, courage involves the proper management of fear, and temperance involves the proper management of temptation and that the degree to which some fears or temptations "get the better of one" is the degree to which one fails to be courageous or temperate

[47] "The Limited Unity of Virtues," *Nous* 30 (3), 1996: 306–29.

[48] My thanks to Daniel Groll for conversation on this point. The example regarding temperance is from Drew Schroeder.

(respectively). With regard to (iii), Badhwar suggests that if one is kind to one's friends, then one will also be courageous, just, and temperate with them as well. Combining this with (i) implies that one can have the special knowledge of, say, knowing how to assess risks when friends are involved but not knowing how to do this when one is alone. If any conclusions follow from the discussion above, we have learned that what is intellectually required for courage, temperance, and justice is not to be had piecemeal, and that we cannot be expected to have it in some situations but not in others. Nevertheless, (iii) simply cannot be true if it is thought that merely being in a particular context, say being with friends, can automatically supply one with special knowledge of the virtues which one lacks in other contexts. So, of Badhwar's three aspects of her limited unity of virtues thesis, (i) and (iii) cannot be maintained.

While being consummately courageous implies being so everywhere, being so does not imply being consummately temperate or consummately just, though it does rule out being gluttonish or insensible, arrogant or servile. Perhaps however, if one really focuses in on the Aristotelian idea of the necessity of *phronesis* for all the virtues, one can derive a stronger unity than the one just described. If *phronesis* is necessary for all the virtues, then this seems to imply the possibility that being a *phronimos* is sufficient for having all the virtues. Russell 2009 calls the necessity and sufficiency of *phronesis* for all the virtues "hard virtue theory," and Annas (2011) defends something similar. The view that falls out of the discussion above is a more nuanced and limited view: being a *phronimos* is necessary for all the virtues, but it is not sufficient for them as well. It is sufficient for not having any of the moral vices and for *becoming* a master at all the virtues, but it is not sufficient for *being* a master of all the virtues.

In place of my model of carpenters, plumbers, and electricians, Annas suggests a different model; namely, the way that a pianist has all the skills for being a pianist and not one skill for fingering and another for tempo (2011, p. 87). So, in the same way that a pianist could not be skilled at fingering while lacking the skill of keeping proper tempo, a *phronimos* could not be courageous while lacking temperance or justice. In reply, one might suggest that a classical trained pianist might be rather inept at extemporizing jazz. But pursuing the matter in these terms would probably require the ability to individuate or count skills, which seems as hopeless as counting possible worlds. And, in fact, Annas does want to draw a distinction that allows us to see how one person can be more proficient at one virtue than the others. This is the difference between what she calls "the circumstances of a life" and the "living of a life" (2011, p. 93). The *circumstances of a life* are those features of our lives over which we have no control, the place in time and location in which we are born, our gender, height, nationality, culture, and so on, while the *living of a life* concerns what one does with one's life and the circumstances into which one is born. She rightly points out that the virtues are always exercised in the circumstances of our lives, and soldiers and care-givers lead very different kinds of life. She then concludes that "There is no such thing as being virtuous in a way which will be appropriate to all kinds of lives, or one ideal balance of virtues such as courage and patience that could be got right once and for all for everybody" (p. 95). She quotes Gary Watson in a footnote on the same page:

> The unity thesis implies that if one has a particular virtue one must have them all; it does not imply that if one has a particular virtue one's life will allow for the manifestation of all virtues equally. Which virtues will receive fuller expression will depend on fortune, cultural context, and one's moral personality. (2011 p. 65)

Undoubtedly, this is true. We explain how different people express the virtues by appeal to the differences between people and the circumstances in which they live their lives. But the problem with this thought is that it seems to be belied by those rare people, like Socrates, who do seem to be fully consummate in all the virtues. We can imagine a soldier coming home from war and being an excellent care-giver, full both of courage and sensitive patience. The question concerns what we are to say of *phronimoi* born in times of peace and never confronted with the sort of danger that requires advanced competence in risk assessment. These people will no doubt not act foolishly. We can imagine them being as courageous as their circumstances have allowed. What they are lacking, however, cannot be wholly chalked up to circumstance or context. What they are lacking are the specific and specialized forms of knowledge that come with *having learned lessons* through prolonged and intense exposure that cannot be learned by those whose exposure is more limited. To employ the Greek idiom, *phronesis* does not exhaust the individual *logos* for each virtue; each *logoi* also contains special knowledge and principles. There is more to the knowledge required for being a master of a particular virtue than what is required for being a master of deliberation, comprehension, discernment, and so on. For courage, particular knowledge of risk assessment is required; for temperance, one must know how one's needs, desires, and appetites can/do/may prejudice one's judgment; and for justice, knowing the difference between sympathy and mercy is necessary. There are epistemic aspects of each of the virtues that go beyond *phronesis per se*, and this shows that being practically wise, all by itself, is not sufficient for courage, temperance, and justice, that the wise person must actually go learn the special things known by courageous, temperate, and just people in order to instantiate these individual virtues. This keeps the virtues from being fully unified, even if there are certain people, like Socrates, who can equally and consummately manifest them all.

One might object to this by saying that while the soldier may not grasp justice as well as a judge, whatever special knowledge is involved in being a judge, it is merely the expression of *phronesis* cast in the "direction" of justice. This idea of "direction" comes from Russell, where he contrasts the idea of virtue as a "trajectory" to the idea of it being a "direction" (2009, pp. 342ff.). Trajectories are limited in shape and distance, while virtues are not limited like this. Learning a virtue is applying one's practical intelligence, one's *phronesis*, in a particular way or in a particular direction. Russell interestingly points out the difference between a theoretical understanding of the virtues, or how we model them, and how we attribute them to particular people (pp. 362ff.). So, Russell might try to save the unity of virtues thesis by replying that even if we do not attribute to people equal amounts of virtue, acknowledging that some of the virtues had by a person might not be as strongly manifested as others, when we consider the model of the virtues, and the way in which *phronesis* plays a necessary and unitary epistemological role in each of them, we may therefore conclude that the virtues are in fact unified at the level of theory, even if not in attributed fact.

As noted above, we can acknowledge the sense in which *phronesis* is sufficient for *becoming* fully virtuous even if it is not sufficient for actually *being* fully virtuous. The problem is that *phronesis* is essential to all practical endeavors that admit of excellence, not just the moral virtues, and it is hard to see what could stop Russell's line of argument from spreading in odd and perhaps even global directions. Let us assume that axiology is involved in *phronesis*, as discussed previously, and that we can draw a principled distinction between being an expert nurse and being an expert torturer, so we do not have to say that being kind and being sadistic are somehow unified. But if *phronesis* unifies the intellectual realm of character building or the construction of a flourishing life, and is also found to be sufficient for all excellent endeavors that are not proscribed by morality, from being a carpenter to a soldier to a nurse, or a parent or friend, we end up with a much broader unity than even Aristotle or perhaps even the Stoics thought. We do not want to say there is a unity to all worthwhile pursuits, but it is hard to see how to stop Russell's "model theoretic" argument from encompassing all this.

Instead, we may return to the prosaic relation of being a carpenter to being a plumber to being an electrician. What unifies them is practical intelligence: *phronesis*. What keeps them distinct is not merely that the rough materials of wood, water, and electricity differ, but the special forms of knowledge and technique that are developed within the various specialties. There are things which courageous soldiers know which just judges do not, and *vice versa*, so being courageous and just, or courage and justice *per se*, cannot amount to the same thing. They cannot be fully unified. Of course, this does not excuse us from striving to learn and manifest all the virtues we can, given the circumstances of life into which we are born. Socrates can still be our model of virtue manifest fully. We may not be as prodigiously talented as he was, but we can try just as hard as we can try. And this will make us as virtuous as each of us can be.

Index

accountability 157, 159 n. 8
achievement:
  instrumental value of 89, 96 n. 19
  intrinsic value of 83, 90
  action
  guidance of 268 nn. 19, 20
  symbolic or expressive aspects 156, 160, 161, 167, 168, 169, 248
  under uncertainty 296
actus reus 162
Adams, R. 289
agency 190–202
akrasia 296–304
Annas, J. 288, 311, 312
Anscombe, E. 61–2
answerability 157
Aquinas, T. 18, 21 n. 15, 290 n. 14
arbitrariness 48, 49
Aristotelian virtue ethics 19–24
Aristotle 55, 81 n. 6, 289, 294, 298, 299 n. 33
Arneson, R. 117 n. 37
attributability 156–7
Audi, R. 37 n. 2
autonomy 27, 28, 278

Badhwar, N. 309
Baehr, J. 287
Battaly, H. 287
beneficence 203 220–3, 284, 295
Bishop Butler 300, 301
Blackburn, S. 307 n. 46
blameworthiness 16, 17, 40, 47, 244, 250–3, 255–8, 275
Bratman, M. 189 n. 25
Broad, C. D. 23
Broome, J. 38
Buss, S. 227, 236 nn. 1, 2, 237 n. 3, 238 n. 5

categorical imperative 103
character 129, 156, 248, 313
cognitive attitudes 67
cognitive limitations 179–202
Cohen, G. A. 99, 112
collateral damage (see unintended side effects)
consequences 274, 275
  actual versus expected 182–4
consequentialism 43, 55, 145 n. 41, 205, 239 n. 10
  act 179
  agent-centered 132–6, 143
  agent-neutral 127, 205, 206 n. 2
  assignment of deontic status to actions 125, 133, 186
  equivalence with non-consequentialist theories 133
  maximizing 205, 206 n. 2, 238
  objections to 127–31, 181, 205 (see also deontological constraints)
  Sidgwickian 43, 54–5
  versus deontology 125, 137–42
consequentializing 123–53
  evaluator-neutral 134, 135
  evaluator-relative 133
  motivations for 123, 127
contractualism 44 n. 8
coordination/cooperation games 72–6

Crisp, R. 13–33, 139
Cullity, G. 207–11, 217 n. 8, 219
Cummiskey, D. 124

Dancy, J. 21 n. 16, 48, 279
Darwall, S. 138, 247
decision theory (see game theory)
“deep” self 157
defeaters 208, 250, 251 n. 19
deliberation 190–5, 200, 307
  joint 75
demandingness 28, 203–33
  source of 205–11
deontic constraints 98, 107 n. 18, 111, 112, 138, 216, 260, 277
deontic status 125, 137, 144, 185, 186, 268 n. 19
  reasons-based conception of 138
  value-based conception of 133, 134
  versus aretaic status 268 n. 18, 274
deontologizing 137–41
deontology 103, 138
  “dirty hands” objection 146, 263, 272
  objections to 102, 278
  versus consequentialism 137–42
Doctrine of the Double Effect (DDE) 114, 115, 118, 260
doctrine of the mean 22
doing/allowing distinction 44, 260, 263–6
Dreier, J. 125 n. 8
Driver, J. 289
Duff, A. 154, 160–6
duty 15, 27, 131 (*see also* moral requirements)
Dworkin, R. 307 n. 46

egoism, rational 141
Eshleman, A. 155
ethical theory 29, 30
Eudaimonism 48 n. 11, 289
excuses 40, 158, 176, 236
expertise 243, 248, 299, 300

Fischer, J. 155
fittingness 20 n. 16, 22, 23, 81
Fitzpatrick, W. 265 n. 13
folk concepts 290
Foot, P. 262 n. 6
Frey, R. 275
Fricker, M. 287
Funder, A. 212 n. 6

game theory 70–6
Gardener, J. 158
Gaus, J. 170 n. 22
Gettier cases 187
Goldman, A. 185 n. 19
good 38, 80, 82, 125

Hall, N. 266 n. 14
harm 106, 113
Hart, H. L. A. 154, 160, 161
Herman, B. 137 n. 34, 145 n. 41, 148–9, 203, 217–19
Heyd, D. 21, 27, 28, 29
Hill, T. 103 n. 10
Hookway, C. 287
Horgan, T. 15, 16
Hume, D. 65
Hurka, T. 80, 81, 96, 289
Hursthouse, R. 299 n. 30, 207
Husak, D. 154

infringement:
  property 99, 100,
  ranking wrongness of 102, 107–11
  rights 104, 105, 106
intentional objects 86–94

intention 159, 189, 267–72
interpersonal relations 44, 45, 55, 173, 239 n. 10, 283
intuitions 262

Johnson, R. 15 n. 4
judgment 303–6
justice 31, 306

Kagan, S. 124, 279 n. 29
Kamm, F. 210 n. 4, 262 n. 6
Kant, I. 103
- Formula of Humanity 110, 264
- Formula of Universal Law 103

Kantian moral theory 218
- conception of persons 110 n. 26
- imperfect virtue 30 n. 30
- and respect 130
- and value 149

Kavka's toxin puzzle 74 n. 18, 268, 269
knowledge 187, 245
- self-knowledge 298–302

Kolodny, N. 61–2, 74, 145 n. 41
Korsgaard, C. 149 n. 54
Kvanig, J. 287

laws 159, 168, 174
liability 158–66
libertarianism 99 n. 4, 117–22
Lidell, H. 20 n. 15
Locke, J. 117 n. 37
Louise, J. 133, 139–40

MacIntyre, A. 289
Mackie, J. 304 n. 40
Markovits, J. 15 n. 4
McKenna, M. 171 n. 23
McMahan, J. 264 n. 10, 268 n. 20, 276 n. 27
Mellema, G. 15 n. 4
mens rea 158, 163, 167 n. 19
metaethics 150, 151, 304, 308 n. 46
Mill, J. S. 40, 303 n. 38
Miller, D. 184 n. 18
Millian liberties 99, 121
Moore, G.E. 82, 90–1
moral development 217
moral epistemology 235–60
moral motivation 17, 18, 28, 175
moral requirability 185–9
moral requirements 39, 181, 237
- application to agents 196
- knowledge-based 245–50
- *pro tanto* versus all-things-considered 224

Morse, S. 154
Murphy, L. 211 n. 5

Nagel, T. 143 n. 39, 151
narcissism 261, 270
negligence 159 n. 9
Nielsen, K. 263
Norcross, A. 263, 274, 276 n. 27
norms 74
- legal 169–70, 172
- and metanorms 173

Nozick, R. 101, 106–12, 121 n. 41

obligations 63, 147, 175, 204, 209 (*see also* moral requirements)
omissions 261, 264 n. 11, 281
options 35, 38, 41, 182, 188 n. 23, 238
organic unities 82, 90–3, 95–6
Otsuka, M. 108, 110, 121 n. 41
ought 15, 37, 38, 151
- all-things-considered 37, 40
- evaluative versus normative 151, 152, 153
- implies can 245
- objective versus subjective 37, 38

Owens, D. 66

pain 93, 210 n. 4
Parfit, D. 74 n. 18, 138 n. 35
paternalism 103, 119, 122
permissions (permissibility) 103 n. 9, 115, 126, 222, 240, 262, 263, 268 n. 20, 280
 all-things-considered versus *pro tanto* 225
 rational 72
Pettit, P. 146
phronesis (*see* practical wisdom)
Plato 48 n. 11, 294, 299
pleasure 81, 93
pluralism (*see* value, pluralism)
Portmore, D. 15 n. 6, 133, 136 n. 32, 139–40, 197 n. 32, 198, 32
practical reasoning 39 n. 5
practical wisdom 26, 288, 310–13
Price, R. 35, 36
Prichard, H. 36, 37
Principle of Equal Consideration of Interests 205, 207, 273 n. 25
promise-keeping 59–78
 and justified beliefs 67, 68, 69
 social practice accounts 63–6
property rights 99–102
punishment 155

quality of will 167–70
Quinn, W. 210 n. 4, 264 n. 10, 265

Railton, P. 101, 110 n. 23, 127 n. 10
rationality 67–8, 70–4
Rawls, J. 63, 306
Raz, J. 35, 51
reactive attitudes 168, 172, 173, 201, 244, 246
reasons 73, 74 n. 18, 197
 agent-centered 183
 conclusive 36, 37, 138 n. 35, 246
 content versus prohibition 168, 169
 contrastive 192, 200
 counterveiling 41, 52, 208, 224
 exclusionary 51, 52
 instrumental versus intrinsic 269, 270
 justifying versus excusing 157
 moral versus non-moral 40–6
 motivating versus normative 166
 objective versus subjective 197, 198, 199
 overriding 36, 37, 46
 practice-based 67
 pro tanto 246
 responding to 171
 sufficient 46
 supererogatory 41, 46
 value-based conception of 35, 38, 39, 45
 weighing model of 45–51
 wrong kind of 62, 66–9
redistribution 102, 115, 122
requirements (*see* moral requirements)
responsibility 248
 moral versus criminal 154–78
 pluralism about 156
risk 114, 116, 296
Robert, R. 287
Rorty, R. 307 n. 46
Ross, W. D. 268 n. 18
rules 26, 303, 304, 305
Russell, D. 288, 307, 312, 313

Sachs, B. 134 n. 28
sanctions 157, 171–5
Scanlon, T. 44 n. 8, 145 n. 41, 167, 173, 264 n. 11
Scheffler, S. 124, 131 n. 20, 134 n. 25, 211 n. 5

Schelling, T. 70–6
Schmidtz, D. 111 n. 27
Schroeder, D. 303 n. 38
Schroeder, M. 74 n. 18, 125 n. 6, 125 n. 7, 133, 134, 135, 147 n. 46, 151
Scott, R. 20 n. 15
self-ownership 98–112
Seneca 30 n. 29
Shah, N. 68
shallow pond 206, 207
Shaw, J. 265 n. 13
Sher, G. 159 n. 8
side constraints (*see* deontic constraints)
Sidgwick, H. 13–15, 19–21, 25
Singer, P. 205, 273 n. 25
Slote, M. 289
Smith, H. 236 nn. 1, 2, 246 n. 13
Smith, M. 137 n. 34, 147 n. 45
social good (see welfare)
social practice 59–62
standard deontic logic 39
Steiner, H. 108
Strawson, P. 157 n. 6, 167 n. 18
strict liability (*see* liability)
Stroud, S. 46
supererogation (*see also* reasons, supererogatory)
  arguments for 24–9
  history of 17–19
  and urgency 25
Swanton, C. 289, 299 n. 30

Tadros, V. 154, 282 n. 32
talent 239 n. 10, 243
Taurek, J. 261 n. 4, 270
Thomson, J.J. 262 n. 6
thresholds 255–8
Timmons, M. 15, 16
toxin puzzle (*see* Kavka's toxin puzzle)
Trianosky, G. 15 n. 4
trolley problems 118, 260

unintended side effects 159 n. 9, 171 n. 23, 260
Urmson, J. 13–19, 24–9
utilitarianism 55 (see also consequentialism)
  and supererogation 24
  objections to 129, 212

Vallentyne, P. 108
value 44, 55, 56, 57
  agent-neutral versus agent-relative theories of 43, 130, 131 n. 20, 205
  as a whole versus on the whole 91
  intrinsic 30, 80
  instrumental 89, 96–7 n. 19
  monism 128
  moral 16, 22, 23, 42, 43, 52–7
  objective versus subjective theories of 119, 120, 121
  pluralism 127, 128
van Roojen, M.
Velleman, J. D. 138
virtue:
  and duty 21
  cognitive versus non-cognitive 302
  epistemology 287
  unity of the 307–13
virtue ethics 13–33, 288
  and practical reasoning 298, 299

Wall, S. 103 n. 9
Wallace, J. 61–2, 74
Watson, G. 158, 304 n. 39, 311
weakness of will (*see akrasia*)
Wedgewood, R. 268 n. 20, 270 n. 23

welfare 100, 107 n. 18, 114–18
Wiland, E. 180, 181
Williams, B. 129, 130, 211, 216,
    292 n. 18, 293
Williamson, T. 49 n. 14
wishful thinking 68
Wolf, S. 157, 237 n. 4
Wood, J. 287
wrongness 40
  moral versus legal 171–4