

**AFTER COMPATIBILISM: ESSAYS ON FREEDOM AND RESPONSIBILITY**

by

**Robert H. Wallace**

---

**Copyright © Robert H. Wallace 2020**

**A Dissertation Submitted to the Faculty of the**

**DEPARTMENT OF PHILOSOPHY**

**In Partial Fulfillment of the Requirements**

**For the Degree of**

**DOCTOR OF PHILOSOPHY**

**In the Graduate College**

**THE UNIVERSITY OF ARIZONA**

**2020**

THE UNIVERSITY OF ARIZONA  
GRADUATE COLLEGE

As members of the Dissertation Committee, we certify that we have read the dissertation prepared by: Robert H. Wallace

titled: After Compatibilism: Essays on Freedom and Responsibility

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy.

*Michael S McKenna*

Michael S McKenna

Date: May 6, 2020

*Terence E Horgan*

Terence E Horgan

Date: May 6, 2020

*Dana Kay Nelkin*

Dana Kay Nelkin

Date: May 7, 2020

*Carolina Sartorio*


Carolina Sartorio

Date: May 6, 2020

*Mark C Timmons*

Mark C Timmons

Date: May 6, 2020

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copies of the dissertation to the Graduate College. 

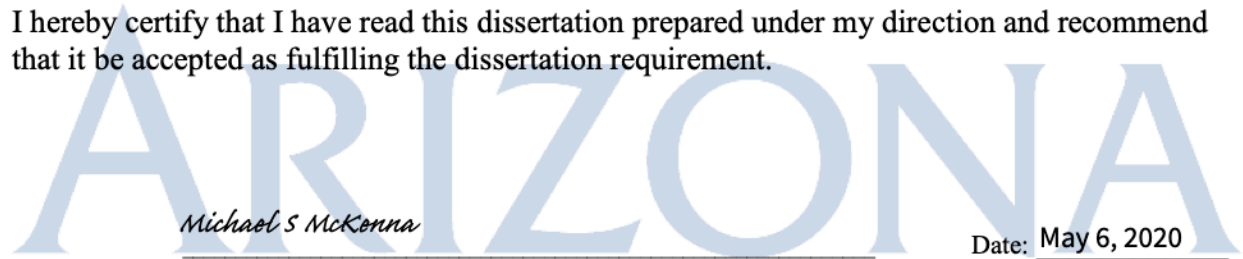
I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

*Michael S McKenna*

Michael S McKenna

Philosophy

Date: May 6, 2020



## Acknowledgements

I have been blessed during my time in graduate school, and I could not have completed this dissertation without the help of many people.

My fellow students and friends have taught me so very much. Thank you to Brandon Ashby, Jacob Barrett, Sameer Bajaj, Bryan Chambliss, Rhys Borchert, Michael Bukoski, Joel Chow, Ben Cilwick, Caleb Dewey, Ding, Luke Golemon, Danielle Guzman, Avital Hazony, Chris Howard, Mario Ivan Juarez-Garcia, Benjamin Keoseyan, Victor Kumar, Robert Lazo, Elizabeth Levinson, Andrew Lichter, Danilo Linhares, Theresa Lopez, Yael Lowenstein, Alexander Motchoulski, Arià Paco Abenoza, David Poplar, Travis Quigley, Sarah Raskoff, Santiago de Jesus Sanchez-Borboa, Alex Schaeffer, Will Schumacher, Dan Shahar, Lucy Shwartz, Anna Bella Sicilia, Jackie Sideris, Eyal Tal, Chad Van Schoelandt, Brandon Warmke, Karolina Wisniewska, Ke Zhang, and YiLi Zhou. I am especially grateful to the members of the free will reading group in spring 2018 for working through chapter drafts: Joe Metz, Josh Cagnelosi, and Phoebe H.Y. Chan. Special thanks to David Beglin, Rosalind Chaplin, Cory Davia, Tim Kearl, Brian Kogelmann, Max Kramer, August Gorman, Joseph Martinez, Nathan Oakes, Jeremy Reid, Kirun Sankaran, Ethan Schwartz, Aaron Segal, Wes Siscoe, Hannah Tierney, Shawn Tinghao Wang, and Sean Whitton for detailed help on the ideas presented in this dissertation. And thanks to my graduate cohort for all their friendship and support over the years: Matt DeStefano, Caroline King, Tyler Millhouse, and Bjorn Wastvedt. I hope I have been at least a little helpful to each of you, since each of you in your own way has been helpful to me.

Thank you to my teachers at the University of Arizona: Julia Annas, Stewart Cohen, Juan Comensaña, Rachana Kamtekar, Shaun Nichols, Connie Rosati, Jason Turner, Jonathan Weinberg. I owe Jenann Ismael a special debt of gratitude for her encouragement and guidance. Thank you to the faculty working on agency and responsibility who welcomed me during my time at UCSD: Lucy Allais, David Brink, and Manual Vargas. Special thanks to Monique Wonderly for your wise advice and friendship. Thanks to everyone who participated in the UCSD Agency and Responsibility Group for working through what became chapters 4 and 5 below. Thank you to the audience at the 2016 Gothenburg Responsibility Project conference for insightful questions and commentary, especially Per-Erik Milam, Derk Pereboom, and Paul Russell. Thanks to David Shoemaker for commentary and criticisms of what eventually became chapter 2 of this dissertation. I'd also like to thank my undergraduate philosophy professors at Kenyon College: Joel Richeimer, Yang Xiao, Juan DePasquale, Hans Lottenbach, and especially my thesis advisor Rebecca Waller. Special thanks also to Donald Rogan, the first person to suggest something I had written was worth publishing all the way back in undergrad. Rest in peace. Thank you for getting me ready for graduate school and beyond! Finally, thank you to Gary Gutting for mentorship during the years I was out of school; rest in peace.

Michael Mckenna, my advisor, has been a constant source of support, teaching, and friendship throughout my graduate career. Thank you for believing in me. You once told me after reading something that would one day become chapter 2 of this dissertation (and then my first publication) that I had a unique way of thinking, and that I ought to take the time to nurture it. I have tried and will continue to try my best to do so. I hope that this goes a little way towards paying back all of the good will you have shown me over the years.

Carolina Sartorio's guidance has been crucial for my philosophical development. Thank you for being so encouraging, from the first day I sat in your metaphysics class to the last time we discussed a paper. Thank you for helping me to see the value of the philosophy article and being a role model for me as a researcher, scholar, and mentor.

Thanks to Mark Timmons for much sage advice on writing, research, and publishing; for lots of philosophical conversations; for being a living embodiment of the idea that a little history of

philosophy is good for both the mind and the soul. Thank you for teaching me to interpret charitably and to look again at things with new eyes. Thank you for all the kindness over the years.

Thank you to Terry Horgan for teaching me how to think about metaphilosophy; for getting me to take alternative possibilities seriously; for all the wonderful dialogue over good beer. Thank you for showing that breadth of interest and depth of insight are not mutually conflicting goals of philosophical research.

I owe many thanks to Dana Kay Nelkin. You generously took me in as your student and offered thoughtful feedback on pretty much everything I had written or was thinking about. I very much appreciate your time and energy. I hope it's clear from this dissertation how much I learned from you in a short time.

Thank you, Mom and Dad, for raising me in a home where love of learning was a cherished value. Mom, thanks for all the time spent thinking and arguing about the news and ideas of the day. My formative years were spent developing the skills I'd need to complete this work. You had no idea what you were in for! Dad, it has taken me a good long while to appreciate how much your continuing liberal arts education meant to be. Thank you for being a model for what lifelong learning looks like. Thank you to my brother, Charlie, for being a constant source of support. You wouldn't know it, but our conversations have had a bigger impact on my philosophical research than you realize.

There isn't enough space to properly thank my wife, Elisa Pelgrift, for all her love and support throughout the trials and tribulations of pursuing a PhD. (It was suggested to me by a friend that I give you an entire page, but even that wouldn't suffice). I can condense it all down to a simple counterfactual: I couldn't have done it without you. That's not as illuminating, though, as putting it this way: what I can do, I can do because of you. Thank you for everything.

I'd like to dedicate this dissertation to my grandparents, John and Julia Hughes.

## Table of Contents

<b>Abstract .....</b>	<b>7</b>
<b>Introduction .....</b>	<b>8</b>
<b>Chapter 1: Abilities-Based Compatibilism.....</b>	<b>16</b>
1.1 Introduction .....	16
1.2 Abilities, Dispositions, and The Gap Problem .....	17
1.3 First Horn: Closing the Gap .....	24
1.4 Second Horn: Living with the Gap .....	31
1.5 An Objection Considered and the Problem Generalized .....	35
1.6 Further Generalizing.....	37
1.7 Conclusion.....	40
<b>Chapter 2: Responsibility and the Limits of Good and Evil.....</b>	<b>41</b>
2.1 Introduction .....	41
2.2 The Basics of Strawsonianism .....	42
2.3 The Really Bad Case.....	45
2.4 A Possible Solution .....	48
2.5 The Really Good Case .....	51
2.6 Elevation and Disgust.....	53
2.7 Peculiar Reactive Attitudes .....	56
2.8 Solutions .....	62
2.9 The Ethics of Disgust.....	64
2.10 Conclusion.....	68
<b>Chapter 3: P.F. Strawson and the Case of the Missing Account of Control.....</b>	<b>70</b>
3.1 Introduction .....	70
3.2 Strawson’s Main Argument.....	71
3.3 The Specter of Incompatibilism.....	74
3.4 Themes from “Freedom and Resentment” .....	77
3.4.a Sentimentalist Deontology .....	77
3.4.b Natural Commitments .....	83
3.4.c The Viewpoint of Human Affairs .....	87
3.4.d Gains, Losses, and Making Sense of a Life .....	92
3.5. Where Do We Go from Here?.....	97
3.6 Conclusion.....	99
<b>Chapter 4: An Argument for Realism about Free Will.....</b>	<b>100</b>
4.1 Introduction .....	100
4.2 Realism, Commonplaces, and Compatibility .....	102
4.3 The Form of the Argument.....	107

4.4. Getting Precise About Practical Absurdity.....	110
4.5 Understanding the Generalization.....	114
4.6 What Makes the Generalization True?.....	117
4.7 Particular Versions of the Argument.....	119
4.8. Assessing the <i>Reductio</i> .....	121
4.9 Conclusion.....	126
<b>Chapter 5: Compatibilism and Responsibility-Based Realism about Free Will .....</b>	<b>127</b>
5.1 Introduction .....	127
5.2 Responsibility-Based Arguments .....	129
5.3 The Easy Case for Compatibilism .....	137
5.4 ... And Why It's Wrong.....	138
5.5 Compatibilism Revisited.....	141
5.6 Begging the Question .....	144
5.7 Argumentative Burdens and Agentive Phenomenology .....	145
5.8 The Deeper Dispute .....	150
5.9 Conclusion.....	152
<b>References .....</b>	<b>153</b>

## Abstract

This dissertation is a series of standalone essays. Together, they form a critique of contemporary compatibilist approaches to the problem of free will and determinism, and they offer an alternative methodology for approaching questions about freedom and responsibility. Compatibilist approaches to the free will problem exist on a spectrum from the more normative to the more metaphysical. Views at the metaphysical end of the spectrum typically understand free will in terms of abilities. In Chapter 1, I argue that these views face a powerful dilemma: they either fail to explain these abilities or fail to show that these abilities are compatible with the thesis of physical determinism. Perhaps a commitment to abilities could be given up, but I argue that takes us too far afield from the intuitive way we understand ourselves as free agents. Compatibilist approaches at the normative end of the spectrum have been largely influenced by P.F. Strawson's responsibility naturalism. Views of this sort begin by carefully attending to the features of our responsibility practices in order to glean the nature of the sort of freedom that grounds apt responsibility ascriptions. In Chapter 2, I defend a version of this view from a decisive objection: Strawsonian compatibilism seems to make evildoers exempt from moral responsibility. Nevertheless, in Chapter 3, I argue that Strawson's program cannot properly insulate itself from metaphysical concerns about abilities. The methodology may actually support a powerful form of incompatibilism about free will and determinism. This casts the entire contemporary project that draws on his work in a suspect light. Drawing lessons from these two failures, in Chapter 4, I offer a practice-based argument for realism about freedom and responsibility—the view that we really are free and responsible—that is neutral with respect to questions about the compatibility of freedom and determinism. The argument does not depend on any particular metaphysical theory of abilities or particular view of our moral practices. I argue in Chapter 5 that, given this realist framework, we have reason to think that whatever ends up being true about the abilities that characterize free and responsible agents, they will be compatible with determinism after all.

## Introduction

This dissertation is a series of stand-alone essays addressing topics related to free will and moral responsibility. In spirit, though, the essays are almost a monograph. The first three essays struggle with the possibility that two prevailing contemporary compatibilist methodologies in the free will debate, the debate about whether or not free will is compatible with the thesis of physical determinism, have failed. The last two essays offer an alternative way of framing questions about freedom, responsibility, and determinism. This alternative starts by thinking in terms of realism and anti-realism (or nihilism). By beginning with this alternative framing in mind, we can motivate reasons to adopt compatibilism in the free will debate in a way so as to avoid the problems of the compatibilist views under discussion.

I've barely gotten started and I already have to anticipate an objection. Why the hullabaloo about how wrongheaded compatibilism is if I'm going to go ahead and endorse it again?

Well, we can locate compatibilist approaches to the free will problem on a spectrum from the more normative to the more metaphysical. Views at the ends of this spectrum enshrine important, if somewhat platitudinous, true principles about what it takes to be a free and responsible agent. On the one hand, persons are free because they are able to do things in a way that is distinctive of persons among agents. On the other hand, persons are aptly held to the demands of morality, itself distinctive of persons among creatures. Nevertheless, there is a tension between these two platitudes about the nature of persons, made apparent when thinking them through in the context of the problem of free will and determinism. Push on the claim about abilities too hard, and we seem prone to philosophical confusion about the nature of the basic normative competencies that characterize ourselves as moral agents. Push on the moral claim too hard, and it suddenly seems like these basic moral competencies require abilities that push the limits of a plausible naturalistic metaphysics. My interest in exploring these contemporary compatibilist views, then, amounts to a desire to explicate and explore this tension within a particular context.

Why think about this tension in the context of compatibilism about free will and determinism? It is important that compatibilism already tends to deal with the tension between our two plausible platitudes by being conciliatory. Compatibilists attempt to reconcile the tension by somewhat dashing our metaphysical aspirations. Indeed, R. Jay Wallace (1994: 6) has called a desire for some freedom beyond what is required to be practical moral agents fetishistic! Better to settle with a modest sense of freedom, then, one consistent with our moral responsibility practices, and nested within a plausible picture of morality. *This* is the sort of resolution to the tension which I aim to supplant.

Maybe a different dissertation by a different author with different metaphysical sympathies would have considered libertarian views at either ends of this spectrum to illuminate this tension. There would however be a sort of red herring to deal with by considering libertarian views alongside their compatibilist competitors. A simple formula for generating a libertarian view of free will goes like this: take a plausible compatibilist view. Now, at the critical juncture, once you locate it, place an indeterministic process or event or choice or what have you. Presto! You now have a plausible libertarian view.<sup>1</sup> I don't mean to be facetious here about libertarianism. My point is simply that the libertarian requirement that free action involve indeterminacy is best thought of as an independent necessary condition on free action, if indeed free action is shown to be incompatible with physical determinism by argument. If free action is not compatible with determinism, it is very seriously worth considering whether or not we meet this condition. But it would be one necessary condition among many others. No one has an exhaustive list of necessary and sufficient conditions here.<sup>2</sup> Nevertheless, the rest of the items on that list will be *prima facie* compatible with determinism. Starting with libertarian

---

<sup>1</sup> Al Mele (2006) and Christopher Franklin (2018) both adopt this sort of “minimalist” take on libertarian views.

<sup>2</sup> Even talking about necessary and sufficient conditions here is sort of a conceit here on my part; we can't even generate a list of necessary and sufficient conditions for action let alone free action (cf. Clarke 2015: 901). Better to think about offering satisfactory explanations of free action rather than an analysis of free action. As I tell my students, talking in terms of necessary and sufficient conditions can nevertheless be a useful way of thinking through a philosophical problem, even if, eventually, conceptual analysis should be abandoned in favor of a more liberal *conception* of the item under investigation.

views of free will, then, seem to only complicate the matter at hand, namely, of trying to synthesize the two plausible thoughts about persons: they are uniquely able to do things and they are uniquely accountable to the demands of morality. Best methodological practice suggests starting with compatibilist views as stand-ins for more general considerations about what it takes to be a free and responsible agent.

If they do not complicate the matter at hand, then at the very least, libertarian views of free will generally do not offer the sort of resolution I am interested in. Earlier, I said that we seem prone to philosophical confusion about the nature of the basic normative competencies that characterize ourselves as moral agents when we consider that persons have unique abilities. Compatibilists tend to deflate these abilities and so arguably misconstrue the nature of our moral competence. (More on this in a moment). Incompatibilists tend conversely to inflate these abilities. They push the limits of a plausible naturalistic metaphysics in defense of ordinary moral competence. This is obviously not conciliatory. I like that about these views. By my lights, however, inflating our abilities is not a good idea. It misconstrues the nature of our moral competence. (Consider, for instance, C.A. Campbell's classic idea in his 1957 *On Selfhood and Godhood*, that free will consists in the ability to will oneself to comply with morality in the face of temptation, basically, no matter what; doesn't this suggest an overly narrow picture of the moral life?)

So, compatibilist approaches to the problem of free will and determinism enshrine two plausible platitudes about the nature of free and responsible persons. They offer a conciliatory way of dealing with the tension between them. But being conciliatory comes with dangers. And this is the subject of the first three essays.

We might for instance be too modest, too willing to squish our ordinary action-theoretical talk into neat naturalism-friendly boxes when we consider the metaphysical aspects of free action. The classic conditional analysis of free will, for instance, suggests that a person is free when she is able to

do otherwise, and she is able to otherwise just in case if she had wanted (or chosen, or tried, etc.) to do otherwise, then she would have (e.g. Ayer 1954, Moore 1912). Indeed, Ayer suggests that all we mean when we say a person could have done otherwise is for that counterfactual to be true, for their action to be voluntary (in a psychological sense), and for no one to have forced the person to perform the action they in fact performed (1954: 117).

This analysis famous fails for a variety of reasons, perhaps most recognizably by counterexample (Lehrer 1968). But the heart of this compatibilist project, of reducing talk of freedom-grounding abilities, which strike some as requiring somewhat ontologically demanding notions of powers or abilities, to less demanding modal notions is ongoing. Fischer and Ravizza (1996) discuss the abilities of agents in terms of dispositions to respond to reasons, and this has in many ways been the dominant compatibilist way of thinking about freedom-grounding abilities since. It remains so even when we talk about causal powers or the causes of action rather than dispositions, for instance (e.g., Nelkin 2011, McKenna 2013, Sartorio 2016).

Although I agree that the ability to respond aptly to reasons is among the moral competencies distinctive of persons (and who wouldn't agree, the idea goes back to Aristotle!), I worry that this project misconstrues the nature of reasons-responsiveness. The sales pitch goes like this, simplifying a great deal across various disagreements: you can have all the free will worth wanting, and it can be bought for the low ontological price of the pertinent counterfactuals, or the pertinent dispositions, or the pertinent causal powers, or the right kind of causal sequence. Again, I don't mean this facetiously. It's a good sales pitch. It offers a tempting philosophical project, a picture of freedom consistent with a plausible naturalistic way of thinking about the world. Perhaps, though, there is something special about the sorts of abilities that make agents free and responsible, which can't be bought by what contemporary compatibilists are selling. Or so I will argue. (More on that in Chapter 1).

We might instead, and somewhat ironically, be immodest in being conciliatory. Consider for instance Moritz Schlick's (1939) insistence that the problem free will and determinism was a pseudo-problem in ethics. What kind of freedom is required for us to be morally responsible? "It is easy to attain complete clarity in this matter." (1939: 151). All we need do is carefully reflect on our moral practices—and in particular, what we count as responsibility undermining—to find out. As he puts the pertinent questions: "What is the case in practice when we impute 'responsibility' to a person? What is our aim in doing this?" (1939: 151). Lo and behold, when we answer these questions, we see that the sense of freedom at plays here is compatible determinism; Schlick had a (deeply wrongheaded) utilitarian view of these practices that was obviously compatibilist-friendly; it led him to an overly simple conditional analysis of free will. Nevertheless, there is something powerfully explanatory in this all-too-quick Schlickian thought. My own suggestion for how to proceed, you will see, shares a kind of family resemblance to it. P.F. Strawson (2008/1962: 35) essentially, and famously, took up this same strategy while aiming to correct its "one-eyed utilitarianism". By careful reflection on our practices of moral responsibility, we can glean the nature of the sort of freedom which underlies apt responsibility ascriptions. More recent articulations of "Strawsonian" compatibilism offer more sophisticated variations of this very same strategy (e.g. Wallace 1994, Watson 2018).

It is, unfortunately, never *easy* to settle the matter. One sort of problem, which I think is surmountable, falls out of attempts to offer an accurate picture of our moral responsibility practices. We seem forced with a never-ending quest to get the extension of the class of morally responsible agents right. Plausible accounts of this kind, ones that make sense of ordinary agents, seem to exempt the very evil and very heroic among us. The quest is not endless, however. It will just take a lot of time and an imaginative look at our moral psychology and our moral practices. But another problem is not surmountable. A detailed look at the actual features of our practices, even when well described, appears to lead straightforwardly to incompatibilist results by appearing to implicate things about the

metaphysics of abilities. (More on this in Chapters 2 and 3). I think this appearance is mistaken. Yet it is very difficult to show why.

My suggestion in light of all of this is to find a way to properly connect our two platitudes. The way to do so is by locating them in an argument for realism about free will that is exceedingly neutral with respect to both (1) substantive views about the metaphysics of abilities and (2) substantive views about the nature of morality. In its most basic form, it's also a simple transcendental argument. Since we *are* competent with respect to the demands of morality, we *must have* whatever sorts of abilities are required to be so competent. These abilities are excellent candidates for what free will actually consists in. Such arguments, I think, have quite a pedigree. Somewhat ironically, I think libertarians about free will have historically been clearer about the real nature of this kind of argument. I am inclined to think that such arguments are as good of a reason as any so far given to be a realist about free will.

I will for the most part not get the weeds about the viability of world-directed transcendental arguments. We could of course rehearse Stroud's (1968, *inter alia*) famous attack on this sort of thing. I think a charitable version of philosophical naturalism ought to accommodate the thought that there is an interdependence between what we actually are and what we have to believe ourselves to be *qua* persons. "Person" is perhaps uniquely among all kinds plausibly understood as both a natural and as a social kind. This interdependence is thus to be expected. By my lights, this undermines the classic worry that transcendental arguments are merely mind-directed when they purport to be world-directed.<sup>3</sup> I make this interdependence apparent by considering the conditions under which the possibility of having the sort practical lives persons normally do could make sense. My preferred version of such arguments, then, is a kind of *reductio ad absurdum*. It is absurd, in a particularly practical

---

<sup>3</sup> Perhaps I am wrong about this. If so, then I can with Strawson (1985) retreat to the not so unpleasant thought that the sort of argument under consideration draws out some of the basic structural features of our conceptual scheme.

way, to think persons are not morally responsible. Since free will is plausible whatever abilities or competences are required to meet the control condition on moral responsibility, we therefore have free will. (Some of this appears in Chapter 3, and more explicitly in Chapter 4).

With this kind of argument in hand, we can go back to questions about free will and determinism. What does the absurdity of anti-realism about free will mean? For one, it shows that the deeper dispute between compatibilists and libertarians is about the epistemic standards for justified belief in freedom and responsibility. The views I consider agree that we in fact have justification for this belief. They just disagree about what evidence we have, and the latter thinks the former is missing something important. This will shift us to matters phenomenological. But reframing the debate this way favors compatibilism.

How so? Earlier, I set aside libertarian views about free will in setting up the tension between the platitudinous thoughts that (1) persons are free because they are uniquely able to do things among agents and (2) persons are uniquely among creatures morally responsible. I did so because libertarians accept a condition on free action which I do not, and which would complicate matters: that an action be free from prior determination. The sorts of considerations at work in the argument for realism seems to show that we don't need evidence that the actions of free agents are free from prior determination. And, well, "since there can only be a *lack* where there is a *need*" (Strawson 1985: 41), a compatibilist friendly picture of freedom and responsibility isn't lacking. The form of compatibilism I endorse is obviously anything but conciliatory. Our metaphysical aspirations are not something to set aside, but something to cast off. Maybe this sounds immodest like Schlick and Strawson, but I don't think so. Nothing I say shows that libertarianism about free will is logically confused or something. I just think compatibilism amounts to good judgement about the kind of evidence that supports realism about free will. (More on this in Chapter 5).

I have thus titled this collection “After Compatibilism: Essays on Freedom and Responsibility”. I mean the “After Compatibilism” bit in two different senses. In the first sense, this dissertation articulates what *comes after* contemporary compatibilism. In the second sense, it is about *going after* compatibilism again.<sup>4</sup>

Even if this dissertation is almost a monograph, it is of course just a series of essays. There is only so much it can do. Nevertheless, it is my suspicion, and a hope for future research, that the kind of methodology I develop here can show that, quite generally, there are many ways the world could be that are consistent with the existence of free and responsible persons. If so, we might more clearheadedly think about abilities, about action, about morality, and about persons. This is work for another time. I am satisfied with the thought that, even if things are not fully settled, there is one thing—physical determinism—to worry a little less about.

---

<sup>4</sup> The title is also a riff on the final chapter of Sean Nichols’ *Bound* (2015) titled “After Incompatibilism”, which argues that even if incompatibilism is true, we can and should retain moral anger and ordinary forms of moral responsibility on practical grounds.

## Chapter 1

### A Dilemma for Abilities-Based Compatibilism

#### Abstract

A common compatibilist view says that we are free and morally responsible in virtue of the ability to respond aptly to reasons. Call the view reasons-responsive compatibilism. Many hold a version of this view despite disagreement about whether free will requires the ability to do otherwise. Although I am a compatibilist myself, I argue that this view faces a significant challenge: reasons-responsive compatibilists cannot adequately explain why the ability to respond aptly to reasons is compatible with determinism. The problem has the form of a dilemma. I argue that this dilemma may generalize to other plausible ability-based versions of compatibilism, views that ground our freedom and responsibility in abilities. The problem therefore places a significant burden on compatibilists in the classic free will debate, which is cast in terms of abilities. I conclude by briefly considering the prospects for an abilities-neutral conception of compatibilism.

#### 1.1 Introduction

There is a popular inspirational quote shared on the internet that captures the nature of free agents: “between stimulus and response there is space. In that space is our power to choose our response. In our response lies our growth and our freedom.”<sup>5</sup> Intuitively, it is this actionable “space” between what happens to us and what we do that makes us morally responsible beings.

This spatial metaphor is, of course, just a metaphor. But it gestures at an important conceptual point about the nature of free agency. Free agents have an ability to choose (or decide) what to do. And in virtue of this ability, free agents settle for themselves what they are going to do. Put otherwise, free agents make a difference to what happens by exercising their ability to choose or act. This difference-making is unlike the happenings that result from the activity of non-agents in the world, for whom what happens next is as if just a reaction to what came before.

Although I am a compatibilist, I will argue that this basic conceptual point about the kind of ability that characterizes free agents is the source of a serious problem for a widespread compatibilist view. The view holds that we are free and morally responsible in virtue of the ability to respond aptly

---

<sup>5</sup> The quote is often misattributed to Viktor Frankl. See: <https://quoteinvestigator.com/2018/02/18/response/>.

to reasons. Call the view *reasons-responsive* compatibilism.<sup>6</sup> Wolf (1990), Fischer and Ravizza (1998), Nelkin (2011), McKenna (2013), Vargas (2013), and Vihvelin (2013), among others, each defend some version of this view despite disagreement about whether free will requires the ability to do otherwise.<sup>7</sup> The challenge I raise is this: reasons-responsive compatibilists cannot adequately explain why the ability to respond aptly to reasons is compatible with determinism. The problem has the form of a dilemma. I defend this dilemma from objections. I argue that the dilemma may generalize to other plausible ability-based versions of compatibilism, which ground freedom and responsibility in abilities. I am therefore skeptical that any ability-based compatibilism can succeed.

## 1.2 Abilities, Dispositions, and The Gap Problem

A traditional way of construing the problem of free will and determinism begins with the necessary and sufficient conditions for moral responsibility.<sup>8</sup> One of these necessary conditions is having a certain kind and degree of control over one's actions. Call the kind and degree of control sufficient for meeting this condition on moral responsibility *free will*. A common assumption among the majority of disputants in the free will debate is that the control condition is to be understood in terms of abilities. So, to say that there is some free will ability is to say that there is an ability, or collection of abilities, such that when had by an agent, that agent meets the control condition of moral

---

<sup>6</sup> "Reasons-responsive compatibilism" usually picks out views in the style of John Martin Fischer and Mark Ravizza (1998). I use the term more broadly. For instance, I include Susan Wolf (1990) and Dana Nelkin (2011) in this camp, who defend what is sometimes called the "Reason view". Kadri Vihvelin's (2013) view is often labeled a form of "dispositional compatibilism", since her view involves the disposition to choose on the basis of reasons. Why do I use the name "reasons-responsive compatibilism" to denote all of these views? I take the idea that free will is to be had in the ability to respond aptly to reasons as the central feature of each view, in spite of disagreement over other things—for instance, about the kinds of reasons one must be sensitive to or whether or if moral responsibility requires the ability to do otherwise.

<sup>7</sup> I omit Sartorio (2016) from this list. Her view might be construed in ability-neutral terms. More on this later.

<sup>8</sup> It is generally accepted that there is at least one other condition: a knowledge condition. It is an open question whether not the conditions for moral responsibility are exhausted by these two conditions. I discuss the problem in terms of necessary and sufficient conditions for the sake of clarity in presenting the problems. Less strictly, what we are after is a theory or explanation of what moral responsibility requires. I should also note that although it is not uncontroversial to frame the problem in terms of moral responsibility, it is also not universally accepted. For some likeminded philosophers, see: Pereboom (2001: xxii), Mele (2006: 17), McKenna (2008: 187), and Sartorio (2016: 7-8).

responsibility. Then arguments are marshalled as to whether some candidate free will ability is compatible with determinism, the thesis that the past and the laws of nature entail one unique future.<sup>9</sup>

How might a compatibilist proceed? They need to offer (1) an intuitively candidate free will ability, which is (2) compatible with determinism. Consider a classic view. Philosophers like Ayer (1954), Schlick (1939), and Moore (1912) attempted to reduce abilities to sets of counterfactual conditionals, e.g., of the form: ‘If I had tried to do otherwise, then I would have done otherwise’.

These counterfactuals were meant to reveal and reductively explain freedom-relevant capacities of the agent. Compare: in a deterministic world there is no obvious threat to the counterfactual claim that if I were to drop salt into water, it would dissolve. This counterfactual claim tells us something important about the nature of salt, namely, that it has the disposition to dissolve in water. Likewise, counterfactuals about what some agent would have done if they had tried to do differently reveals something important about our nature as agents: So, we move from abilities to determinism-friendly counterfactuals, identifying the free will ability as an ability to do otherwise. This is a clever tactic. But this sort of account fails by extensional inadequacy (e.g. Lehrer 1968). It renders the verdict that agents are able to do what they cannot do (like eat candy that they are psychologically incapable of wanting) The view falsely suggests that agents are able (in the wide sense) to do what they lack the opportunity to do.

How might one keep the cleverness of this strategy, without falling prey to its extensional failure? Instead of focusing on counterfactuals, we could focus on dispositions as a reductive base for abilities. What is a disposition? Let’s consider an example. Fragility is the disposition of some object to break (and so manifest fragility) when put into certain kinds of situations, like being dropped. In other words, dispositions *manifest* under specifiable *stimulus conditions* (cf. Martin 1994, Lewis 1997,

---

<sup>9</sup> More formally, determinism is the thesis that two propositions, one describing the past at some given time, and another describing the laws of nature, together entail a proposition describing the one unique future.

Manley and Wasserman 2007, *inter alia*). Dispositions are presumptively compatible with determinism. We can see why by thinking about the close relationship between dispositions and counterfactuals. In a deterministic world there is no obvious threat to the counterfactual claim that if I were to drop salt into water, it would dissolve. And this is true of salt which never comes into contact with water. So, salt's solubility is not threatened by determinism. If one were to offer a reductive analysis of the free will ability in terms of dispositions, one could secure the ability's compatibility with determinism.

This is the strategy that many reasons-responsive compatibilists take. For the sake of argument let's assume that the ability that gives us control sufficient for moral responsibility is the ability to respond aptly to reasons (where this ability involves being able to respond to a relatively high degree of such reasons).<sup>10</sup> In other words, the free will ability is reasons-responsiveness. Here is an apparently promising way, it seems, to secure the compatibility of free will and determinism: reduce the reasons-responsive ability to the dispositions necessary and sufficient for having that ability. They are presumptively compatible with determinism. And if the ability is identified with these dispositions, then the agent's ability is not mysterious in the least. Her ability to respond aptly to reasons is just her being *disposed* to respond aptly to reasons. We can moreover learn about the nature of her rational dispositions by imaginatively checking what reasons she would take as sufficient in counterfactual situations. Fischer and Ravizza, for instance, appear to take up this strategy.<sup>11</sup> First, they analyze reasons-responsiveness as dispositions to recognize and react to sufficient reasons and follow through by explaining these dispositions in terms of a collection of true counterfactuals (1998 sections IV.2

---

<sup>10</sup> I'll omit this qualification in the discussion to follow.

<sup>11</sup> McKenna (2019: 23, ft. 15) claims that Fischer and Ravizza never commit to a reductive view. Nevertheless, I believe it is the most promising interpretation of their view, since it is the dialectically easier way to secure the compatibility compared to the alternative McKenna suggests for them. I'll discuss that alternative in section 6. Thanks to an anonymous reviewer for raising this issue.

and IV.3). This provides an explanatorily fleshed out view of what the ability to respond to reasons is, how we can know about it, and why it is compatible with determinism.

A recent instance of this reductive strategy (disregarding areas of disagreement with Fischer and Ravizza) is found in Vihvelin (2013). She argues that our narrow abilities, abilities in the sense of those things we can do but are as yet unexercised, ‘are structurally like the so-called intrinsic dispositions of medium-sized objects—dispositions like fragility, elasticity, solubility, and so on’ (2013: 169). How so? They are both three-place relations: fragile glass has a disposition to break when dropped, and I have the ability to speak French given the opportunity. She goes on to say that ‘our narrow abilities’, what in us makes us able to do things, ‘are either intrinsic dispositions or bundles of intrinsic dispositions’ (2014: 169). And she goes on to argue that the abilities relevant to free will are wide abilities to respond to reasons. Wide abilities involve not only narrow abilities but also opportunities to exercise them. Opportunities on this view are understood as situations amenable to the manifestation of our dispositions. And since dispositions are compatible with determinism, we have a compelling compatibilist account of abilities.<sup>12</sup> This reductive approach may go some ways towards explaining, as Vargas (2013: 139, ft. 6) put it, the ‘uncontested consensus about the possibility of reasons sensitivity even under determinism.’

Despite the cleverness of this strategy, some incompatibilists have been skeptical about the sense of ability to be gleaned from this kind of reductive project (e.g. Van Inwagen (1983), *inter alia*). The source of this skepticism stems from a simple point: dispositions do not seem to be the same thing as abilities. Here is Peter van Inwagen’s archetypal way of making the distinction between dispositions (causal powers and capacities) and abilities (agentive powers):

---

<sup>12</sup> To be clear, Vihvelin argues that free will and determinism are compatible prior to her endorsing this dispositional view of abilities by attacking incompatibilist arguments. That our abilities are explicable as bundles of dispositions is meant to explain *why* the ability to respond aptly to reasons is compatible with determinism. More on this in section 4.

For a man to have the capacity to understand French is for him to be such that if he were placed in certain circumstances, which wouldn't be hard to delimit, and if he were to hear French spoken, then, willy-nilly, he would understand what was being said. But if a man can *speak* French, it certainly does not follow that there are any circumstances in which he would, willy-nilly, speak French. The concept of a causal power or capacity would seem to be the concept of an invariable disposition to react to certain determinate changes in the environment in certain determinate ways, whereas the concept of an agent's power to act would seem not to be the concept of a power that is dispositional or reactive, but rather, the concept of a power to *originate* changes in the environment. (1983: 10-11)

In other words, abilities involve something more than the mere reactivity displayed by dispositions.<sup>13</sup>

But what is this something more? (And does van Inwagen beg the question?)

Well, the sense of 'can' involved in the statement 'the fragile vase can break when dropped' involves no power of the vase to instigate novel changes. Instead, the vase's fragility involves a reaction to the condition of it striking a hard surface. One could think of capacities, skills, and talents in a likewise manner. For instance, one cannot help understanding spoken French, if one is fluent in the language. I myself am disposed to become angry when someone insults me, disposed to laugh at a good joke, and disposed to sweat when I run. Dispositions like these, whether had by objects or persons, involve *stimulus conditions* and *manifestations* of inner natures. A stimulus condition is a set of circumstances that primes a reaction. And to manifest something is to perhaps express and make apparent some inward nature. Now, a disposition can be interfered with (finked or masked) such that it will not manifest in its stimulus condition.<sup>14</sup> For example, a fragile glass will not break when wrapped in bubble wrap. Absent such interference, dispositions manifest, as it were, willy-nilly.<sup>15</sup>

---

<sup>13</sup> Three points. First, van Inwagen means 'willy-nilly' in the sense of happening like it or not, rather than happening haphazardly. I'll use it in this sense too. Second, ordinary talk of abilities is quite permissive. We can say perfectly well of my car that it is able to handle rough terrain in virtue of its four-wheel drive and suspension. The abilities at issue in this discussion are the sort at issue in the free will debate, which are relevant to human agency. Finally, I should note that many dispositions are multi-track. They are best characterized by multiple pairs of stimulus conditions and manifestation conditions. Van Inwagen talks about 'invariable dispositions' in his criticism of the compatibilist view, but this might just be a mistake. The dispositions under discussion are all plausibly multi-track dispositions.

<sup>14</sup> For examples, see Martin (1994) and Lewis (1997).

<sup>15</sup> To simplify things, I'll ignore the possibility of indeterministic dispositions. I'll return to them later in section 5.

By contrast, one is able to perform some action if one is put together the right way so as to have the power to do said action and one has the *opportunity* to *exercise* that power. An opportunity is a set of circumstances that makes an action possible. Opportunities are not mechanism triggers. Opportunities are not finked or masked. They are seized upon or missed. One *elects* if and when to speak. When I speak French, I do more than manifest or express my inward knowledge of its grammar and vocabulary in certain circumstances. I do more than manifest the necessary features of myself relevant to speaking French. I have control over these elements. I direct them and govern their expression such that I have, in the relevant sense, the ability to speak the language. I myself settle the question of what I will say.

So, manifestation and exercise are conceptually distinct. And so too are the concepts of opportunity and stimulus. And we can highlight this distinction by thinking about what dispositions and abilities do. Dispositions and abilities can both make a difference to what happens. However, there is apparently an important dissimilarity in the way that each can make a difference (cf. Nozick 1981: 311-313). Dispositions can make a difference insofar as some outcome would have been different if the disposition had not been made manifest. They can also make a difference insofar as their non-existence, absence, or their being interfered with would not bring about the same outcomes.<sup>16</sup> But abilities can do more than contribute to a change in outcomes given circumstances; instead, abilities are the origin of such changes.

As a conceptual point, one can see how it alone might give rise to the thought that if determinism is true then no agent can ever make a difference in this way. (Perhaps this thought lurks in van Inwagen's remarks). Yet this presupposes too much about the pertinent notion of origination. I take it that the pertinent notion leaves open the compatibility question. Having an ability does not

---

<sup>16</sup> Sartorio (2005, 2016) argues that causation involves this kind of difference making.

necessarily mean one has a power to originate changes in a way that simply requires that there are no antecedent sufficient conditions for whether or how one originates those changes. It's just that when an agent exercises her ability, she settles *that* such a change occurs, given the opportunity to do so.<sup>17</sup>

In other words, there is a conceptual gap between our concept of dispositions on the one hand and our concept of agentic abilities on the other. The intuitively freedom-grounding aspect of these abilities seems missing from dispositions. So, this aspect of these abilities seems unexplainable by dispositions as a suggested reductive base. Call this the gap problem. One need not endorse any strong conceptual constraints on successful reduction in order to advance the gap problem. It's just that dispositions seem either wrong as an analysis of abilities altogether, or deeply ill-suited to explain them.<sup>18</sup> One might think that the gap problem poses a serious problem for any reductive analysis of the free will ability. So, it poses a problem for reductive accounts of the free will ability as the ability to respond aptly to reasons.

One might think that the gap problem poses a serious problem for a reductive analysis of abilities, and thus the project of the compatibilist account of the free will ability as the ability to respond aptly to reasons. Herein lies a dilemma. There are two ways of developing reasons-responsive compatibilism to get around the gap problem. Each strategy fails.

First, one could attempt to close the conceptual gap between abilities and dispositions and show that dispositions are in fact a suitable reduction base for abilities. The most plausible way to do so is to invoke what I will call *active* dispositions: dispositions that involve the activity of the agent.

---

<sup>17</sup> I acknowledge the apparent tension between this claim and the claim that the difference-making conception of abilities leaves open the compatibility question. But the tension is only apparent as it is, again, the result of assuming too much about the pertinent notion. I'll discuss this further in section 5.

<sup>18</sup> Clarke (2009: 338-339, 2015: 901) argues for a similar problem. He suggests that having many kinds of abilities, and in particular the ones at issue in the free will debate, cannot simply be a matter of having a disposition (or bundle of dispositions). Abilities involve both underlying competencies (plausibly construed as dispositions) but also something more. When it comes to the abilities at issue in the free will debate, perhaps this something more is the choice to exercise them. So, dispositions are necessary but not sufficient for having the sort of abilities at issue. The gap problem is a distinct worry, but it explains why Clarke's point is right: abilities make a difference in a way that dispositions do not.

(By contrast, we could think of fragility as a *mere* disposition). Then one would show that this additional activity within the disposition is sufficient to close the conceptual gap. In what follows, I will argue that this project fails. Reductive analyses of the free will ability either make it such that an agent's actions are not settled by the agent herself, or they presuppose that agents have abilities of the sort which are plausible candidates themselves for free will. These analyses therefore fail. Since they fail, they cannot be put to work in favor of compatibilism.

In light of this, one could give up the project of reductive analysis but try to retain the useful element of the reduction strategy, namely, the appeal to dispositions. One might use whatever dispositions are required to exercise the free will ability as a kind of forensic evidence for the existence and nature of the ability. Critically, this evidence must be taken to show that the ability is compatible with determinism. I argue that this option, which amounts to acknowledging and trying to live with the gap between dispositions and abilities, offers a poor answer to the compatibility question. Dispositions do not exhaustively explain abilities. This leaves it an open question whether or not the freedom grounding elements of the ability to respond to reasons are to be had in the dispositions necessary to have and exercise that ability. Thus, we cannot conclude that some ability is compatible with determinism just from the fact that the dispositions required to exercise it are. The essential freedom grounding elements of the ability may not be (and incompatibilists should assert that they are not).

### **1.3 First Horn: Closing the Gap**

The first horn of the dilemma problem at hand—developing a successful reductive view of abilities— itself leads to a dilemma. There are two sorts of reductive accounts that might close the gap between abilities and dispositions by appeal to active dispositions. One could offer an analysis of the free will ability as the ability to respond to reasons in terms of active dispositions where the activity of the agent is the *manifestation* of the disposition. (E.g., you could be disposed to *choose* on the basis of the reasons

you recognize). Or, one could offer an analysis where this activity is located in the *stimulus conditions* of the disposition. (E.g., you could be disposed to respond aptly to reason *when you try to*). But neither option is satisfactory.

Consider the first way of proceeding. For instance, Fischer and Ravizza (1998: 62) seem to offer a view where agentive activity is located in the manifestation of the pertinent dispositions. They work in terms of dispositions to recognize reasons and to choose in accordance with those reasons.<sup>19</sup> This sort of view seems to animate the presumptive case in favor of reasons-responsive compatibilism, for these dispositions seem to capture agentive activity and are apparently compatible with determinism. One might be tempted to say that a view like this just shows that there is no gap problem at all. The view says there is no gap. There is nothing leftover or left out when we reduce.

Although at first it might seem that the gap problem disappears on such an account, I worry that it just buries the issue. Dispositions manifest when the stimulus condition occurs, willy-nilly. This seems to be conceptually ill-suited to explain abilities insofar as the stimulus conditions prompt the manifestation of the disposition. The problem is that dispositions seem unable to highlight a sense of control moving through the processes of stimulus condition to manifestation. Consider dispositions to recognize reasons and choose in accordance with those reasons. First, we do not settle what reasons we recognize as such. And given this, our choices seem not to be settled by us either—that is, unless we have an ability to choose in response to our reasons.

Isn't this a stretch? It might seem to some that the gap problem boils down to the idea that the stimulus conditions of a disposition are something done to the thing which has said disposition (like *dropping* a fragile glass). Dispositions seem to involve a kind of passivity on the part of the

---

<sup>19</sup> They talk in terms of a 'cognitive' power to recognize reasons and an 'executive' power to choose. Both are general dispositions of an agent's mechanism of action rather than abilities of the agent. This executive power is explicitly cast in terms of reacting to incentives recognized by the cognitive power (1998: 75). The successful manifestation of the cognitive disposition is the stimulus for the executive one.

disposed. Actions, on the other hand, originate from the agent. If one thought that this was the heart of the gap problem, one might reply like this. It is important that the agent's reasons are the stimulus conditions for my choosing. Why? They are *internal* to the agent (cf. Clarke (2015: 898), Vihvelin (2013: 172)). It is not as if something outside of the agent is pushing her around! This makes an active disposition to respond to reasons rather unlike a disposition like fragility. Indeed, these active dispositions seem to be tracking an internal locus of control on the part of the agent.

Fair enough. The gap problem is neither about the external conditions of action nor about the agent's lack of activity, however. The problem as I understand it is about explaining the right kind of activity, the sort difference-making which characterizes the freedom-grounding aspect of abilities. The free will ability is typically understood to be an ability to perform a basic action, an action we can take without doing something else. These are typically understood to be abilities to choose, decide, or try. Hence, we can be directly in control of and responsible for the performance of basic actions like choosing, deciding, or trying.<sup>20</sup> But which reasons I recognize is not a matter of my taking a basic action by deciding, choosing, or trying. (At least not directly). So, nothing about the stimulus condition of my dispositions to choose in response to the reasons I recognize is a matter of my taking a basic action. And if my choosing in accordance with the reasons I recognize is the (willy-nilly) manifestation of a disposition, then it looks like I do not settle the matter of what I do. An agent's reacting to the reasons which she recognizes cannot be a *mere* reaction to background conditions.<sup>21</sup>

To close the gap, we need the right sort of activity to feature in the pertinent dispositions. Otherwise, they will be mere dispositions. Plausibly, though, the right kind of activity, the kind of activity that would ground an agent's freedom, is just the exercise of an ability in response to her

---

<sup>20</sup> For instance, Pereboom (2001: xxi) suggests that freedom and responsibility apply to decisions.

<sup>21</sup> Switching to an agent-based reasons-responsive theory (e.g. McKenna 2013) won't resolve this problem. If we characterize the agent's reasons-responsiveness in terms of dispositions, then we seem committed to her (willy-nilly) responding to a condition which was not settled by herself.

reasons, like an ability to choose. Exercises of abilities do not happen willy-nilly. They do not happen whether one likes it or not. So now it looks like we will have to invoke the exercise of an ability as the manifestation of the disposition to respond to reasons. If so, then these dispositions cannot be a suitable reduction base for the pertinent abilities. They involve abilities! Indeed, if we have said nothing about the nature of these abilities to choose, then perhaps they require the falsity of determinism. If so, then we cannot marshal this kind of reductive analysis as part of the case in favor of compatibilism.

Reductive reasons-responsive compatibilists will have to try something different. Let's consider the other kind of account that a reasons-responsive compatibilist might offer, which reduces abilities to active dispositions where the pertinent activity is in the stimulus condition of the disposition.

I will use Vihvelin's (2013) view as an example. Vihvelin sees herself as offering an account of active dispositions to act in intelligent and goal-directed ways. Indeed, she distinguishes between the pertinent active dispositions and mere dispositions like fragility, solubility, and the like (2013: 178). She maintains that the free will ability is the narrow ability to choose on the basis of reasons. Extrapolating from her analysis of abilities as dispositions, this narrow ability consists in being disposed to choose on the basis of reasons. These choices need occur only in a suitable proportion of cases where one has an opportunity to choose on the basis of reasons, has some intrinsic causal basis for this ability, and tries to choose on the basis of reasons (2013: 187). This seems to capture the right kind of activity. Why? The 'tries to choose' condition of the analysis here is important. For it allows the agent to settle the stimulus condition of the disposition to respond aptly to reasons. This gets around the problem for views that only locate the relevant agentive activity in the manifestation condition.

You will notice that this is a sophisticated version of the failed classical compatibilist conditional analysis of abilities cast in terms of dispositions rather than counterfactuals. Recall that

such views were extensionally inadequate and so couldn't motivate compatibilism. Vihvelin rightly has a more modest aim in mind. Her dispositional account is meant to 'fill in the details' of a compatibilist conception of what abilities ground the facts of freedom within a commonsensical picture of free agency (2013: 214).<sup>22</sup> Dispositions are supposed to explain *why* the free will ability is compatible with determinism not show *that* it is so.<sup>23</sup>

Still, if the modest aim of offering a compatibilist-friendly analysis of abilities fails by way of the gap problem, *a fortiori* so too does the less modest aim of employing such an analysis as part of the case for compatibilism.

This modern, modest take on the classic strategy won't work, unfortunately. It faces the same difficulties as its forerunner. It turns out that these difficulties are in fact explained by the gap problem.

Chisholm (1964) articulated the following challenge to the classical conditional analysis. The classical analysis is supposed to tell me whether or not I could do otherwise. It says I can do otherwise when, if I had tried to do otherwise, I would have. But a trying to do otherwise is just an instance of doing otherwise, because a trying is itself a doing. It is a particular kind of doing, a mental action, and a basic action at that. Trying is plausibly an action we can take without doing something else. Hence, it is the sort of thing we can have direct control over and be directly responsible for. The ability to try is therefore, plausibly, an ability to perform a basic action. This is the sort of ability which might *be* the free will ability. And I may or may not be able to try. To parrot Davidson's (1973: 114) understanding of the argument: the antecedent of the pertinent conditional must not contain any verb which makes sense of the question: can someone do *it*?

---

<sup>22</sup> Vihvelin does not offer an analysis in the sense of giving necessary and sufficient conditions. Rather, she thinks of herself as offering an analysis vis-à-vis plausible ontological reduction as a kind of research program for compatibilists (2013: 170).

<sup>23</sup> In spite of my criticism, I am sympathetic to Vihvelin's defense of commonsense, agent-causal compatibilism. She rightly claims that the issue of determinism is orthogonal to what ordinary people mean when they say that they have abilities and opportunities. I only worry about her treating abilities like dispositions.

In order for the analysis to be informative, then, it must assume that I am able to try. That I am able to try is exactly the kind of question we ought to be answering in the free will debate, and unfortunately, it has gone unexplained. Indeed, the counterexamples to the classical analysis are just cases where an agent cannot try to do something, but nevertheless, if she *were* to try to do it, she would (e.g. Lehrer (1968)). Thus, the agents in the counterexamples appear unable to perform the action in question.

In other words: the classical analysis appeared informative *only because* it assumed that an agent could perform a basic action, a trying or the like. There are, after all, many actions we can take only because we try. However, the classical analysis offers no answer to the question: is the agent *able* to try? The analysis therefore does not rule out that indeterminism is part of what it takes to be able to try. It is simply silent on the matter.

The same challenge faces dispositional accounts of abilities that involve basic actions in the stimulus condition. Recall that (simplifying a good deal) Vihvelin's analysis reduces the ability to respond aptly to reasons to a disposition to respond aptly to reasons when one tries to choose on the basis of reasons. Analyses of the free will ability are supposed to tell us what it means when we say I am able to try to do things. They cannot presuppose that I am able to try in order to be informative. And it looks like Vihvelin's take on the classical strategy does so. It doesn't tell us what it means for an agent to be able to try to choose. In her case, the problem is also related to a purported extensional failure. Clarke (2015: 896) has suggested that Vihvelin's analysis fails by counterexample in cases where the intrinsic causal basis of an ability will be lost were the agent to try to exercise it.

Perhaps these objections are mistaken. The agent in the counterexamples cannot try to do the target action. Nevertheless, there is a sense in which she is able to perform the action. Why? If she were to try to, she would succeed, just as this kind of reductive analysis suggests (cf. Vihvelin (2013: 201-203)).

I remain unconvinced. Reflecting on the gap problem clarifies that the sense in which the agent in the counterexamples remains able to do things is not the pertinent sense. The reductive analyses leave open the possibility that the exercise of some of an agent's abilities (that is, that they are exercised) could fail to be settled by the agent herself in virtue of her trying, deciding, choosing, or intending to exercise them. This is just a failure to explain the aspect of agentic abilities that seem to ground an agent's freedom and responsibility. To exercise this sort of ability is to make a difference by settling that some action occurs given the opportunity to do so.

In general, a proponent of this sort of analysis—the sort where we take agentic activity to the stimulus condition of the disposition to act—has two options when it comes to basic actions like trying. Let's construe a trying as the acquisition of an intention or desire to do something. Well, either the agent settles that she acquires an intention or desire, or she doesn't.<sup>24</sup> In the first case, the proponent of the analysis cannot invoke an unexplained ability to try (or choose, or decide) to try, in virtue of which the agent settles that she tries. And if the proponent explains this ability to try to try as a disposition to try when one tries to, then we can simply reiterate the problem. When is an agent able to try to try? On the other hand, if the proponent instead argues that a trying can initiate from a process which the agent herself does not settle, then I say this process is not control-implicating. The trying isn't settled by the agent herself. (An agent doesn't settle that she tries when a trying is caused by an overwhelming and unendorsed desire, for instance).

We should conclude that neither kind of reductive analysis cannot plausibly explain what the pertinent reasons-responsive abilities are. Putting the agentic activity in the manifestation condition seems to either entail that an agent can exercise the pertinent ability or makes it such that an agent

---

<sup>24</sup> Vihvelin offers a defense of both options. She can consistently do so because she holds the view that a process counts as a trying if it causally leads to the beginning of an action. Such a process may or may not be initiated by an agent's trying to try (2013: 176-180).

does not settle whether or not the pertinent ability is exercised. That won't do as a reductive base. On the other hand, putting agentic activity into the stimulus condition either presupposes that agents can perform abilities of the sort which are themselves candidates for being free will or, again, makes it such that an agent does not settle whether or not the pertinent ability is exercised. Given this, the view which reduces the ability to respond aptly to reasons to a kind of active dispositions cannot offer a plausible view of how free will could be compatible with determinism.

#### 1.4 Second Horn: Living with The Gap

Luckily for compatibilists, there is a compelling alternative to the reductive strategy. Consider Wolf's view in *Freedom within Reason* (1990). On her view, the ability by which we have met the control condition for morally responsible (right) action is the ability to act in accordance with Reason, and so be guided by the True and the Good. Wolf then suggests a test to check if an agent has this ability. To have some ability is to have the necessary capacities, skills, and talents to exercise that ability (ibid. 101-103). To lack these necessary components is to lack the ability. So, to know if an agent had control sufficient for moral responsibility, we check to see if an agent has the capacities, skills, and talents (i.e., dispositions) to be appropriately reasons-responsive and whether or not these capacities (etc.) have been interfered with or hampered. Indeed, Wolf rejects the need for a reductive analysis. She instead supplies a "*characterization* of what is involved in attributing an ability to someone" that purports to show the "*per se* irrelevance" of determinism to an agent's having abilities (1990: 101).

Here is the general strategy: advance a partial conception of the free will ability, argue that this conception is sufficient for theorizing about moral responsibility, and then show that there is no compelling reason to think that it is incompatible with determinism. On this way of advancing compatibilism, the reasons-responsive ability is not identified with dispositions to respond to reasons; instead, the relevant counterfactuals and dispositions provide forensic evidence for the presence of an agent's reasons-responsive ability. Moreover, the pattern of counterfactual response and the presence

of dispositions provide us with information about what this ability is like, and so helps explain how some agent is able to act as she does (or how she would act otherwise). This information might also help us understand when and why an agent is unable to act (or would be unable to act differently) in some particular situation; for instance, a disposition required to exercise some ability might be impeded in some manner. This provides an explanation of how the counterfactuals and dispositions relevantly relate to our abilities. In other words, we have good evidence about what the free will ability is and whether or not it is compatible with determinism without the need to posit a risky analysis. The view avoids both the problem of extensional inadequacy since it holds fixed a plausible non-reductive conception of the free will ability. Better still, it has learned to stop worrying and love the conceptual gap between dispositions and abilities. Recent defenders of this “evidentialist” strategy include Nelkin (2011) and McKenna (2019).<sup>25</sup>

For example, I will focus on Nelkin (2011), who has offered a compelling and sophisticated version of this strategy. She argues for a compatibilist version of reasons-responsive agent-causation. Building off Wolf, an important feature of her view is that it promotes an intentionally ambiguous conception of ability: an interference-free ability (2011: 67-68). For an ability to be interfered with, an agent must have the *general* ability to do something—she has to be constituted in a certain way so as to have the power to do such-and-such. But to have an interference-free ability need not require of that agent that she have a *specific* ability—to be in fully favorable circumstances for exercising her general ability.<sup>26</sup> Interference-free abilities require the unimpeded manifestation of capacities, dispositions, and talents. They need not require *fully* favorable circumstance for exercising one’s abilities, however; for instance, one might be in the presence of a counterfactual intervener (2011: 67-

---

<sup>25</sup> McKenna calls the reductive enterprise “conservative” since it is ontologically conservative, whereas the “evidentialist” strategy is “liberal”, involving a more permissive idea of what is required of a compatibilist theory of free will.

<sup>26</sup> This distinction between specific/general abilities is structurally similar to the distinction between wide/narrow abilities discussed in §3.

68). Nelkin offers reasons to think that this sense of ability is sufficient for theorizing about moral responsibility. Grant her this point for sake of argument.

Let's consider why one billiard ball moves in a certain direction when another one strikes it. On Nelkin's view, "billiard balls...exert certain causal powers in their circumstances when struck by a pool cue, and it can be inevitable what their effects are. Given their own natures and causal powers, and the natures of and causal laws governing everything else in the vicinity, their effects are determined" (2011: 89). If billiard balls can be causes in virtue of their causal powers, why can't agents? She continues, "agents, while having unique natures and causal powers, such as being able to act on certain sorts of reasons, can, at least in principle, exercise those causal powers even when, given their natures and circumstances, it is determined how they will act" (2011: 89-90). What explains the pattern of movement of a billiard ball, or the actions of an agent for reasons, are the causal powers in play.

Notice that Nelkin compares the causal powers of physical objects with the causal powers of agents. First, let's say that objects are causes in their own right as substances in virtue of their causal powers and natures. These powers are (presumptively) unproblematic given determinism in the following sense: determinism does not interfere with their operation since they are grounded in the objects themselves. The causal powers of agents and the causal powers of objects appear to be of the same type. Ultimately, they are both causal powers. So, we should conclude by analogy that the causal powers of agents are unproblematic given determinism, as determinism does not interfere with them in a likewise manner. This is an appealing picture. Abilities are (at least) powers of difference-making. Grant Nelkin that even ordinary objects can be genuine difference-makers, even if determinism is true. If a billiard ball has this power, surely, we do too! Thus, Nelkin endorses a kind of rationalist agent-causal compatibilism. We are free because we cause, in virtue of our both rational nature and our causal powers, actions for reasons.

I think we should concede to Nelkin that determinism does not obviously interfere with all causal powers; however, in her descriptions, we find the language of “exercise” in her discussion of an agent’s causal powers (2011: 89-90). This language is absent in her description of the causal powers of objects. Here one might argue that the causal powers of agents and the causal powers of objects are dissimilar. The causal powers of objects are, if anything, dispositions. The causal powers of agents are abilities. The conceptual gap between them remains, even if we say that ordinary objects have powers to make differences, for these powers do not amount to anything like the sort of control that characterizes morally responsible agents. Objects are not *origins* of difference-making the way that free agents are. Agents exercise active control over what they make happen. Therefore, Nelkin’s view can be met with criticism: what is this ability, and what accounts for our being able to exercise it, over and above our having the requisite capacities and the like? Without a full answer to this question, we cannot know if the ability is compatible with determinism. Nelkin could be right that the exercise of an ability to respond to reasons requires unimpeded dispositions. This does not resolve the question of whether this is all that the exercise of this ability requires, however, and it might not capture the heart of what makes reasons-responsiveness a plausible candidate for the free will ability. What has the reasons-responsive theorist said to rule out that the ability to respond aptly to reasons is incompatible with the thesis of determinism? Not enough. What makes the ability to respond aptly to reasons freedom-conferring may be something beyond the dispositions needed to have and exercise it (for as I argued above, they do not seem adequate to ground freedom on their own). And this “evidentialist” compatibilist reasons-responsive strategy under consideration has nothing more to offer by way of explaining what this extra thing would be.

Now, one might want to say here that the incompatibilist’s offer of indeterminism as that extra special element is of no help either. Indeed, Nelkin argues as much (2011: 75). I agree. My aim here is not to add to the dialectical burdens we compatibilists already shoulder. Nevertheless, even if no one

has a worthy proposal for what the needed additional element is, it remains true that the evidentialist compatibilist's strategy is incomplete.

To put the challenge succinctly: the modal properties necessary for having an unimpeded ability are not necessary *and sufficient* for having that ability and exercising it in a deterministic world. That the dispositions needed to exercise our freedom-relevant abilities are a kind of forensic evidence about the abilities does not resolve the relevant questions. It is quite possible that some further element needed for having and exercising that ability is not compatible with determinism. We have no evidence that dispositions exhaustively capture the essential freedom grounding elements to be had in this kind of ability. Since the view offers no such evidence, it offers no full answer to the compatibility question.

### 1.5 An Objection Considered and The Problem Generalized

Skeptical readers will have been worried that the gap problem begs the question against compatibilism, and after setting out the arguments against the reductive and non-reductive views in question, I can now answer their skepticism.

Vihvelin herself gives voice to this worry in a particularly clear way. She suggests that van Inwagen appears committed to the claim that 'the concept of a disposition is the concept of something that is compatible with determinism whereas the concept of an agent's ability or power to act is the concept of something that is incompatible with determinism' when he distinguishes abilities from dispositions (2014: 173). If this were so, then van Inwagen plainly begs the question. Perhaps I do too.

I don't, though. I have defended no principle which entails the falsity of compatibilism. Perhaps it looked like I begged the question since an upshot of my view is that an adequate account of the free will ability looks more friendly to libertarians than compatibilists have generally recognized. Abilities make a difference in a special way. And libertarians face comparatively less pressure to offer a reductive view of this difference-making than us compatibilists do. Nevertheless, there are possible libertarian reductive accounts of the free will ability in terms of *indeterministic* dispositions. And the gap

problem is in principle generalizable to these views. This shows that the gap problem does not beg the question.

To see this, let's consider an incompatibilist dispositional view of the free will ability. Indeterministic dispositions are those dispositions whose manifestation may but needn't occur given the presence of the stimulus condition and the absence of intervening factors (cf. Clarke (2009: 326, ft. 3)). A libertarian could be motivated to adopt a view of the free will ability in terms of these indeterministic dispositions because libertarian causal powers seem mysterious. Indeterministic dispositions are, perhaps, comparatively less mysterious, and so seem to offer a nice way of filling in the details. Such a view faces the gap problem too.

There is, after all, a difference between the radioactive atom that has an indeterministic disposition to decay or not decay, and the ability of an agent to speak French or not. For there is surely no sense of control on the part of the atom in whether or not it will decay, whereas one elects if and when to speak. And this is because the stimulus condition of the atom's decay does not provide an opportunity. The stimulus could trigger a reaction, or it could not. Conversely, if an agent exercises her ability to speak French, she had an opportunity to do so. Moreover, *she* settles that she speaks. If you are a libertarian, you may think she settles this *because* nothing antecedently determined her to do so. (Since I am a compatibilist I disagree, but let's set that aside).

Can our libertarian close this gap between indeterministic dispositions and abilities? No. Let's say that she offers a simple view like this: the free will ability is the indeterministic disposition to choose in response to reasons. The problem here is that she does not settle the reasons she recognizes. And her choice cannot be a mere reaction to her reasons. A reaction is still a reaction, even when it is not determined to happen given prior conditions. If she does settle her choice in response to reasons, then, it seems to be because she exercised some ability.

What if the libertarian says instead that the free will ability is the indeterministic disposition to choose when one tries on the basis of reasons? Now we have run into the other problem. Aren't we trying to explain abilities to perform basic actions like trying? We can't presuppose an unexplained ability to try.

Let's now imagine a libertarian who, in light of this problem has a different view. She says that indeterministic dispositions are required to have and exercise the ability to respond to reasons. (Maybe the brain is indeterministically disposed to do certain things, say). She takes this to be a kind of forensic evidence about the presence and nature of the free will ability, namely, that it is incompatible with determinism. In other words, she does not offer a reductive account of the ability in terms of these dispositions. Now the problem is something like the inverse for Nelkin's compatibilist view: adding indeterminism to the dispositions does not settle the pertinent question. What makes the ability to respond aptly to reasons freedom-conferring may be something beyond the dispositions needed to have and exercise it. Indeed, as Nelkin herself notes, indeterminism doesn't seem like the special element needed to explain freedom-grounding element of abilities. Why would it help when added to the dispositions needed to exercise them?

The gap problem, therefore, does not beg the question against the compatibilist. It is a general problem in the metaphysics of abilities, albeit one with profound dialectical consequences for compatibilism.

## **1.6 Further Generalizing**

So far, I have argued that a mainstream compatibilist strategy, which identifies the free will ability with the ability to respond aptly to reasons, is at best incomplete. Reductive analyses of the free will ability in terms of dispositions fail. On the other hand, using dispositions as a kind of forensic evidence to indicate the presence of the free will ability does not answer the central question: is the free will ability

compatible with determinism? And this argument does not beg the question. It is in principle generalizable to corresponding incompatibilist views.

In this section, I want to briefly point out that the problem I raise appears general in a different way. None of the foregoing entails that other compatibilist projects, ones that are not based around dispositions, will fail. The problem, though, is that for any given compatibilist project a new gap problem seems to appear. And a “reductive” and an “evidentialist” way of responding to that gap naturally suggest itself. The essential problem of each strategy seems to remain even when we change the metaphysics.

Consider variations on a compatibilist project based on *causation* and not dispositions. One way to secure the compatibility of some agency-relevant ability and determinism is to reduce that ability to patterns of causation. Alfred Mele’s compatibilist proposals could be construed in this manner. In his (1996: 193), free action critically involves the nondeviant production of action on the basis of rationally formed deliberative judgment. “On the basis of” seems not uncharitably construed as a causal notion. (In any event, one might neatly adapt his proposals in this manner).<sup>27</sup> Or, on the other hand, some further piece of evidence could help settle the evidence in favor of compatibilism. Again, causation appears useful. Perhaps the right sort of causal pattern in the actual sequence of events is evidence of a freedom-relevant ability. Since causation occurs in deterministic worlds, the pertinent causal pattern can be taken as evidence that the freedom-relevant ability is compatible with determinism too.

Now for the problems. The worry for the reductive strategy is that it failed to account for the freedom-grounding element in our concept of a free agent’s ability. There is a “space” between what happens to us and what we choose to do within which agents can make a difference by settling what

---

<sup>27</sup> See Mele (2006: 170) for another example.

comes next. Now, if our abilities are to be successfully reduced to causal relations, then I fear we will once again lose the freedom-grounding element in our concept. For there is no “space” between causes and effects; causes are not opportunities for effects to arise. A compelling way to get around this new gap would be to reduce abilities to patterns of causation, and then (perhaps) reduce causation to pertinent counterfactuals (cf. Lewis 1973b). Here, we would face the same problems discussed in §3. A plausible way of developing this idea would be to put the agent’s activity in either the causes or the effects (or both). If in the effects, then it is no longer clear that it is the agent which settles what she is going to do. (This is an inversion of the luck problem for event-causal libertarians).<sup>28</sup> If in the causes, then just like putting activity in the stimulus condition of a disposition, we risk presupposing what we aim to explain or failing to explain it altogether.

On the other hand, causation *per se*, even causation by reason, does not seem sufficient evidence that our freedom-grounding abilities are compatible with determinism. The causal theory of action cannot decide an answer to the traditional problem all on its own. (Hence, for instance, Nelkin’s appeal to the causal *powers* of agents). So, the causal “evidentialist” alternative looks incomplete.

Perhaps the focus on abilities is the real problem for the compatibilist. Maybe this commitment could be given up. Carolina Sartorio (2016)’s argues that the freedom of an agent is grounded in the reasons, and crucially the absence of other reasons, given a suitably wide range of reasons, that are the causes of her action in the actual sequence of events. She goes on to argue that there is “no difference in freedom without a difference in the relevant elements of the causal sequence” (2016: 32). Notice that there has been no mention of abilities so far. Her view can be developed in an ability-neutral way. One could forego an abilities-based approach altogether and explain freedom

---

<sup>28</sup> The luck argument against event-causal libertarianism, alleges (to greatly oversimplify) that if an agent’s actions are not determined by prior causes, then her performing said action is (at least in part) a matter of luck, and so not free. See Mele (2006) for a version of this argument and an extended discussion on how libertarians could deal with it.

directly in terms of causation by reasons alone. Maybe the relevant abilities supervene on the actual sequence, but these are not explanatorily basic. Perhaps other ability-neutral views are possible, or perhaps some of the views discussed above could be modified in ability-neutral ways.

For those of us who would like to respond to the traditional problem of free will, of reconciling some special freedom-grounding ability with the thesis of physical determinism, this may be asking too much. A successful account of freedom does seem at least a little beholden to the traditional (and intuitive) view, that if we are free and responsible, it is because we have some special ability characteristic of free agents. Of course, this is only one metric with which to test a theory. Compatibilists should consider the ability-neutral alternative carefully.

## **1.7 Conclusion**

Abilities-based compatibilism faces a serious challenge in arguing that the free will ability is compatible with determinism. This paper is a warning to fellow compatibilists. We face a gap between our intuitive ideas about abilities and the metaphysical notions, compatible with determinism, that we might appeal to in explaining the nature of these abilities. This problem that has not been sufficiently recognized, and so has not been aptly responded to.

## Chapter 2

### Responsibility and the Limits of Good and Evil<sup>29</sup>

**Abstract:** P.F. Strawson's compatibilism has had considerable influence. However, as Watson has argued in "Responsibility and the Limits of Evil", his view appears to have a disturbing consequence: extreme evil exempts an agent from moral responsibility. This is a *reductio* of the view. Moreover, in some cases our emotional reaction to an evildoer's history clashes with our emotional expressions of blame. Anyone's actions can be explained by his or her history, however, and thereby can conflict with our present blame. Additionally, we too might have been evil if our history had been like the unlucky evildoer's. Thus, our emotional responses to the evildoer compromise our standing to blame them. Since Strawson's view demarcates moral responsibility by moral emotional responses, his view appears to be self-defeating. In this paper, I defend the Strawsonian view from the *reductio* and self-defeat problems. I argue that two emotions, disgust and elevation, can be moral reactive attitudes in Strawson's sense. First, moral disgust expresses neither blame nor exemption from responsibility. Instead, moral disgust presupposes blameworthiness but is instead a distinct response to the extreme wrongdoer. Secondly, moral disgust involves self-directed attitudes that explain away our apparent lack of standing to blame the evil agent. The structure of disgust as a reactive attitude is mirrored along the positive dimension by the emotion that Jonathan Haidt has called "elevation", a feeling of moral inspiration. I conclude by defending my view from objections about the moral appropriateness of disgust.

#### 2.1 Introduction

P.F. Strawson's compatibilism about free will and his attendant account of moral responsibility have had a considerable influence on contemporary work.<sup>30</sup> This is so even among those who do not count themselves as followers of Strawson's overall compatibilist program. Even opponents address Strawson's main themes: the moral emotions and our practices of interpersonal expectations and demands.<sup>31</sup> Despite this influence, there are two damning problems for Strawson's theory. First, it appears to restrict the class of morally responsible agents to those who are in fact members of the moral community. However, as Gary Watson has argued in "Responsibility and the Limits of Evil" (2008/1987), this has a disturbing consequence: extreme evil *exempts* an agent from moral responsibility. If so, why shouldn't we view this as a *reductio*? Second, in some cases our emotional

---

<sup>29</sup> Published in *Philosophical Studies*, 176 (10): 2705-2727. (2019). Reprinted here with permission from Springer publishing.

<sup>30</sup> For example, see Bennett (1980), Darwall (2006), Fischer and Ravizza (1998), McKenna (1998, 2012), Russell (1992), (2004), Shoemaker (2015), Wallace (1994), Watson (1987), and Wolf (1981).

<sup>31</sup> For example, see Pereboom (2001), (2014) Smilansky (2000), and G. Strawson (1986).

reaction to an evildoer's history clashes with our emotional expressions of blame. We might then worry that anyone's actions can be explained by his or her history, and thereby, conflict with our present blame. We might also worry that had our history been like the unfortunate evildoer, we too might have been evil. This compromises our standing to blame. Together, this would undermine the expression of blame, generally, and so be self-defeating. It is critical to the Strawsonian project to respond to these problems. I will do so in a novel way. As I see it, we have an impoverished vocabulary of reactive attitudes. Expanding our vocabulary can dissolve the problems. To do so, I will revisit the example of extreme evil discussed by Watson, which motivates the challenges, and then provide a contrasting example of my own, one of extreme good. By examining these cases at two different ends of a spectrum, I will identify a peculiar pair of reactive attitudes. Relying upon them, I will then provide an unorthodox defense of Strawsonian compatibilism.

## 2.2 The Basics of Strawsonianism

Strawson's theory is about moral responsibility in the accountability sense of being on the hook for one's actions, so to speak.<sup>32</sup> It has three components. First, Strawson thinks that the practices of holding someone morally responsible are expressed by and consist of emotional *reactive attitudes*. Some emotions—resentment and indignation—are responsive to the intentions of other persons in a way that many emotions are not. We might get upset if our car is a lemon, but we couldn't be genuinely indignant *at* the car! We get indignant *at*, say, a co-worker when we find out that he or she has unfairly spoken ill of our other colleagues.

When holding another responsible by way of a reactive attitude like resentment, we are reacting to whether or not someone has shown us due consideration in their intended actions. This is the

---

<sup>32</sup> This kind of moral responsibility is opposed to moral responsibility in other senses, like moral assessment of character (attributability responsibility) or moral assessment of judgement-sensitive attitudes (answerability responsibility). For an extended discussion see Shoemaker (2015).

second component: moral responsibility tracks the *quality of will*—the good or ill will—with which a person acts. “Quality of will” is best understood as referring to the concern or lack thereof that a person shows to others, especially as manifested in her actions. This concern is constituted by an agent’s attitudes and intentions (Strawson 2008/1962; McKenna 2012; 59-60; Shoemaker 2015). In turn, in being disposed to the pertinent reactive attitudes, we *interpret* agents’ actions as expressing these quality-of-will constituting attitudes and intentions.

Why focus on interpretation? Sometimes we interpret each other’s actions incorrectly. Hence, we have need for excuses, justifications, and exemptions. We could be *excused* from being held responsible—“I was shoved and fell into you!” Alternatively, we could provide *justification* for our actions—“if I didn’t shove you that car would have hit you!” In either case, we are invited to see a different quality of will behind the (putative) injury. Exemptions show that someone is not the proper target of the reactive attitudes at all. An exempted agent is “*incapacitated* in some or all respects for ordinary inter-personal relations” (Strawson 2008/1962; 25). The exempted agent is seen through what Strawson calls objective attitudes. He or she has become an object that demands explanation or requires management, even if temporarily. This person is not expected to meet the demands of morality. Examples might include very young children or the seriously mentally ill.

Here is the third component of the view: there is nothing more to being morally responsible than being the appropriate target of reactive attitudes as set by our practices of holding one another responsible. Our expression of reactive attitudes signifies the moral demands we place on one another. To lack either the capacity for relevant quality of will or the capacity to engage in interpersonal relations would *dispel* the reactive emotions that express our moral demands and would invite an objective attitude.

We needn’t go into detail about how these components make for compatibilism. Instead, our focus is the capacities needed to be a target of reactive attitudes. This is essential to responsible agents

on the Strawsonian view; however, he does not specify what these capacities are. Building on a proposal first suggested by Gary Watson (1987), contemporary Strawsonians have attempted to advance his view by arguing that reactive attitudes communicate responses to interpretations of quality of will. This provides a way to precisify the requisite capacities. Call the following the argument from communication, which is due to the interesting work of Colleen Macnamara: (1) morally responsible agents are eligible candidates for praise and blame. (2) Praise and blame in the form of the reactive attitudes are forms of communication and, in paradigmatic cases, take a praised or blamed agent as an intended addressee. (3) In order to be a candidate for praise and blame, then, one has to be a candidate addressee of praise or blame. Therefore, (4) in order to be a proper candidate for address, one must have the capacity to understand the meaning communicated by the reactive attitudes. What indicates that someone genuinely understands the meaning of a received reactive attitude? Typically, they need an understanding of what an appropriate emotional response to a given reactive attitude is. For instance, someone who knows that blame should be accompanied by guilt and a desire to make amends when faced with indignation is a proper target of address in the form of blame. They need not actually feel guilt.<sup>33</sup>

McKenna's (2012) view of the reactive attitudes shows us that persons with these capacities to understand reactive attitudes constitute a moral community (but importantly McKenna's view is that one could have these capacities *and* be outside the moral community—more on this later). By analogy to the linguistic competence needed to hold an intelligible conversation within a linguistic community, the morally responsible agent is involved in an interpretive enterprise whereby different patterns of conduct come to indicate states of mind, intentions, and so on. An agent embedded in

---

<sup>33</sup> This is an expanded paraphrase of Macnamara (2015)'s excellent characterization of the argument. I have made an amendment. Macnamara thinks that to be the appropriate target of address is to undergo the right sort of response to a reactive attitude. I disagree. People often understand what they should feel but do not concurrently have those feelings. Watson (2008/1987), Shoemaker (2007), and Darwall (2006) each propose a similar argument.

these practices understands *both* their own actions and the actions of others by reference to the set of patterns that have come to indicate good or ill will. The knowledge of one requires the knowledge of the other. Why? In order for the reactive attitudes to work they need to be understood by members of the moral community as expressing demands, for instance. Their expression requires a shared framework of value and practice in order to be intelligible. Pleas to be released from these demands require this shared framework too. So, a moral community is constituted by a set of agents capable of understanding, expressing, and responding to common moral expectations by way of interpersonal emotional reactions. Incapacitation from moral responsibility is therefore the incapacity to participate in a moral community (cf. McKenna 2008a/1998).

With the preceding sketch of Strawson's theory of moral responsibility in place, I turn now to two cases, one really bad and one really good. Using them, I will illustrate the two challenges to the Strawsonian view noted above and also call attention to two reactive emotions on the basis of which I will attempt to defend the view.

### 2.3 The Really Bad Case

It is natural to think that acts of evildoing are appropriately addressed by the reactive attitudes expressive of blame. Consider the following case:

**Harris:** Robert Harris and his brother planned to rob a bank. They hijacked the vehicle of two 16-year-old friends. Robert promised to leave the friends some money in the car for having used it in the robbery. The car ride was amiable. One of the teenagers wished Robert "good luck." As the teenagers left, Robert raised the rifle and shot one of them in the back. He chased the second down a hill, shooting him four times. He knelt over him and shot him in the head. Harris proceeded to laugh. Harris and his brother left the scene and he began to eat the teenager's lunch 15 minutes after the murders. Harris offered his brother some food. He refused—nauseated from the murders. Harris mocked his weakness. He proceeded to suggest that they impersonate police officers and inform the teenagers' families of the deaths. He went on to suggest that they drive near the scene of the murder in order to kill police officers. Harris,

during final preparations for the bank robbery, noticed blood on the rifle and said, “I really blew that guy’s brains out.” He again began to laugh.<sup>34</sup>

How do people respond to Harris? By all appearances, he is clearly blameworthy if anyone is! In prison, he continued to be a man who cared for nothing and no one. His fellow inmates on death row planned to have a party upon his death (Watson 2008/1987: 125).

Yet, as Watson points out, Strawsonianism seems to suggest that Harris is *ineligible* for moral responsibility. How so? Harris expresses unfathomably ill will by showing no consideration for others. He communicates with us. Yet, as Watson points out, “not all communication is dialogue;” Harris repudiates the moral community and places himself outside of it (2008/1987: 128). His being outside the moral community is precisely what makes him so evil. His evil consists in part of the fact that our reactive attitudes towards him are pointless—at best they are met with “icy silence” and at worst with “murderous contempt” (ibid. 128). Harris is not a suitable target for our communicating our reactive attitudes—or so it seems. Due to the structure of the Strawsonian theory, Harris appears to be exempt from morally responsibility. But if he is exempt, it is not because of a lack of interpersonal capacities. If he is exempt, it is because extreme evil is its own kind of exemption, and this is absurd.

I have just gone through the first problem: extreme evil ends up being an exempting condition from moral responsibility. There is another problem. Consider Harris’s childhood:

**Harris, cont’d:** Robert Harris was born prematurely. His father had kicked his mother in the stomach after accusing her of infidelity. Robert spent the first months of his life in an incubator at the hospital. Robert’s father beat and abused all his children and his mother. He sexually assaulted his daughters. He never accepted Robert as a son. Over time, his mother grew to blame Robert for her own abuse and came to hate him. Later, she would say that she felt that Robert’s crimes were her fault. She was never able to love him. His sister reported that Robert was starved for attention. He would seek out his mother’s touch, only to be kicked away. She went on to report that he was the most sensitive of all 10 siblings. As a child, he cried when Bambi’s mother was shot when watching the eponymous film. Robert suffered from learning disabilities and was teased at school. At age 14, Robert was sent to a youth

---

<sup>34</sup> This is paraphrased from Miles Corwin’s *Los Angeles Times* article, “Icy Killer’s Life Steeped in Violence,” as quoted in Watson (2008/1987)

detention center for stealing a car. He was raped several times and attempted suicide twice.<sup>35</sup>

The process from heartbroken child to heartless murderer was brutal. Now it no longer seems clear that it is appropriate to blame Harris, at least not without significant attenuation. Watson describes our reaction to Harris as a mixture of sympathy for the child and antipathy towards the man, leading to ambivalence. This clash is accompanied by the unsettling thought that “one’s moral self is such a fragile thing. One tends to think of one’s moral sensibilities as going deeper than that (although it is not clear what this means). This thought induces not only an ontological shudder, but a sense of equality with the other: I too am a potential evildoer.” (2008/1987: 132) Here is the second problem: Harris was not responsible for his history. No one is—at least not early on. Watson contends, rightly, that we are often ignorant of the historical considerations that shape us. He also discusses possible ways to understand the sensitivity of our reactive attitudes to historical considerations within the Strawsonian paradigm.<sup>36</sup> However, the fact remains that when one focuses his or her attention on histories, one comes to feel like “it is not *one’s* business to blame” (ibid. 137). Faced with Harris’s history, we might think to ourselves, “Who am I to judge? It was only luck that kept me from being as evil as Harris. I was spared such a cruel childhood. If I had been in his position, I might have been no different.” And what is true for you is true for anyone. No one is in a position to blame if we are subject to such deep moral luck. If a history can make us feel like no one is in a position to blame, how can we maintain a commitment that anyone is blameworthy? For a Strawsonian it is impossible to pull apart the propriety of reactive attitudes from responsibility because the theory says that *to be responsible* is to be the appropriate target of those attitudes. If no one can appropriately blame by

---

<sup>35</sup> Again, this is paraphrased from Miles Corwin’s *Los Angeles Times* article, “Icy Killer’s Life Steeped in Violence,” as quoted in Watson (2008/1987)

<sup>36</sup> What is at issue here is whether or not responsibility requires an “ultimacy” condition, that one be the sole originator of oneself. For the details of Watson (2008/1987)’s suggestions see pg. 133-134 and pg. 137. For a further compatibilist discussion of the ultimacy conditions see McKenna (2008b)

addressing other persons with reactive attitudes, then it looks like no one can be blameworthy. The theory appears to be self-defeating.

## 2.4 A Possible Solution

Consider the following defense of the Strawsonian theory offered by McKenna (2008a/1998). First, we can get the right result with a minimal revision of the basic theory. So long as someone has the capacity to participate in the moral community, then he or she is the appropriate target of reactive attitudes. *Pace* Watson, it is clear that Harris does understand the values of our community. He must in order to mock and repudiate them. It follows that he has the capacity to participate in the moral community in a way that, for instance, a frightening animal does not. So we should hold him blameworthy. Second, the skeptical force of Harris's "ontological shudder" is unwarranted. If the skeptical worry arises from the suspension of our reactive attitudes, we need only admit that the appropriateness of those reactive emotions is distinct from what emotions we in fact have. In the Harris case, we can understand that certain negative emotions are appropriate, even when the emotions we actually experience are sympathetic. If the skeptical worry arises out of concerns for moral luck or equality, then you must realize that given who you are now you could not become like Harris. You lack the potential for evil (unless, of course you are like Harris, but that is a different sort of case). Even if we ought to be compassionate in holding Harris accountable, "the moral order" must find its voice through blame (2008/1998, 217). We owe that to his victims.

Could this be the solution the Strawsonian is looking for? I will argue that it is not. Nevertheless, McKenna's response moves us in the direction of a better reply.

Let's assess McKenna's response to the *reductio*: we can hold those outside of the moral community responsible so long as they have a (here and now) capacity to be members. This response strikes me as insufficient. We must conclude that Harris does indeed have the capacity (here and now) to meet the expectations and demands of morality. Yet, contrary to McKenna, I think that it is in some

sense appropriate for us to *stop* engaging Harris with moral demands while maintaining a commitment to his blameworthiness.<sup>37</sup> Why? At least at for some time, he did not care for our indignation. He was unresponsive to moral demands. A perfectly appropriate response to such a person is to *stop trying to get them to respond*. Consider the following case:

**Cheater:** Charles consistently cheats when playing little league baseball. The other little leaguers call him out on his cheating, but Charles does not care. He continues to cheat in spite of this. Eventually, the other children stop trying to penalize Charles. Instead, they just don't let Charles play with them anymore.

By excluding Charles, the little leaguers protect the integrity of the game in a way that is distinct from either acting so as to hold him accountable or exempting him from blame. Our response to the Harris case, and our protection of “the moral order,” is more like **Cheater** than we might initially imagine. To block the *reductio* worry, a reactive attitude that expresses this response is needed: Harris is an appropriate target of moral demands, but it is also appropriate for us to withdraw our efforts to communicate those demands to him.

Regarding the self-defeating problem, I agree with McKenna's distinction between the eliciting of a reactive attitude and the understanding of the appropriateness conditions of having a reactive attitude. It successfully replies to the worry about the presence of sympathetic emotions when faced with Harris's history. However, we do ourselves a disservice by acknowledging that we cannot become like Harris. Still, we would also do ourselves a disservice if, like Watson, we concluded that we could be like Harris. Both responses miss the heart of the “shudder.” Why? Here, my answer is crucial to the view I will advance: *We feel degraded when faced with evil*. This degradation is not directed at any particular feature of oneself—an assumption shared in the disagreement between Watson and

---

<sup>37</sup> On the conversational model I have adopted from McKenna (2012), there is a sense of engagement relevant to the features of one's interlocutor in an instance of moral address. How we decide to engage another person with moral demands ought to depend on what they are like. I will have more to say on this point in the concluding section.

McKenna.<sup>38</sup> Rather, it is directed at our common humanity. For instance, there was no special feature of the soldiers of the Einsatzgruppen death-squads that we might use to explain their participation in genocide. One might wonder, if they were able to become the instruments of genocide, might I? Might anyone? And so, one comes to *understand*, not simply wonder, that *we* are able to become evildoers. This challenges our status as moral beings. It is degrading.

Both Watson's and McKenna's explanations of our reactions to Harris are tempting because they are explanatorily powerful in cases dissimilar to **Cheater**. Consider this example:

**Insult:** Tom and Tanya are coworkers. Tom, for no good reason, calls Tanya a jerk. Tanya, however, knows that Tom has had a particularly bad day at work—he just lost the big account and it was not his fault.

How might Tanya respond to Tom? There are at least two plausible answers:

**Insult, cont'd 1:** Tanya feels antipathy toward Tom for calling her a jerk for no reason. At the same time, she feels sympathy for him, because Tom's unfairly losing the big account was a tough break. She remains conflicted.

**Insult, cont'd 2:** Tanya thinks it would be appropriate to resent Tom for calling her a jerk, but she just can't bring herself to do it. Tom's unfairly losing the big account through was a tough break, and her current emotional state is just sympathetic towards Tom.

I see no *prima facie* reason to favor either of Tanya's possible reactions. Local variation in practices and personal temperament license either response. But **Harris** is not comparable to **Insult**. Tom, we can safely stipulate, would care about Tanya's reactive attitudes. Harris is, like the child in **Cheater**, a target of demands but unmoved by them. What the Strawsonian needs is a different solution to a different kind of case.

---

<sup>38</sup> One could feel degraded for all sorts of particular features of oneself, although here the difference between shame, humiliation, and degradation becomes obscure.

## 2.5 The Really Good Case

McKenna's account draws our attention to the form of a satisfactory solution to Watson's *reductio* and self-defeating problems. Finding this solution will be easier to reach if we reflect on a different kind of example, the opposite of **Cheater**:

**Paragon:** Patricia consistently follows the rules when playing little league baseball. She also encourages other players to do so. She is furthermore a paradigm example of good sportsmanship. Regardless of how well she plays, the other little leaguers want her on their team. When she plays well, Patricia makes sure to credit the team.

There is an attitude that is appropriate for the little leaguers to have towards Patricia that is distinct from both praise and admiration—one of inclusion—that is the opposite response of the children in **Cheater**—one of exclusion. Let me be clear: Patricia is clearly praiseworthy, but she is the appropriate target of a further attitude (more on this later). Moving to serious matters, consider the following case:

**Williams:** On January 13<sup>th</sup>, 1982, Air Florida flight 90 “crashed into the barrier wall of the northbound span of the 14th Street Bridge” between the District of Columbia and Virginia, “and plunged into the ice-covered Potomac River...Four passengers and one crewmember survived the crash.” they were all located in the same section of the downed airliner, a section of the aft cabin that separated during the crash. A fifth passenger was located in this section of the plane, and was able to get out of the plane and into the water with the other survivors. Although they were not far from shore, ice on the river made rescue by boat impossible, and eventually, a rescue helicopter was sent. The rescue rope from the helicopter was dropped to that fifth passenger, Arland Williams. He did something amazing. He passed the rescue line on to another passenger, and did so repeatedly, until he drowned when the plane wreckage upon which he stood shifted and sank in the water.<sup>39</sup>

I hope Williams's actions inspire you. He had showed the (arguably) greatest love we can have for other persons by giving up his life to save others. Williams is not simply admirable and praiseworthy.

He merits a further response, like the child in **Paragon**.

---

<sup>39</sup> The preceding is directly quoted from the National Transportation Safety Board, “Aircraft Accident Report: Air Florida, Inc., Boeing 737-222, N62AF, Collision with 14th Street Bridge, Near Washington National Airport, Washington, D.C., January 13, 1982,” and paraphrased from McDougall (2007).

One powerful reaction to Williams came from the essayist Roger Rosenblatt (1982), who eulogized Williams as “the man in the water.” Contemplating Williams’s last seconds, he writes:

For at some moment in the water he must have realized that he would not live if he continued to hand over the rope and ring to others. He had to know it, no matter how gradual the effect of the cold. In his judgment he had no choice...

... He was there, in the essential, classic circumstance. Man in nature. The man in the water. For its part, nature cared nothing about the five passengers. Our man, on the other hand, cared totally. So the timeless battle commenced in the Potomac...

Since it was he who lost the fight, we ought to come again to the conclusion that people are powerless in the world. In reality, we believe the reverse, and it takes the act of the man in the water to remind us of our true feelings in this matter. It is not to say that everyone would have acted as he did, or as [the other rescuers] Usher, Windsor, and Skutnik. Yet whatever moved these men to challenge death on behalf of their fellows is not peculiar to them. Everyone feels the possibility in himself. That is the abiding wonder of the story.... If the man in the water gave a lifeline to the people gasping for survival, he was likewise giving a lifeline to those who observed him.

The odd thing is that we do not even really believe that the man in the water lost his fight.... The man in the water pitted himself against an implacable, impersonal enemy; he fought it with charity; and he held it to a standoff. He was the best we can do.

Rosenblatt notes that Williams exemplifies the best of what we are, and so he inspires in us the possibility that we too could be good. Williams reminds us of our human dignity.

Williams *looks* like a paradigmatic example of the praiseworthy in the same way that Harris *looks* like a paradigmatic example of the blameworthy. In order to make the cases analogous, we need to examine his life story:

**Williams, cont’d:** Arland Williams grew up in a small town. His nickname “Chub,” was “more about personality than pants size, about being a grinning, gosh-golly, aw-shucks kind of guy who wasn't even riled by everyone calling him Chub.” He went through ROTC in high school and was educated at a military college. A classmate reported the expectations for future officers: “That's an unbreakable code. You go last. Your people go first.” Never seeing military action, he instead took a post in the United States for his two required years of service. He then became a banker like his father. Williams spent the next two decades of his life checking the numbers of other bankers. At the time of the accident, he was a bank examiner during a banking crisis and was going through a divorce. During this turmoil, he began seeing his high school girlfriend again. She reported an interesting conversation with Arland. She expressed her

expectation of “100 hundred percent” commitment in a relationship. He replied with hesitation. “You have to keep a little for yourself. That's what I've learned.”<sup>40</sup>

Williams’s actions are explainable by his history. Williams had loved ones. They provided a reason to keep hold of the rope. Yet, his military training prepared him for emergencies. How do we react to Williams now? Watson speaks of the “ontological shudder” we feel upon hearing Harris’s backstory. Williams instead inspires a steadying. His story does not diminish our wonder. We too can be good.

Given this, it is odd that the cases have similar features. Both Harris and Williams perform actions that are difficult for us to imagine doing. They are both unreactive to some powerful reasons. Harris is unmoved by moral demands made on him. It is by *immunity* to normal self-regarding moral considerations that Williams inspire us. Given this, Williams seems to be, by the structure of the Strawsonian view, somewhat outside the moral community—or maybe it would be better here to say, beyond it or at its upper regions, a place few of us inhabit. Harris’s upbringing and Williams’s training seem to explain their lack of receptivity to a certain class of reasons, for Harris moral reasons and for Williams prudential ones. Despite the symmetry in the cases we feel degraded by Harris and empowered by Williams. This demands explanation.

In the next section, I will give one by arguing that there are two specific responses to Harris and Williams. I will go on to argue that these responses should be understood in terms of distinctive reactive attitudes, attitudes that have not hitherto been recognized as such in the Strawsonian literature on moral responsibility.

## 2.6 Elevation and Disgust

The literature on free will and moral responsibility lacks a term for the sort of responses found in **Cheater, Paragon, Harris, and Williams**. Luckily, we have a psychological science friendly to the interpersonal paradigm. Research in positive psychology details a family of “other-praising” emotions.

---

<sup>40</sup> Paraphrased and quoted from McDougall (2007).

Familiar among them are gratitude, admiration, “appreciation, awe, esteem, and respect” (Algoe & Haidt 2009: 107). Additionally, Jonathan Haidt, among others, propose the existence of an emotion called “elevation,” which is “a pleasurable feeling, sometimes involving warm or pleasant feelings in the chest, that trigger[s] desires of doing good deeds,” involves a “spiritual” or self-transcendent feeling, and promotes altruistic behavior (Algoe and Haidt 2009, 106; cf. Haidt 2003a; Haidt & Morris 2009; Schall, Roper, & Fessler 2010). This builds on the “broaden and build” paradigm of the positive emotions, which hypothesizes that positive emotions expand a person’s “thought-action repertoires” by encouraging exploration, play, and cooperation (Fredrickson 2001, *inter alia*). Elevation animates Rosenblatt’s “The Man in The Water” and it is an appropriate response to Williams.

We have an opposing reaction to Harris. Elevation, on Haidt’s view, is contrasted with “social disgust.” He writes:

In all of its components, elevation appears to be the opposite of social disgust. Where social disgust is caused by seeing people blur the lower boundary between humans and non-humans, elevation is caused by seeing people blur the upper boundary between humans and God (i.e., saints, or people who act like saints). Where disgust makes people close off and avoid contact, elevation makes people open up and seek contact. Where disgust creates negative contamination...elevation creates positive contamination (e.g., people want to touch living saints, or in some cultures to collect the hair, clothing, or bones of dead saints). (Haidt 2003b, 852-879)

One appropriate reaction to Harris is a kind of disgust.<sup>41</sup> The literature on disgust is too expansive to canvass here. Rozin, Haidt, and McCauley (2008) provide one way to frame the issue. They consider disgust to be a scalar defensive response ranging from bodily protection to the protection of the social order. Forms of disgust that seek to protect the body include “distaste” as a defense reaction to poison. “Core disgust” is a defensive reaction to possible contamination from disease. “Animal disgust” is a defensive reaction to death and sexuality—things that might remind us of our animal nature (cf. Rozin

---

<sup>41</sup> My discussion owes much to Strohminger (2014)’s thoughtful guidance through contours of the current debates about disgust. My presentation here follows hers closely. A reader looking for more information about this growing topic should begin with her excellent overview.

and Fallon 1987). Social disgust might be a better name for disgust over offenses of social norms rather than moral offenses. For specifying something relevant to moral responsibility we might choose “moral disgust,” reactions aimed at protecting the moral order itself. I claim that Harris morally disgusts us.

Behaviorally, “disgust is manifested as a distancing from some object, event, or situation, and can be characterized as a rejection;” phenomenologically it is typically manifested as revulsion (Rozin, Haidt, and McCauley 2008: 758-759). This comports well with other theories of disgust. Daniel Kelly (2011) develops a story of how the psychological-physiological disgust system responses that protect the physical body from contamination could be co-opted to protect the “soul” from spiritual defilement. Martha Nussbaum (2004) endorses a similar story in her work on disgust and legal theory. She carefully separates disgust from both anger and indignation as an emotion whose “core idea...is that of contamination of the self; the emotion expresses a rejection of the possible contaminant” (99).

Disgust paradigmatically manifests as an attempt to create *distance* between oneself and the harmful contaminant. This may operate by a kind of magical thinking, which may be a heuristic: once one has contacted the contagion one is always in contact with it and that shared properties indicate a shared identity (Strohming 2014: 483). For instance, people dislike eating chocolate shaped like dog feces (Rozin, Millman, & Nemeroff 1986), and people prefer not to touch objects handled by AIDS patients even after extended periods of time (Rozin, Markwith, & Nemeroff 1992). Objects that share properties with something disgusting—even when it is known that no actually harmful properties are shared—can make an object disgusting. Elevation has the opposite effect. We want to be *closer* to the elevating person.

Once someone like Harris is killed, imprisoned, or banished, the moral order regains stability because the moral contaminant is removed. This is structurally similar to the children’s response in **Cheater**. Animal disgust, reminding us of our animal nature, may make us feel degraded. One thinks,

“I am just an animal.” Moral disgust likewise can elicit self-directed degradation. Disgust involves magical thinking. We do share properties with Harris, even if we do not share any harmful properties, and so we feel the “ontological shudder.” We *cannot* conceptually distance ourselves from him. Conversely, Williams elevates us, our response to him is structurally similar to the response in **Paragon**, and we think that we too can be. We *want* to include him in our moral community, even in death.

Disgust is often considered morally dangerous (cf. Kelly 2011, Nussbaum 2004). I am, for now, using the term “moral disgust” descriptively, not prescriptively.<sup>42</sup> As a matter of fact, if someone does not respond to our blame, we may become morally disgusted. I will return to the prescriptive status of disgust.

## 2.7 Peculiar Reactive Attitudes

Elevation and moral disgust—disgust henceforth—have the right features to be reactive attitudes. Strawson demarcates the reactive attitudes by tying them to our “demand [for] some degree of goodwill or regard on the part of those who stand in...relationships to us” (2008/1962, 23). Elevation and disgust respond to quality of will. As McKenna points out, there is an intelligible quality of will expressed by Harris’s actions—a terribly evil one. Williams’s quality of will is extraordinarily good. I wish to argue that elevation and disgust are suitable response to these cases but are not themselves praise and blame. However, someone might resist this conclusion by suggesting that elevation and disgust are non-standard forms of praise and blame. Let me be clear: disgust is a response to a blameworthy agent. But, it is a response that has, in some sense, left blame behind. Disgust expresses a judgment that someone should be excluded—not exempted—from participation in the moral

---

<sup>42</sup> As Giubilini (2015) points out, “moral disgust” is ambiguous between the two uses.

community.<sup>43</sup> This is distinct from actively holding someone accountable for his or her actions, but it presupposes that the agent in question is blameworthy. Consider **Cheater** again. The players dismiss the cheater from the game, not because he is ineligible, but because he won't follow the rules. This response presupposes that he is a legitimate rule breaker! Thus, I claim that disgust presupposes our commitment to an agent's blameworthiness, but it itself does not constitute a way of holding someone accountable.<sup>44</sup> We are trying to distance ourselves physically and conceptually from the agent who repudiates morality. Thus, a fitting target of disgust is an agent who repudiates the demands of morality while generally being able to comply with them.

Elevation expresses inclusion in the moral community, and this presupposes a commitment to the praiseworthiness of the agent in question. A fitting target of elevation is an agent who exceeds the demands of morality by showing exemplary quality of will. Their quality of will shows better than expected regard for others often due to a lack of consideration for his or herself. This, on the face of it, seems to be the same conditions for praiseworthiness (cf. Pereboom 2014: 127). Recall that I claimed in **Paragon** that the exemplary little leaguer merited a further response beyond praise. I take it to be an open question whether or not elevation "leaves praise behind," but I maintain that elevation and praise are *distinct* responses to praiseworthy agents. There could be an asymmetry between elevation and disgust in this way, and we would have a natural explanation for this insofar as praiseworthy agents do not ignore or repudiate our moral address. Still—to take Haidt (2003b)'s

---

<sup>43</sup> I follow R. Jay Wallace (1994) in thinking that reactive attitudes have beliefs about quality of will, involving judgment, as their objects.

<sup>44</sup> The person who considers disgust to be a non-standard form of blame is suggesting something close to Scanlon (2008)'s sense of blame: blame is whatever response is appropriate from one person towards another person who through some action has indicated that they hold attitudes that impair the relationship between them (122-123, 128, 138). On this kind of view, disgust is a form of blame because it is an appropriate response to an impaired relationship. Harris holds attitudes that make it appropriate for us to be disgusted and distance ourselves from him. However, my analysis of disgust indicates that something is very wrong with this view. We maintain our commitment to Harris's blameworthiness, but we take no action towards him that constitutes blame. *We give up trying to express blame him.* To call this response "blame" in the fullest sense is absurd.

example of elevation related behavior quoted above—collecting someone’s bones seems to be a very different sort of thing than praise. I will leave this asymmetry question open but maintain a commitment to the distinctness of elevation and praise.

One might be skeptical that disgust is a reactive attitude because one might think that is not a form of moral address. In her defense of contempt, Macalester Bell argues that disgust is not a form of moral address because it does not make any *demands* on an agent. She writes (2013: 187-88):

Disgust presents its object as a contaminant and therefore as incapable of a response that could be given uptake: Any response that the target of disgust could offer would be unwelcome because contagions should be kept at arm's length. We don't want slime talking back to us (this is the stuff of horror movies!), nor do we want those we deem disgusting responding to our evaluation of their disgustingness. Moreover, if the message implicit in disgust is that its object is a contaminant that must be avoided, it is difficult to see how this message could be given uptake because it is not clear what would count as taking this claim seriously.

Bell is right to see disgust as a problematic bearer of moral demands in the same way that resentment, indignation, or contempt are such bearers. But this does not mean that disgust fails to be a form of moral address. When we are morally disgusted, we have *already* addressed demands to an agent and they have failed to meet them. But they also continue to ignore our demands. By expressing disgust, we communicate a powerful message: we won’t deal with you anymore. And this is a kind of demand, in the sense of an ultimatum. Moreover, we can get over our moral disgust of someone when they let us know that they recognized our demands, and this suggests that we really were communicating after all. So, I do not think we should cede to Bell that disgust isn’t communicative of a demand and therefore not a kind of moral address.<sup>45</sup>

---

<sup>45</sup> It shouldn’t surprise us that disgust at slime is non-communicative. Slime lacks the relevant capacities to understand the meaning of reactive attitudes. Disgust is a primitive emotion, and like other moral emotions it only gains moral significance as it becomes a part of our moral practices and becomes interconnected with our moral concepts.

Even if disgust and elevation are distinct from praise and blame, one might think that their distinctiveness is accounted for by being aretaic or characterological responses to agents.<sup>46</sup> Here is one clear way of spelling this out: David Shoemaker (2015) has recently defended a theory of responsibility whereby different kinds of emotional reactions pair with different aspects of “quality of will.” On this view, emotions expressive of accountability, like resentment and indignation, are tied to someone’s regard for others. However, emotions like admiration and contempt are tied to someone’s character and are expressive of another dimension of responsibility: attributability (cf. Watson 1996). Why not say that we admire Williams and have contempt for Harris? Does this explain the cases?

I am not convinced. The cases involve the maintenance of a commitment to moral responsibility in the accountability sense, and this demands a solution that focuses on accountability. Harris and Williams seem blameworthy and praiseworthy for their *actions* respectively, and yet they are inapt candidate for moral address by way of reactive attitudes. This is why the cases are puzzling. Contempt and admiration are evaluations of character. They are therefore unsuitable explanations of the puzzle found in the cases I have been considering. However, the deeper worry seems to be that disgust and elevation are not what I say they are, namely, reactive attitudes tracking regard for others in actions.

Let’s start with elevation and Williams. Williams is in many respects utterly ordinary. His character is not obviously admirable. He explicitly endorsed some selfishness in his relationships (recall his comment about not committing oneself fully). It is precisely because his character is not extraordinary that we find his action so extraordinary. I claim that *he* elevates us as an exemplar of what *we* are capable of. His character may modify our response to him, but he is not his character. His character may provide a backdrop by which our responses to him are modified. This modification

---

<sup>46</sup> Thanks to David Shoemaker for pressing me on this point in conversation.

may be especially apparent in cases where evildoers have (as Kant would call it) a “revolution of the will” and perform an extremely good action (cf. Kant 1998/1794 6:48). In cases like this, we cannot admire the evildoer’s character (it’s evil). But it precisely because they are an evildoer that their sudden regard for others shines all the brighter.

On to disgust: I have inherited the Harris case, and admittedly it is harder to pull apart his character and his regard for others in his actions. But disgust and contempt are distinct in the following case:

**Cheater II:** Cassandra and Carol have been friends and coworkers for 20 years. They work in public service. Cassandra has always looked up to Carol and admired her for her hard work and dedication. However, Cassandra finds out that for the past 10 years ago, Carol has been taking bribes. Cassandra confronts Carol, who thinks that accepting the bribes is morally justified. Cassandra thinks it is wrong but cannot do anything more; her evidence for the bribery would not convince anyone else. Cassandra genuinely believes that Carol is a good person who has made a mistake in her moral reasoning. But she cannot help distancing herself from Carol, feeling a sense of revulsion at the thought of her actions.

Cases like this show that we can separate out disgust and contempt.<sup>47</sup> Cassandra does not have contempt for Carol, even for her habitual transgressions. Instead, she feels “yucky” about her longtime friend. In light of a good character, especially transgressive actions—a dedicated public servant taking bribes, for instance— can make disgust feel especially pronounced.<sup>48</sup>

There is something to be said about why the aretaic reading of disgust and elevation is so tempting. Strawson gives the reactive attitudes a tripartite division: attitudes in reaction to another’s quality of will towards oneself, vicarious attitudes in reaction to another’s quality of will towards another person, and self-reactive attitudes related to the demand on oneself to show others good will (2008/1962: 28-29). Disgust and elevation can be other-directed *and* self-directed *and* vicarious. In

---

<sup>47</sup> It also shows that the exclusionary effect of disgust need not be total. More on this later.

<sup>48</sup> Empirically, contempt and disgust appear to have different (though related) characteristic and universally recognizable facial expressions. This suggests that they are distinct basic emotions (Ekman and Friesen 1986, as cited by Bell 2013).

the above I described their other-directed features. As they are self-directed, they share features with other self-directed emotions like shame and guilt. Arguably, guilt as a self-directed emotion is a response to a specific feature or action, whereas shame is a global negative self-assessment. It is the difference between saying, “I did a terrible thing, I ‘X’d’” and “I ‘X’d’ and therefore I am terrible” (cf. Lewis 2008). Elevation and disgust are like shame. Notice that praise and blame *lack* these self-directed features. For instance, when I am indignant, I direct no attitude towards myself. Elevation and disgust are also vicarious attitudes. Remember, the “magical thinking” of disgust: shared properties imply shared identity. Not only us, but also *everyone*, is judged to be bad in the same way. Elevation seems to involve similar, but positive, thinking.

A *global* evaluation is not necessarily an *aretaic* evaluation, but aretaic appraisals are often global. We can be interested in a specific feature of someone’s character and we can be interested in his or her character as a whole. And if the above cases are right, there can be global evaluations that are distinct from evaluations of character. For instance, consider Williams again; his life was not structured around selfless cares and commitments for others. His character is in many ways ordinary. He just performed one all-consuming act, and this makes us *see* him as the man in the water. That epitaph is, if anything, a global evaluation. In a single moment he was *for the good* in some fundamental sense. His character and his history just drop out of our picture of him in the face of his regard for the lives of others. Their disappearance is part of what makes the action so elevating—some actions make us feel *unbound* by who we are and where we come from. Disgust is structurally similar—I’ll elaborate in the next section.

I contend that elevation and disgust are reactive attitudes. They are emotional responses to regard or lack thereof for others as expressed in intentional action that are distinct from, if intimately related to, evaluations of character. My suggestion that we accept them as reactive attitudes is an

unorthodox amendment to Strawsonian views. I will now argue that their acceptance can solve the two problems. The amendment is warranted.

## 2.8 Solutions

Strawsonians need an account of why Harris is the appropriate target of reactive attitudes generally *and* why it is appropriate to stop addressing him with moral demands. This is the form of the solution to the *reductio* and self-defeat problems.

So, how can we maintain our commitment that Harris is blameworthy, but why is it also appropriate for us to stop engaging him with moral demands? He is generally an appropriate target of blame because he can understand the values of the moral community. Yet, it is not appropriate to hold him morally accountable because he repudiates our moral demands. I have argued that Harris is the appropriate target of disgust. With disgust comes a defensive reaction against that agent as a moral contaminant. Our response to Harris tries to re-establish the purity of the moral-interpersonal world by exclusion. In this case, permanent exclusion: Harris was sentence to death.

The forgoing implies that Harris is not exempted. Like Strawson, I maintain that exemption involves viewing someone from an objective attitude. Exemption means that someone is not the sort of thing that can be addressed by reactive attitudes. In fact, it would be *easier* to view Harris objectively. “It was just his childhood!” we might tell ourselves. But thinking that Harris ought to be exempted misunderstands the *personal* nature of the case and his crimes. Given this, the Strawsonian who accepts disgust avoids the *reductio*. It is not absurd to think that Harris elicits in us a complex response that is not blame, but is also not exemption, and is instead exclusion. What about Williams? *Exceeding* the criterion for responsibility should not exempt someone from responsibility. I agree. He is an apt target of praise, but also the distinct attitude of elevation.

Given disgust, the Strawsonian need not fear the self-defeat problem. We do make a negative self-assessment when faced with Harris that involves some generalization about our common nature.

This generalization, however, is not one about our particular histories, and this was the source of the self-defeat problem. I have argued that elevation and disgust do not focus on character or history primarily, even if they are global responses to others. Moreover, even if we feel sympathy towards Harris because of his past, this does not dispel our disgust. Why? History and character can *heighten* our feeling of disgust even in the presence of sympathy. Like how violations of norms about sex and death can remind us of our animal nature, Harris's history reminds us of our own moral fragility. Disgust makes us feel *bound* to who we are and where we came from. Insofar as Harris makes us feel this way, we are morally humiliated.

But for every case that instills doubt about ourselves, we can find another to restore our faith, like the case of "The Man in the Water." The lesson to learn from this is that the moral community has no well-defined boundaries. Persons sometimes fall above and below the threshold of ordinary moral expectation. Sometimes this can happen by an agent's own doing. Like Harris, persons can remove themselves from the moral community. Internal to our practices there are reactive attitudes that express our responses to such persons. When we see someone who rejects moral expectations, we exclude them. When we see someone do *better* than those expectations, we raise the expectations we have on ourselves.

Strawson himself has more to say. He considers "something far above or far below the level of common humanity," i.e., something that breaks free of our "common roots in our human nature and our membership in human communities." He calls the thing that lacks the self-directed attitudes a saint and the thing that lacks the other-directed attitudes a moral idiot. We imagine Harris to be a moral monster. But Williams and Harris are not Strawson's imagined beings even if they seem like them. As Watson reports: "in his last years, Harris either remained, or became once again, capable of friendship and remorse. His crimes were monstrous, but he was not a monster after all. He was one of us" (Watson 2008/1987, 141). Why does Watson tell us this? Harris mouthed the words "I'm sorry"

to one victim's father moments before he was put to death by the state (ibid, 140). So, he was not *really* a monster. Given this, maybe disgust is immoral. Should someone like Harris really face fatal exclusion from the moral community?

## 2.9 The Ethics of Disgust

The answer to the preceding question is difficult. Here is the problem. Disgust is my way out of a theoretical problem for the Strawsonian responsibility theorist. But what if disgust itself is morally problematic? Many authors have worried about the morality of disgust as a response to human persons. It often misguides our moral lives. Immigrants, people of different sexual orientation, race, and so on, have both presently and historically been the targets of unmerited disgust. For instance, Daniel Kelly is skeptical of the possibility of genuine moral disgust because disgust has become associated with immoral attitudes (2011, 128). And Martha Nussbaum is positively anti-disgust. Indeed, one might say that she is disgusted by disgust. She writes: "I would argue...that even the moralized form of disgust is an emotion that is highly problematic. It must be contained and perhaps even surmounted, on the way to a genuine and constructive social sympathy" (2004: 105).

The disgust skeptic has a serious worry. Consider this argumentative strategy: Bell (2013) argues (to my mind, persuasively) that contempt can be both a fitting and reasonable response to persons who embody the "vices of superiority", like arrogance or hypocrisy. Moreover, holding those persons in contempt is the best moral response to those vices available. We can say similar things regarding disgust. Moral disgust can be fitting towards persons who are genuinely unmoved by moral demands. It can be reasonable to be disgusted if we have sufficient evidence that the target of our disgust is truly unmoved. Although I do not endorse moral disgust as the required response to the moral fault of being unmoved by calls for accountability to the demands of morality, I do think disgust is uniquely good as a response to this fault.

This will not convince the disgust skeptic. She may object to this view in several ways. For instance, disgust skepticism can be motivated by disgust's exclusionary effects. On the view I have offered, disgust has the functional role of protecting the moral community by excluding persons who are not open to moral address. This may seem unduly harsh, especially in light of Watson's post-script to "Responsibility and the Limits of Evil". This problem can be resolved if we endorse the view that the exclusion brought about by disgust comes in degrees. And we should endorse that view. Consider again **Cheater II**, in which Cassandra is disgusted by Carol's corruption in the public service. Cassandra does not think that Carol should be excluded from the moral community, full stop. (She *should* not, in any event). Cassandra nevertheless seems warranted in distancing herself from Carol. What Carol has done *is* morally disgusting. It is not as morally disgusting as eating the lunches of one's murder victims. In extreme cases of evil-doing, the appropriate degree of exclusion may indeed be removal from the moral community. The degree of exclusion which a morally disgusting act merits therefore seems commensurate to the degree to which the action itself merits disgust.<sup>49</sup>

Let's consider a different way to motivate disgust skepticism. One might wonder how disgust can ever be a fitting response to a human person if one thinks that disgust is inherently dehumanizing. Consider that dehumanization is a possible outcome of the Strawsonian view of exemption anyways. From an objective attitude, another person is as an object of casual or explanatory inquiry. We ask, "how it is structured and/or how it functions" (Bennett 2008: 53). The objective attitudes at their most extreme—although often accompanied by positive non-reactive emotions—can dehumanize persons by reducing them to non-agents (cf. Strawson 2008/1962: 24). That the objective attitude is dehumanizing is consistent with Strawson's description of it as reducing human behavior to causal understandings of "programmed mechanisms" (Strawson 2011: 140). These kinds of explanations, at

---

<sup>49</sup> I offer no calculus here; in real life cases, the degree of appropriate disgust, and therefore appropriate exclusion, will likely be a matter of good moral judgement.

least some of the time, can be morally degrading. Despite this, we do not think that exemption is never a fitting response to human persons. Likewise, we are disposed to experience disgust. It is surely true that some token instances of disgust in some domains, like those Kelly and Nussbaum find objectionable, are highly problematic and unfitting. But it does not follow that all tokens of disgust are.

I think we should be unmoved by the worry that disgust is, in itself, morally problematic because it is dehumanizing. But even if disgust can be an appropriate moral response, perhaps disgust is generally problematic in some other way. Disgust is an admittedly blunt instrument in morality's toolbox; one might worry that it promotes bias in moral judgement.<sup>50</sup> If we are frequently disgusted, we may be more likely to be disgusted by all the wrong things.

The worry appears empirically supported. In a series of interesting experiments, Jones and Fitness (2008) found that individuals who were highly sensitive to disgust were prone to several biases in their judgment.<sup>51</sup> Jones and Fitness believe that this is evidence for the view that individuals who are highly sensitive to disgust exhibit what they call *moral hypervigilance*, “a range of cognitive, emotional, and behavioral tendencies that are intended to reduce the risk of exposure to moral contaminants” (2008, 614). They found that, for instance, persons who were highly sensitive to disgust made higher estimates of the probability that suspects described in crime vignettes were culpable and were more likely to recommend lengthy sentences for criminals. Moreover, persons who are sensitive to disgust also had comparatively inflated perceptions of community crime levels. Perhaps most interestingly, Jones and Fitness found that highly disgust-sensitive persons also displayed a greater tendency to attribute evilness to criminals. These findings fit well with my view that moral disgust presupposes

---

<sup>50</sup> Thanks to an anonymous reviewer for raising this important objection and pointing me towards the relevant empirical research.

<sup>51</sup> Jones and Fitness used versions of the disgust scale presented in Haidt et al. (1994) to measure disgust sensitivity. Importantly, this disgust scale does not specifically focus on one kind of disgust, e.g., moral disgust.

accountability but involves a distinct exclusionary response especially aimed at evildoing. This seems all the worse for my view, though. Persons who were sensitive to disgust displayed systematic differences in their moral judgment than persons who were less sensitive to it. And these judgements were harsh. Disgust seems to lead to hypervigilance against moral contamination, and correspondingly, a tendency towards drastic and potentially unwarranted countermeasures.

In response to this worry, it should first be noted that Jones and Fitness themselves state that their use of the term “bias” is not meant to imply that persons who are highly disgust-sensitive are making errors in moral judgment (2008: 615). They did not measure the accuracy of the judgements in their experiments. They instead use the term to refer to dispositions to make judgements of a certain sort. But let’s assume that highly disgust-sensitive individuals are making moral mistakes. What then? Jones and Fitness do not investigate the relationship between the emotional *state* of disgust and moral hypervigilance. Rather, they investigate the relationship between moral hypervigilance and a personality *trait*, disgust sensitivity.<sup>52</sup> So, even if high sensitivity to disgust does produce problematic bias in moral judgment and hypervigilance, this does not mean that any token state of being disgusted is morally problematic. On my view, moral disgust has an important moral role to play in policing the boundary of the moral community. It is, however, a weapon of last resort. It should be used sparingly and only after calls for accountability via blame fail. Given this, it is perfectly possible that persons who are highly sensitive to disgust may misuse this powerful moral tool and so become inappropriately hypervigilant. Strawsonians take the moral emotions to play a key role in our moral practices of accountability. We do not endorse the stronger thesis that we ought to inculcate in ourselves dispositions for hard feelings. We may even say that doing so is morally wrongheaded if these dispositions promote moral error.

---

<sup>52</sup> Specifically, Jones and Fitness label disgust sensitivity a trait-like construct that describes the ease and intensity with which one is prone to experience disgust (2008, 614).

To see why such a response should convince, compare with anger. Strawsonians take forms of anger (resentment and indignation) to express (or constitute) blame. That is, *states* of anger express blame. But *trait* anger, being disposed to respond to situations with anger, seems correlated with biases in moral judgement. Epps and Kendall (1995) showed that high trait anger in adults is associated with hostile interpretation bias—the tendency to interpret others as having hostile intent when social cues are ambiguous. Wenzel and Lystad (2005) found that although both anxious and angry individuals demonstrate negative interpretation biases when asked about ambiguous but potentially threatening scenarios, the bias is more extensive among angry people. In Hazebroek et. al (2001)’s study, individuals with high trait anger blamed a wrongdoer more than individuals with low trait anger when presented with videos of social interactions. This difference was especially striking in cases where the wrongdoer’s intent was ambiguous.

Do these results mean that Strawsonians should accept that anger is morally problematic? No. Nothing so far shows the accuracy of the judgements in question. And even if these biases are morally problematic, it only shows that angry people are prone to moral error. Reflection on ordinary life makes this apparent enough. Sometimes, though, we need anger to stand up for what is right and to hold wrongdoers to account. We should say similar things about disgust. We may need it to respond to those who are unresponsive to morality’s demands, in spite of the danger. We should try not to be characteristically angry or disgusted people.

Briefly put, we have need for the more difficult moral emotions even if they come with the risk of error in moral judgment. The lesson to draw about disgust and anger from this is simple: *use with caution.*

## 2.10 Conclusion

I have argued that two challenges to Strawsonian compatibilism fail. The first was a *reductio* problem that evil agents are exempted from moral responsibility. The second was a self-defeat problem that a

person's history dispels reactive attitudes. My response was to suggest the adoption of elevation and disgust as reactive attitudes. This solution, although with questions unanswered, is plausible and explanatorily powerful. It is not an ethically problematic solution either. But beyond contributing a novel solution to these problems, I hope to have shown the theoretical fruitfulness of being true to our experience as moral agents. Our moral reactions to others, especially in extreme cases, are varied, subtle, and messy. This is a feature of our lives to be welcomed rather than ignored.

## Chapter 3

### P.F. Strawson and The Case of the Missing Account of Control

**Abstract:** In opposition to alternative approaches to the problem of free will, P.F. Strawson's (2008/1962) strategy in "Freedom and Resentment" focuses on the nature of moral responsibility directly (that is, without talking about the metaphysics of control or freedom). It is widely thought that this approach is friendly to compatibilism. Perhaps by looking at the relevant norms of holding responsible we can extrapolate a conception of the relevant kind of control demanded by such norms. We may find this control to be compatible with determinism. In spite of my own sympathies to Strawsonian compatibilism, I argue Strawson's methodology looks promising in favor of incompatibilism in a way that has gone unappreciated.

#### 3.1 Introduction

To some, P.F. Strawson's "Freedom and Resentment" is the most important contribution to the free will debate in the 20th century. To others, it badly misses the point. What gives? Many take the problem of free will to be a problem in the metaphysics of agency, albeit one with profound implications for moral theory.<sup>53</sup> Many others, however, are amenable to theorizing about free will in relation to moral responsibility. Perhaps the best way to get a grip on the nature of free will is to focus on the control needed for morally responsible agency. Of course, though, such control might be manifest in non-moral actions too.<sup>54</sup> In opposition to either of these approaches Strawson instead focuses on the nature of moral responsibility directly (that is, without talking about either control or freedom) in an attempt to address the problem. One might see this as a metaethical inquiry quite apart from investigation into the nature of free agency.<sup>55</sup> Thus, one might think that Strawson misses the point. But many contemporary philosophers have warmed to this way of approaching questions about moral responsibility, and by extension, free will. After all, Strawson's project looks theoretically appealing. It appears to offer a way of getting compatibilist conclusions without directly taking on the burdens of the free will debate.

---

<sup>53</sup> See for instance Vihvelin (2013), van Inwagen, (1983, *inter alia*) Ginet (1997, *inter alia*), Campbell (1957).

<sup>54</sup> See for instance, Fischer and Ravizza (1998) Mele (2006), Pereboom (2014), McKenna (2012), Sartorio (2016).

<sup>55</sup> The careful reader of Strawson is sure to note the paucity of the phrase "free will" throughout the entirety of the essay.

Incompatibilists are not so sure that Strawson's approach wholly avoids these burdens, and have challenged Strawson's account (e.g., Watson 1987, Pereboom 2014, *inter alia*). The charge is that Strawson's view leaves open, or even supports, incompatibilist concerns about the nature of moral responsibility. This problem for Strawson can be traced back to his lack of an account of control.

I aim to show that this problem for Strawson is worse than has been generally received by trying to find a plausible version of Strawson's missing account of control. It is not obvious that the appropriateness conditions for those moral emotions are compatibilist-friendly given Strawson's own commitments. To show this, I will not offer a rigorous survey of this extensive literature on "Freedom and Resentment". Nor will I engage with the (perhaps even larger) literature that takes serious inspiration from Strawson's view. Instead, I will exploit crucial themes from "Freedom and Resentment", offering charitable reconstructions in light of Strawson's later work, and extending the line of thought when needed.

To be clear, this worry doesn't decisively refute Strawsonian compatibilism in its various contemporary guises. Nevertheless, I take this to problematize part of Strawson's legacy, the general idea that a direct focus on responsibility could obviously vindicate compatibilism.

### **3.2 Strawson's Main Argument**

Strawson's main argument begins with three theses. Each involve the shape of our moral responsibility practices. First, Strawson accepts that to hold someone responsible is to have towards this person a *reactive attitude* like resentment or indignation. These attitudes express (or constitute) our moral demands towards others. Second, he argues that reactive attitudes are reactive to *quality of will*, the good or ill will with which a person acts. In other words, Strawson thinks that what we are responsible for is the moral concern (or lack thereof) we show to others in our actions (cf. McKenna 2012). Lastly, Strawson endorses the idea that it is from the standpoint of holding accountable that we are to understand responsibility and desert (2008/1962: 22-23). So, importantly, even though many

paradigmatically non-responsible persons—the mentally ill or the very young—can perhaps express quality of will, they are not participants in the sort of interpersonal relationships within which we find the demand for good will to be appropriate. In other words, they lack quality of will of a morally salient sort. Thus, they are not morally responsible agents. Since praise and blame consist in conveyed reactive moral sentiments, they play a foundational communicative role in our moral practices. They express to others the moral demands to which we hold them accountable. This makes fitting the theoretical methodology at work in the view. We ought to take the commitments and practices that we adopt as moral interlocutors as our guide in figuring out the facts about who can be expected to meet the demands of morality. Indeed, Strawson contends that this framework of moral emotions is part of what it is to be interpersonally engaged with others. He writes: “in the absence of any forms of these attitudes it is doubtful whether we should have anything that we could find intelligible as a system of human relationships” (2008/1962: 36).

The argument now continues as a proof by cases. Here is the first case: excuses and justifications. These kinds of pleas are employed to show that some purportedly blameworthy agent is not so because they did not act from ill will—or at least from the sort bearing on the grounds for holding to account. Thus, one might explain that one tripped and bumped into someone else to forestall the blame warranted by (what might have been understood as) an angry shove. Likewise, one might justify one’s actions as necessary. If I plead with you by saying that I only shoved you so that you wouldn’t be hit by an oncoming car, I am trying to tell you that my action involved no ill will towards you, but rather involved good will. It was justified. Thus, excuses and justifications are attempts to show others that an action that looked like it was motivated by less than good will for others was not what it first appeared to be. They invite us to see the putative injury in a new light (ibid. 23).

A more extreme plea is available within our practices: exemptions. This is the second case. Exemptions show that someone cannot be blameworthy because they lack the capacity to engage in adult interpersonal relationships. They are mentally ill, a small child, and so on. So, while excuses and justifications show that no morally salient ill will occurred in an agent so capable, exemptions show that it is wrongheaded to expect appropriate good will (of the pertinent sort) from that agent at all, since they are unable to participate in the relationships within which such a demand makes sense. In these latter cases, we view the agent in question with more or less *objective* attitudes, rather than reacting to them with attitudes constitutive of participation in interpersonal relations. At the extreme, these objective attitudes involve seeing someone as an object “to be managed or handled or cured or trained” (1985: 25). To use Strawson’s own terminology, excuses and exemptions are appropriate for normal persons. Exemptions indicate either temporary or permanent abnormality by way of incapacitation from normal interpersonal relations.<sup>56</sup> Instead of inviting us to see a putative injury in a new way, exemptions ask us to “see the *agent* as other than fully responsible” (2008/1962: 23).

The argument now proceeds quickly. In the first case, does the truth of physical determinism provide grounds for excuses or justifications for normal agents?<sup>57</sup> No. Determinism does not mean that everyone acted without ill will or that everyone acted with sufficient good will. It would be absurd to think that no blameworthy actions ever occurred just in the sense that no one ever showed unreasonable disregard for others, something Strawson claimed would involve the universal reign of good will, not determinism. Likewise, in the second case. Does determinism mean that everyone is incapacitated in such a way as to be exempted from responsibility, as in viewed from a purely objective

---

<sup>56</sup> Strawson’s use of the term “abnormal” has caused some interpretive controversy. See Bennett (1980) for criticism.

<sup>57</sup> Determinism is the thesis that that two propositions, one describing the past at a given time (usually the distant past for effect), and another describing the laws of nature, together entail a proposition describing the one unique future. Strawson proceeds without a formulation of determinism. Given the time, he may have had universal necessitating causation in mind.

viewpoint? No. This would render everyone “abnormal” and unsuitable as a participant in interpersonal relationships. This consequence is absurd. As such, we should conclude that determinism would not make us abandon the normative standpoint of praise and blame. There is no obvious connection between determinism and the sorts of considerations that release agents from responsibility. So, this general metaphysical thesis is irrelevant to our way of going about things. In fact, Strawson thinks it would be practically inconceivable for us to give up our ordinary way of relating by way of reactive attitudes because of determinism (2008/1962: 26).

It should now be apparent how Strawson’s compatibilism is deeply immodest. Strawson’s argument suggests that the basic moral concerns expressed in our interpersonal moral practices and rooted in our human emotions and attitudes—irrespective of the “local and temporary features of our own culture”—are not threatened by the truth of physical determinism (2008/1962: 36). Even if they were threatened, Strawson thinks, our “human commitment” to such concerns are impossible to relinquish (2008/1962: 26). And given that we have these basic concerns, it would be practically irrational to give them up if we could.

### **3.3 The Specter of Incompatibilism**

Should this argument convince? I think not. The argument is problematically incomplete. Strawson’s explicit aim is to correct the compatibilists of his time by disabusing them of a theoretically inadequate consequentialism.<sup>58</sup> As far as I am concerned, point taken. The problem comes with his other aim: trying to show libertarians that a richer conception of morality, with desert-based notions, need not require justification by libertarian free will.<sup>59</sup> It is here that the argument as I read it is in need of support.

---

<sup>58</sup> See for instance Schlick (1939) or Ayer (1954).

<sup>59</sup> See for instance Campbell (1957).

Strawson does not specify what sorts of capacities are needed for engagement in our interpersonal practices of moral responsibility. He does not tell us what constitutes the normality appealed to in the cases. This has been the subject of extensive debate. For he seems to leave open the possibility for an incompatibilist to assert that libertarian free will is in fact a condition of normal, adult human agency such that, if we lack it, then we all should be exempt from moral responsibility.<sup>60</sup>

Why might one take libertarian free will to this condition? Well, it seems that there are existing pleas in our moral responsibility practices that seem to count universally if determinism were true. Determinism is the thesis that that the (even distant) past and the laws of nature entail the unique future. If there is one unique and unavoidable future, it looks like “being unable to do otherwise” counts as an excuse in every case. Note that one only needs a token instance of a plea to count in every instance in order to reject Strawson's argument. As Russell (2017) argues, even if we could never dispel the natural fact that we feel fear, we could find out that each token instance of fear we experience is unwarranted. (Imagine a person who is afraid of oranges and nothing else). Likewise, even if we couldn't dispel our reactive attitudes, we could learn that each token instance where they naturally arise in us is unwarranted too. Perhaps at every *token* instance where we might start to feel indignation, we could come to see the other person as being physically unable to have done otherwise than they do. "Aha!" we might say, "I forgot about the truth of determinism for a moment". This might expel our indignation toward them. How could we blame someone who could not have avoided wrongdoing? That the generalized excuses are tokens and not types of excuses is important. For Strawson seems to have argued that we could not dispel types of responsibility responses. He has not yet ruled out that in every token case of resentment, say, might be inapt.

---

<sup>60</sup> See: Wiggins (1973), Bennett (1980), McKenna (1998), Watson (2004), and Russell (2017).

Notice that the excuse “I couldn’t have done otherwise!” might be taken to express a metaphysical criterion for free and responsible action, namely, be a condition that requires the existence of actionable alternative possibilities. Consider a different common plea: that someone is unfortunate in formative circumstances. If someone was horribly abused as a child and then grew up to be a murderer, we might find that our resentment or moral indignation at them is somewhat suspended. Why do the reactive attitudes seem less appropriate in these cases? That person’s history *explains* their ill will towards others. Now, though, we might worry that if someone’s history explains their quality of will in one case, why couldn’t *anyone’s* history explain their quality of will?<sup>61</sup> This commonplace plea shares a similar form to the worries raised by manipulation arguments.<sup>62</sup> If one is manipulated, one is not free and responsible because one is not the source of one’s actions in the right way; likewise, if determinism is true. Pleas of this sort seem to latch onto a different metaphysical criterion on free and responsible action: an ultimacy condition that free and responsible action requires that one be the ultimate source of one’s own actions.<sup>63</sup> So, some excuses and exemptions internal to our practices seem to apply universally if determinism is true. And all the worse for Strawson, they latch onto concerns extensively discussed in the traditional debate about the metaphysics of free will.

Our sensitivity to these criteria suggests that it is quite possible for us to see (by one route or another) that everyone’s behavior is not up to them. And if this were true, on Strawson’s own view objective attitudes would be appropriate towards every person. It is hard to see why this appropriateness would be practically irrational given what we have learned about ourselves. It is at least unclear that it would be psychologically impossible. Crucially, though, it does not seem practically

---

<sup>61</sup> This point is part of Gary Watson’s famous exploration of the topic in his “Responsibility and the Limits of Evil: Variations on a Strawsonian Theme”, reprinted in (2004).

<sup>62</sup> See Mele (2006) Pereboom (2001) and (2014).

<sup>63</sup> Cf. Kane (1996: 33-37)

inconceivable. And given that it is not practically inconceivable, the practical irrationality of a change in the basic concerns of human interpersonal life is no longer obvious.

This kind of reasoning is sometimes referred to as the *generalization strategy* against Strawson. By taking an appropriate case of suspended reactive attitudes, we can generalize to every other token case. This reveals that the kinds of concerns at play in our practices of moral accountability are in fact metaphysically laden. They are sensitive to determinism in the form of excuses and exemptions. This suggests that the capacities needed for moral responsibility are not compatible with determinism.

In other words, Strawson needs to specify the kind of control needed for moral responsibility. Otherwise, it is not obvious that, for all Strawson has said about our moral lives, the control we implicitly assume therein is amenable to compatibilists.

### **3.4 Themes from “Freedom and Resentment**

Even if there is no clear account of the control needed for moral responsibility in the text, there are several thematic suggestions we can glean from Strawson that might help us fill in this critical omission. I consider four below.

#### *3.4.a Sentimentalist Deontology*

Strawson’s strategy is to look first at the conditions in which it is apt to hold persons morally responsible. Then he attempts to show how determinism does not fit into the set of cases in which it is not apt to do so. These aptness conditions would be a good place to start looking for the missing account of control.

Strawson’s account of these conditions is not neutral with respect to competing conceptions of morality. The moral norms to which we are accountable are deontological in nature; he treats them in a roughly sentimentalist fashion. If one recalls the history of philosophy, this kind of combination of metaethics and normative ethics should strike one as odd. But this combination is psychologically

realistic. Indeed, a growing body of literature suggests that deontological moral notions are psychologically rooted in our emotional capacities rather than our rational ones.<sup>64</sup>

Strawson's account suggests that blameworthy wrongdoing is wrong action which *manifests* a lack of sufficient good will towards others. Indeed, one can get into a "Strawsonesque" mindset such that what makes a wrong action *wrong* is that it displays the absence of sufficient good will. (This might offer an appealing account of *subjective* wrongdoing). For on his view our emotional responses are in response to our perception that others are not acting with moral concern for ourselves or others. It is the presence or absence of good will, Strawson seems to suggest, which is our basic concern when it comes to our moral dealings with other persons. It is the basic thing which we police through blame and encourage through praise.

The sort of control one would need to be held to a standard of displaying good will towards others therefore seems to have three components: emotional, volitional, and intellectual. Responsible agents must desire what is in others' interests or want what is best for them in the spirit of fellow-feeling. In other words, they must have the emotional powers needed to have good will towards others. They must also have emotional powers to react to quality of will by way of emotions resentment, indignation, gratitude, and so on. Moreover, responsible agents have the sorts of powers that enable them to perform actions voluntarily so as to display good will towards others. (Note that I am not saying here that we need this power to *intentionally* express our good will towards others. You can unintentionally show a lack of good will towards others by voluntarily acting in ways which inadvertently fail to show some persons due moral consideration). Finally, Strawson's account of control must have a particularly social component. This is best understood in terms of intellect or understanding. In order to act so as to *manifest* one's concern for others, one has to understand that

---

<sup>64</sup> See for instance Haidt and Greene (2002) and Nichols (2004).

actions can mean things to other persons. Thus, a morally responsible agent understands how to assess the actions of others. She can later to correct any misunderstandings by pleading. And she understands her actions are likewise assessable by others.<sup>65</sup>

Straightaway we should notice that the power to act voluntarily so as to display moral concern (or lack thereof) does not seem to be a *two-way* power. Acting so as to display adequate moral concern does not seem to require that one also be able to act so as *not* to display adequate moral concern. The most virtuous person imaginable may be psychologically unable to express a lack of concern for others. It may therefore be *inevitable* that she does what's right. But we would still praise her when she does good for others. Why? Her actions display her own good will for others. The inevitability of her actions does nothing to explain why she acted thus-and-so. Now some think that there is an asymmetry between praise and blame. Perhaps blame requires fair opportunity to avoid wrongdoing.<sup>66</sup> But Strawson's picture of our basic concerns suggests symmetry. Imagine someone who was psychologically constituted so as to inevitably show a lack of good will towards persons of a specific race. Indeed, imagine that this person shows *gleeful* disregard towards them in her actions. Is she the target appropriate target of our resentment? Perhaps some will think not. (Much more on this in a moment!) But I think by parallel reasoning to the case of the virtuous agent, the picture of control we have painted for Strawson would answer in the affirmative.

So, the kind of control needed to fit into Strawson's basic picture of moral life appears *not* to require an ability to do otherwise. Our moral concern with other's actions has to do with the moral consideration displayed in those actions. Hence, excuses, justifications, and exemptions each in their own way show that what *looked* motivated by a lack of good will was not. Control—mediated by the role of quality of will—is displayed in the actual sequence of events that led up to the action. At the

---

<sup>65</sup> Cf. Russell (2017) and McKenna (2012).

<sup>66</sup> For arguments in favor of this asymmetry see Wolf (1990) and Nelkin (2014)

very least, the generalization strategy in terms of the excuse “she couldn’t have done otherwise!” seems diffusible in light of this point. If excuses of this form only operate in special circumstances, then this excuse does not generalize when determinism is true. An agent’s lack of an alternative can feature in an explanation of her quality of will. My being violently coerced into wrongdoing implies no ill will on my part towards those wronged, for instance. Generalizing, the class of things that can make an action inevitable is not coextensive with the class psychic forces that can bring said action about (cf. Frankfurt 1969). Given this, universal inevitability would not rule out the possibility that actions manifest a lack of good will as a matter of what actually brought those actions about. At least saying something like this would be a natural development of Strawson’s argument.<sup>67</sup>

What about exemptions? Consider a common distinction in the free will debate between *general* and *specific* abilities. To have a general ability is to be constituted so as to be able to do a certain thing. To have a specific ability to do a certain thing is to possess a general ability to do that thing, but also the opportunities to exercise one’s ability. If, plausibly, exemptions only require that one’s general ability to act with the control needed to manifest one’s quality of will, then exempting conditions do not generalize when determinism is true. For determinism should not interfere with our general abilities. Why is this plausible? It is only fair to hold those persons who can fully participate in our moral practices to their demands.

Notice that on this way of construing pleas, the historical sources of a person’s quality of will appear to be of minimal relevance. They explain why an agent *has* an ability to act with the control needed to manifest her quality of will. But what matters to us, following Strawson’s read on our moral practices, is that there *is* quality of will displayed in action. And so, we might think it does not matter *where*, in the history of a person’s formation, good will (or lack thereof) came from. This reply therefore

---

<sup>67</sup> Michael McKenna (2004) pointed this out, including the convergence between Strawson and Frankfurt.

answers a *metaphysical* question about whether or not free agency is “snapshot” or “historical”. Like Frankfurt (2002), Strawson seems like he should deny that the sources of our moral competence are relevant to our current control. Thus, on the form of the reply to the generalization strategy above, Strawson’s picture of control is snapshot.

This reply on Strawson’s behalf has the structure of an appeal to best explanation. When we look at what it takes to appropriately participate in our moral practices, we see that these conditions are compatibilist-friendly.<sup>68</sup>

Is this sufficient? I think not. For it seems as though the sources of our actions matter beyond giving rise to one’s abilities. Consider, for instance, Derk Pereboom’s “internal challenge” to Strawson, based on his 4-case manipulation argument (2014: 74-79). Pereboom’s manipulation argument challenges compatibilists by specifying cases in which an agent, by stipulation, meets any compatibilist set of necessary and sufficient causal conditions for free and moral responsible action, and yet seems non-responsible for her action. I will not set out the cases in their intricate detail, which has become necessary in response to various objections. (My point depends on the intuitive pull of Pereboom’s argument not its ultimate success). The first case is one of manipulation: an agent’s action is deterministically produced by manipulators at the time of the action. The second case is manipulation at the point of origin. The manipulators “wire” the agent from the beginning of her life so as to perform the manipulator’s preferred action. The third case involves the training practices of the agent’s community deterministically bringing about the action. And the fourth case is explicitly casual determinism in virtue of the past and the laws of nature. Again, each case is such that the agent, by stipulation, meets any given compatibilist conditions for free and responsible agency. Pereboom thinks

---

<sup>68</sup> This is the general form that R. Jay Wallace’s (1994) Strawsonian defense of compatibilism takes. McKenna (2004) develops the affinities between Strawson’s view and Harry Frankfurt’s rejection of the principle of alternative possibilities and his “snapshot” account of control.

that in each case we have made the judgement that the agent is not responsible. And the best explanation of this, Pereboom argues, is the salient feature of each case. That feature is causal determination by factors outside of the agent's control.

Notice that Pereboom's argument is cast in terms of best explanation. Pereboom's argument presents us with countervailing intuitions about the basic concerns expressed in our moral practices. They are intuitions to the effect that the control discussed above is insufficient for moral responsibility. Responsibility has a historical dimension, or more broadly, a dimension pertaining to the sources or origins of our actions which can be undermined by external causal determination in a way which is at odds with a "snapshot" approach. Now, my point is not that the Strawsonian proposal sketched above fails outright. It is just that it is once again incomplete. This incompleteness is methodologically troubling. For it seems as though Strawson's picture of morality suggests that the sources of an agent's control are not relevant to her moral responsibility. At any rate, it would be natural for Strawson to say so. Yet, Strawson's explanatory project is difficult to square with this view. For he has set out his view as one in which moral responsibility is to be explained by the appropriateness conditions of holding responsible. But it is not appropriate to hold manipulated agents responsible. Pereboom's argument suggests that, if determinism is true, we are all as if manipulated. Worse, *we* intuitively judge it as such. And this cries out for explanation. Indeed, one might suspect that, at heart, this judgement is explained by fairness too, for it seems deeply unfair to blame agents who are causally determined to be who they are, and so act to as they do.

This suggests that Strawson's project is badly incomplete. It does not offer a "snapshot" approach to the metaphysics of agency. Rather, it appears to need support from an independently plausible "snapshot" metaphysics of control apart from inquiry into our moral practices. For our practices enshrine a powerful, seemingly incompatibilist intuition about moral responsibility (and so freedom).

### 3.4.b *Natural Commitments*

A natural way of reading Strawson undermines his point. The pertinent sense of control needed for moral responsibility is to be explained by the shape of our actual moral practices. Our actual practices seem to enshrine an incompatibilist conception of the sources of our freedom. Yet Strawson himself suggests being particularly unfortunate in formative circumstances as an example exempting condition (2008/1962: 25). Evidently, he thought this didn't threaten compatibilism. But the compatibilist-friendly thought discussed above—that our formative circumstances matter only insofar as they give rise to our ability to manifest quality of will in our actions—was insufficient in meeting the incompatibilist's challenge. How should we make sense of this?

The lesson from Strawson's sentimentalist deontology is that pleas have a special function. They are ways of initiating a corrective. Typically, we can safely infer from certain act-types that a person has a certain quality of will. If someone punches you, it safe to infer that they had insufficient good will towards you. If you confront this person and they plead with you, they are trying to show you that they did not in fact act from ill will (or that they are not the sort of person who has quality of will of the pertinent sort).

Now when we think of exemptions in the form of unfortunate formative circumstances, we are talking about has to do with the sources of action. But notice that the target of this "source" concern is *not* the competence we display in acting thus-and-so. Rather, the concern has to do with *why* we act thus-and-so. Consider again the manipulation worry. It has to do with *why* we are made up of the various compatibilist-friendly psychological forces which stir us to action.

It is natural to think about this in terms of our values. Consider the following view. What makes an action free is that it, in fact, springs from our values in the right way.<sup>69</sup> Now, this kind of

---

<sup>69</sup> Watson (2004) offers a view of this sort.

view requires something action-theoretic too. For it seems not merely that we have values, but moreover, a power to *act* on those values. If otherwise, we could not explain actions which are voluntarily done *against* our own values.<sup>70</sup>

On this view, we are the source of our actions insofar as our values are our own. But it is plain to see how pleas about manipulation and history and self-formation might arguably generalize if determinism were true on such a view. For even if we are in fact able to act in accordance with our values, one could argue that those values are not really our own in deterministic worlds. For, arguably, it is not up to us that we have our values. They are ultimately the product of forces outside of our control. And so, the “special” circumstance of the plea that one was raised in unfortunate formative circumstances ends up being just like any other formative circumstance.

Let me go back to one of the historical reference points to highlight the extreme difficulty of this problem for realists about free will and responsibility—those of us who think that we really do have free will and are moral responsible. For instance, Kant suggests in places that the *whole of one’s life*, including one’s formative circumstances which cause oneself to be a certain way, are to be understood in the context of a *transcendental choice* to conform the ground of one’s maxims to the moral law or not (*KvP* 5:97-98). To contemporary ears this might sound incredible! But this only goes to show how seriously Kant took the kind of worry about freedom and responsibility we are discussing here. (At least someone familiar with the literature on free will might read into Kant’s thought on the subject this level of seriousness; Kant himself may have had other concerns).

Consider a more contemporary reference point. Robert Kane (1996) has argued that unless the sources of our actions are formed by ourselves in free choices, we lack free will. He calls the pertinent actions *self-forming actions*. Unless it is possible for us to performing genuinely self-forming

---

<sup>70</sup> See Velleman (1992) for the problem this is meant to solve. McKenna (2019) offers something like this.

actions, we lack free will and so are not morally responsible. Indeed, to some compatibilists, his defense of this possibility involving quantum indeterminacy might sound nearly as incredible as Kant's transcendentalism. Nevertheless, Kane's driving intuition—that freedom requires self-formation, at least in some crucial respects—does seem enshrined in ordinary moral thought and practice.

At the very least, these ontologically demanding responses to the problem offer some abductive evidence that this is indeed a serious challenge to realists about freedom and responsibility.

Can the Strawsonian offer an adequate compatibilist conception of self-formation? Perhaps. Notice that the values-based view locates the source of our free actions in our values. But one needn't have this view about the location of the sources of free action. One could locate the source of our free actions in a different part of the pertinent sense of control. Frankfurt (1979, 2002), for instance, locates the genuine source of action in the motivational component, in our desires.

Recall that we offered on Strawson's behalf a kind of control that had three components: emotional, volitional, and intellectual. Strawson seems to suggest that the real source of our actions is the emotional component. Perhaps this is not surprising. Frankfurt and Strawson both have a Humean streak. Each (implicitly in Strawson's case) offers an actual sequence conception of responsible action. But Strawson did not offer a positive a "snapshot" metaphysics of agency to compliment it. In fact, his methodology seems to undermine it.

Must Strawson have this weakness? Let's focus on the affinity with Frankfurt again. Consider Frankfurt's positive Humean explanation of the sources of our action. For him, it is *our own* higher-order desires which the requisite springs of action (1969). Strawson, on the other hand, seems to think that it is our *proneness to moral attitudes and feelings* which are the springs of our actions (at least, the pertinent set of actions, the ones for which we can be held morally responsible).

Strawson contends that this proneness is part of our shared humanity. It is a natural commitment that we have, on a par with a commitment to induction as belief formation mechanism

(2008/1962: 35 n. 7). We can neither choose nor give up such commitments. He puts the point in his work later *Analysis and Metaphysics* in a particularly clear way:

Our proneness to moral attitudes and feelings is a natural fact, just as the sense of freedom is a natural fact. I have remarked that they are linked, and it is time to say more about the link. In speaking of the sense of freedom, I connected it closely with the sense of self. Our desires, decisions, actions are not in general felt as alien, as things that simply happen in or to us, like a pain or a blow. They are we. Our awareness of them is awareness of ourselves. I remarked that we attribute to others this same sense of freedom and this same sense of self. We see others as other selves, and are aware that they so see each other. But this is not a matter of a conclusion drawn by analogical reasoning. In a variety of ways, inextricably bound up with the facts of mutual human involvement and interaction, we *feel* towards each other as to other selves; and this variety is just the variety of moral and personal reactive attitudes and emotions which we experience towards others and which have their correlates in attitudes and emotions directed towards ourselves (1992, 137-138).

They are we! Powerful stuff. Here's the upshot. We might say that source-based pleas function only insofar as they show that there is something defective about a person's proneness to the attitudes and feelings which form our moral system, and which are the springs of moral action. (Hence, they are abnormal or are temporarily thwarted in acting normally).

In other words, our *shared* proneness blocks the generalization worry. To think that this kind of "special" excuse generalizes is to make a mistake, for it is to assume that we would all lack what as a matter of natural and normal development we are all prone to experience. And in turn, this proneness shapes our experience of others as free and responsible persons. As he says in "Freedom and Resentment", "this commitment is part of the general framework of human life, not something that can come up for review as particular cases can come up for review within this general framework" (2008/1962: 28).

Perhaps I spoke too quickly. This is only powerful stuff if you have compatibilist leanings. To some ears this will sound deeply troubling. For if we neither choose nor could give up these feelings and attitudes, in what sense can they ground our moral responsibility for particular actions? In what sense are they *our* feelings and attitudes? It seems now that we have a direct route to incompatibilism

rather than a nice way of blocking the generalization worry. We are not the source of the feelings and attitudes which constitute us. They are given to us by nature, or more locally by our formative circumstances. Indeed, is it not the determination of our choices by nature that drove Kant all the way to transcendental freedom, to a freedom that transcends the natural causal order?

### *3.4.c The Viewpoint of Human Affairs*

We have not made progress in blocking the source generalization worry. But we have at least cleared something up. The problem, it seems, is that Strawson wants to say that the source of our responsible actions are our common human attitudes and feelings. Since we are apparently not the sources of these human attitudes and feelings, though, this maneuver seems to lead to source incompatibilism.

Let's now consider these attitudes and feelings more closely. Strawson repeatedly emphasizes that the range of attitudes to which he attends in the paper is varied but specifically imbedded. They depend on the wide range of human interpersonal relationships: "as sharers of a common interest; as members of the same family; as colleagues; as friends; as lovers; as chance parties" (2008/1962: 22). They are also intimately connected. Reactive attitudes can be personal—resentment at the poor treatment of ourselves—vicarious—indignant at the poor treatment of others—and even directed at ourselves (2008/1962: 23). And these attitudes are not logically but rather humanly connected (2008/1962: 23). These attitudes "have common roots in our human nature and our membership of human communities (2008/1962: 23). Beyond this, Strawson makes the claim—*pace* breezy interpretations of his paper—that within this framework of human nature and human community our practices, there is "endless room for modification, redirection, criticism, and justification." (2008/1962: 35). Strawson elaborates on this point:

No doubt to some extent my own descriptions of human attitudes have reflected local and temporary features of our own culture. But an awareness of variety of forms should not prevent us from acknowledging also that in the absence of any forms of these attitudes it is doubtful whether we should have anything that we could find intelligible as a system of human relationships, as human society. (2008/1962:36)

In brief, Strawson brings to light the role that the moral sentiments play in the human form of life. We have moral attitudes and a human need to manifest them. The expression of such emotions will have local variation. No doubt this is explained by the variety of human moral communities and the different ways of relating therein.

Given this, it seems naïve to interpret reactive attitudes as mere dispositions to react emotionally in light of apparent ill will. For it is not simply our nature to be disposed to some range of attitudes. What matters is that these attitudes have cognitive significance within human communities. Why? These communities will set the norms for the expression of good will and ill will towards others. It is not just that we are naturally committed to these reactive attitudes. These attitudes are responses to the actions of others *understood to be* meaningful expressions of their inner attitudes. Reconstructing Strawson's view this way makes some sense of his bold claim that that "only by attending to this range of attitudes can we recover from the facts as we know them a sense of what we mean, i.e. of all we mean, when, speaking the language of morals, we speak of desert, responsibility, guilt, condemnation, and justice." (2008/1962: 34). For the emotions reveal the meanings of the language of morals.

It makes *some* sense of it. Strawson's view of the emotions now seems somewhat demanding! It seems natural here to invoke "thick" concepts. For perhaps our reactive attitudes are a mode of understanding the world in both descriptive and normative terms. Some dispute the division of concepts between "thin" and "thick", where "thick" concepts have inextricably linked normative and descriptive aspects.<sup>71</sup> But let's set these worries aside. For perhaps in light of the thick content of our reactive attitudes we can revise the practices and norms governing the appropriateness conditions for expressing such emotions by reflecting on our moral experience. We can recognize *that this action was*

---

<sup>71</sup> For an overview of this literature see Väyrynen (2017).

*resenable* without feeling resentment; we can question whether or not it is appropriate to feel resentment in any particular case, at all.

To develop Strawson's thought here we will have to offer a more fleshed out sense in which our emotional powers are connected to our powers of understanding. Consider the following view.<sup>72</sup> Say that at least some emotions involve a unified experience of both a cognitive and an affective element. To feel that something is *rude* is to feel offense at *what is offensive* in a given situation. And this, it seems, does not merely involve offense at some descriptive state of affairs. Rather, it involves an understanding that some of these facts express an attitude, perhaps some kind of lack of respect.<sup>73</sup> One is tempted to say here that to feel offense in this way is to have a non-moral reactive attitude. Now, to experience something as rude is one thing. But once the experience is over, we can coolly reflect on the content of that experience. Even beyond this, we can think about how we ought to respond to rudeness.

Let's say something similar about resentment. To experience resentment over something is not merely to respond to some set of descriptive facts. Rather, it is to respond to an interpretation of these facts as expressing a certain attitude: lack of sufficient good will (i.e., a lack of *moral* respect). We have expectations about the actions of others that are imbedded in social contexts. We understand some act-types to show good will or a lack thereof. That our actions mean something to others requires that actions be understood cognitively. Nevertheless, our first reactions to them are emotional in the form of reactive attitudes. This suggests that reactive attitudes involve cognitive content. Since they are bearers of cognitive content, we can abstract away from any particular experience of them and

---

<sup>72</sup> Here I draw on Linda Zagzebski's (2003) view. Rudeness is her example. The general picture I paint of abstracting away from the experience of the emotional response to abstract thinking about it mirrors her story of how this goes but does not go into all of the details of her particular account.

<sup>73</sup> See Foot (1978) as cited by Zagzebski.

reflect on their content. In doing so, we may revise our views about what we ought to do (and about what we are obligated to do).

I side with R. Jay Wallace (1994) on this point about the cognitive content of the reactive attitudes. It is very difficult to get the reactive attitudes to be all that Strawson wants them to be without going cognitivist. If you disagree, here is an easy amendment. We can say instead that the relevant cognitive content can be used in a normative assessment of the appropriateness or fittingness of reactive attitudes (as in D'Arms and Jacobson 2003).

Recall the point of this discussion was to find resources for Strawson. How does this view of emotions as bearers of thick ethical content help block the generalization strategy in terms of source freedom? Well, now we can say the following. It is inescapable that we feel reactive attitudes. And it is moreover inescapable that we have concerns about the good will of others, about justice, and about desert as embodied and understood by way of these moral emotions. Within the framework of commitments made manifest in our experience of these emotions, we see room for variation and modification. Thus, we might come to discover that some histories make persons inapt targets of emotions like resentment and indignation. That is to say we might realize better the structure of our commitments. Nevertheless, in these emotions our questions about the sources of our freedom bottom out, so to speak. They are the basic moral experiences from which we harvest material for moral reflection. We could not give up the entire emotional-conceptual standpoint. It is the standpoint of human affairs.<sup>74</sup>

The incompatibilist, however, can marshal a very plausible argument with the resources described above, since nothing said about the nature of reactive attitudes directly speaks to the matter

---

<sup>74</sup> For further discussion about this by Strawson see his (1985).

at hand when it comes to the generalization strategy, namely, getting from our practices of moral responsibility to a compatibilist-friendly account of agentive control.

Grant that we can abstract away from the “hot” experience of the emotions to reflect on their cognitive content. Now it seems the content of our emotions involves sensitivity to a metaphysical condition of freedom. Again, why do we feel such mixed emotions when confronted with someone was horribly abused as a child and then grew up to be a murderer? Why do the reactive attitudes seem less appropriate in these cases of tragic formative circumstance? It seems as such because a person’s formative history *explains* their ill will towards others. We can realize this by reflecting on the content of our experience. When we learn about a particularly tragic formative history, we feel pity, and perhaps a feeling that if we had been so formed, we too might have been radically different given those circumstances. So doesn’t anyone’s history explain the psychic sources of their quality of will? And doesn’t this kind of pleading share the same intuitive force as the intuitions found in manipulation arguments? So, the incompatibilist can assert that a certain class of pleas having to do with the sources of action involve a commitment to a metaphysical criterion on free and responsible action: an ultimacy condition that a free and responsible person must be the source of her own actions. Thus, if determinism were true, such pleas would generalize.

Again, the plausibility of this line of reasoning is *helped* by the view of emotions and moral knowledge described above. The more we look to Strawson's aid, the more harm we do to him. For we *feel* differently about persons who were not self-formed in pertinent ways. We *realize* the unfairness of holding persons responsible by reflecting on the content of our moral experiences. The incompatibilist's concern about generalized pleas arises *within* the standpoint of human affairs. It arises, it seems, from our basic moral experience.

### 3.4.d Gains, Losses, and Making Sense of Life

Having tried to offer Strawson a picture of the moral emotions that justifies everything he says about them, we have once again helped the source incompatibilist.

I now turn to something of last resort. Strawson raises a puzzling concern in “Freedom and Resentment” about the “gains and losses” that would come with revising the basic terms of human interpersonal engagement (2008/1962: 28). This would involve a radical shift in the basic terms which we understand what to do in the most basic practical sense. It would involve a radical shift in how we would see normal human persons. Strawson thinks that this kind of revision cannot be motivated by physical determinism. Perhaps there is something about the cognitive content of our reactive attitudes which explains why he thought so. And perhaps this has to do with the very terms by which we can even understand a human life as such.

For guidance about these questions, we might consider thinking about what happens to practical questions when a way of life ends. Jonathan Lear’s *Radical Hope: Ethics in the Face of Cultural Devastation* (2006) takes a critical look at the Crow worldview during the transition from nomadic life to life on the reservation. I cannot go into the full details of Lear’s rich account, but the general form of the problem he discusses will be quite instructive here. In particular, the book addresses a puzzle. The last Crow chief, Plenty Coups, recounted his life to Frank Linderman. Now, Plenty Coups was very active in life after the Crow moved to the reservation. But his account of his life to Linderman stops when the Crow moved to the reservation. “Nothing happened after that”. Far from offering a merely psychological account of this wild claim, Lear argues that there is a coherent sense in which Plenty Coups’s claim was literally true. For the central practical aims by which Crow persons understood what to do in life—hunt buffalo, engage in war with the Sioux, and so on—were no longer available to Crow persons on the reservation. It is not as if, Lear suggests, things stopped happening. For instance, it is not as if Crow persons stopped cooking meals. The problem is that these happenings

lost their practical significance in terms of what it would mean to live a good life (e.g., “I am making this meal to prepare the war party for a raid on the Sioux). So, in this sense, even though there were happenings, nothing really happened after the Crow moved to the reservations. The basic concerns of Crow life were unrealizable after that. Living was, therefore, unintelligible in a distinctly practical way.<sup>75</sup>

One can see how Strawson’s concerns about “gains and losses” might be taken in a likewise manner, but at a more fundamental level. Part of our basic concerns in human interpersonal life are that we treat each other with sufficient good will. And we understand—at least in part—that this basic aim involves understanding our actions in a practical sense. That practical sense is the meaning of our actions to others. We *aim* to treat others in a way that accords with good will towards them. Actions which fail to do so are defective insofar as we, as human persons, aim to live in moral community with others. And our disapproval of such actions—our very awareness that these actions are defective—is grounded in our experience of the reactive attitudes. Part of the point of blame on the Strawsonian picture is to communicate to others that we have taken their actions express this defect. It also has the important secondary effect of signaling to the wider community that we stand against such actions.<sup>76</sup> And in blaming, we aim to make the person feel that what they did was incorrect *by their own lights: we make them feel guilty*. Hence the distinctive sanction of resentment and indignation: other-induced self-corrective. This function relies on our shared aims and concerns as human persons. These shared aims involve the interconnectedness of other-regarding and self-regarding attitudes which Strawson explicates at length.<sup>77</sup>

---

<sup>75</sup> Lear goes on to offer an interesting virtue-theoretic account of how Plenty Coups solves this problem. This solution, however, is to the problem of the *actual* end of a particular human way of life. The current question here is, rather, under what conditions would a recognizably human way of life persist?

<sup>76</sup> Angela Smith (2012: 44) makes this point.

<sup>77</sup> This reading of Strawson’s “gains and losses” does his view more credit than glossing his turn of phrase in terms of utilitarian calculation.

Here is the point. Even if it were true that determinism would be grounds to revise our basic terms of interpersonal engagement, it would be practically irrational to do so. For the kind of revision required would make unrealizable the basic concerns by which we understand our actions. And these are the basic concerns and corresponding emotional responses which Strawson thinks we *cannot* give up. Again, they are we. Now we have moved to the last part of the sort of control implied by Strawson's picture of the moral life: the socially imbedded understanding of the meaning of our actions. It is here that we must locate the real source of our actions. Perhaps it is because we understand the meaning of our actions in terms of our human aims and concerns, we can rightly see ourselves as the *authors* of our lives.<sup>78</sup> This idea, at the very least, gestures at a source conception of control consistent with Strawson's picture of our moral practices.

There is a kind of radical revisionism that this point plausibly rules out. If we were to stop interpersonal engagement, our lives would not be recognizably human. Thus, revision is practically irrational but also practically inconceivable. But the problem with this way of motivating Strawson's big picture is just that he thinks interpersonal engagement requires reactive attitudes. It is possible for us to imagine things as being otherwise. To some, these alternatives will appear *better*. At least to some of us. Consider Pereboom's argument (2001, 2014, 2017) that human life could go on without moral responsibility (in the basic desert sense) and perhaps for the better. He argues that the meaningfulness of our life's projects and our personal relationships would remain largely intact. Relationships might be *better* constituted by human emotions of sadness, regret, and love without hard feelings like resentment and indignation. We might do better morally by abandoning systems of harsh sanction and angry blame. The point here is not that Pereboom's view is right. That claim would need much more than my altogether too brief recapitulation here. The point is instead that *there is a conceivable alternative*

---

<sup>78</sup> See McKenna (2008a) for a similar thought.

to our way of going about things. And at least to some, this alternative is not merely warranted on the grounds that determinism is true, but moreover, on the grounds that it would realize human interpersonal aims better. That this is arguable undercuts the move offered on Strawson's behalf that, somehow, the very content of our moral-emotional experience makes such thoughts practically inconceivable.

How might we draw on Strawson to respond? It is hard to see how without giving up a core commitment of Strawson's. To see why, consider a parallel problem and a parallel response: Martha Nussbaum's (1987) response to what she calls "the Marxist objection" to Aristotelian virtue ethics.

Say that a good life is a virtuous life, a life in which one manifests the virtues. Virtues, say, are ways of being excellent in shared and universal spheres of human activity. For instance, courage is just being excellent with regards to dealing with fear. There are, then, objective facts about what counts as being courageous. Nevertheless, we could have different competing theories of what courage is. These would be competing theories of what it is to be excellent with regards to dealing with fear. So, our ethics can be modified through reflection.

But the Marxist will object to this line of reasoning. She will argue that some ways of being virtuous, as commonly understood, do not depend on universal spheres of human activity. Generosity, for instance, requires the institution of private property. And although private property seems like something natural, perhaps something born out of our common humanity, it isn't. So virtues are, in fact, relative to human forms of life. Generosity is merely a bourgeois virtue. Given this one example, one might start to think that the virtues we take to be universal and objective are not so. There are no fixed points to be found in human nature upon which to ground ethical reflection.

Nussbaum rightly sees that the Marxist's question is "profound". "What circumstances of existence go on to define what it is to live the life of a *human being*, and not some other life? (1987: 49-50). In response, she points out that radical transformations in the way we live, like giving up private

property, often have a tragic dimension. We solve some problems only by removing others. If private property is a problem and so we abolish it, we lose out on certain kinds of freedom of choice. We also lose the possibility of one kind of fine action: generous action. Nevertheless, Nussbaum thinks there is a limit to these transformations. For every structure of life, she thinks, is limited in some way, closed off to some kinds of value.<sup>79</sup>

We might analogously concede the “profound” question to the hard determinist on Strawson’s behalf. Perhaps philosophers like Pereboom are right to point out that there would be benefits to giving up the reactive attitudes. But this would have a tragic dimension. It would make impossible certain ways of interpersonal engagement (cf. Shabo 2012). We might then say that the gains and losses to human life warrant preservation of life with reactive attitudes. And so, our belief in whatever kind and degree of control is implied by their appropriateness is warranted too.

There are, however, two problems with such a reply. For one, modelling a Strawsonian response on Nussbaum would be to concede something Strawson would not wish to concede. For Nussbaum grants the Marxist the openness of the profound question: what is a human life, in the relevant sense? For Strawson, determinism can play *no role* in how to answer such a question. Again, on his view, the reason why it can play no role is tied to the reactive attitudes. A system of human interpersonal life without reactive attitudes would not be recognizable as a human society, he says. But a hard determinist can disagree on the grounds that some recognizable human aims will be furthered by our relinquishing our commitment to these attitudes. Indeed, she may argue that the most important human aims of love and fellowship will be enhanced. And a libertarian can better reply to this hard determinist rejoinder, it seems, by denying the warrant of such a revision. She will argue that determinism is false.

---

<sup>79</sup> See her *Fragility of Goodness* (1986), esp. chapter 11.

Notice moreover that the libertarian about free will can agree with this modified Strawsonian reply. She can agree that the “profound” question about the nature of human life is open and countenance the tragic loss of the reactive attitudes. But she seems perhaps better able to articulate *why* this loss would be tragic in a way that pushes against the hard determinist. She can say that human persons have libertarian free will. *That* is the kind and degree of control implied by the appropriateness of reactive attitudes. Thus, she may insist that the gains and losses of the hard determinist revision are quite clear. We would be denying our very nature as free beings if we revised our practices by abolishing reactive attitudes. For humans have free will. That *power* is part of a recognizably human life. It is a normal condition of humanity.

### 3.5 Where Do We Go from Here?

I have argued that Strawson’s sentimentalist deontology, his appeal to natural commitments, his division between an objective viewpoint and a viewpoint of interpersonal affairs, and finally his appeal to gains and losses are of no help in blocking the source generalization worry. They only seem to make things worse! Is there anything left to say on Strawson’s behalf drawn from “Freedom and Resentment”?

Here is a thought. Strawson’s argument relies on our conceptual command over “the facts as we know them” about who is and who is not morally responsible for his or her actions (Strawson 2008/1962: 19). But we might depart from Strawson at this juncture. Strawson seems to think we could not glean any intelligible form of human life from a system that lacked reactive attitudes and the basic moral demands they express. We might instead, in Strawsonian spirit, reframe his point in terms of the moral-conceptual frameworks we would recognize as human, and our competence within those moral frameworks. And this is exactly what the final theme about gains and losses suggests.

Consider this skeleton of an argument. It is not competence within moral practices, construed in Strawsonian fashion, which makes universal exculpation by determinism “practically

inconceivable”. Rather, if determinism is true and universally exculpates, it means that we have been wrong, all along, about ordinary facts about ourselves and others. And these are the sorts of facts that, ordinarily, we can safely presuppose. To say that we have been wrong about these facts clashes with our given competence in making judgments about who is morally responsible and who is not. In other words, it clashes with our command over the facts as we know them. For instance, can we really believe that in every purported instance of *apparent* competence with the demands of morality is really just apparent? Maybe not. Notice that this offers a tantalizing possibility. If such an argument were successful, it could rule out the worry that we lacked the right kind of self-forming possibilities. For if we lacked them, we would not be competent in the ways in which we know that we are.

Now, this “facile” argument certainly needs some flesh on its bones! The main point here is not to defend it but to show how one might begin to formulate an argument that is *neutral* about competing conceptions the kind of control we need to be morally responsible, the thing which grounds our competence with respect to the demands of morality. The point is that we might adopt very different conceptions of control while marshalling a “Strawson-inspired” compatibilist argument. We could even give pride of place to conceptions of agency typically favored by those who begin, contra the project considered in this paper, with considerations of the metaphysics of agency.

My suggestion for Strawsonians, then, is this: we might block incompatibilism by a kind of self-assuredness in the very moral competence that determinism is alleged to threaten. Our moral competence, it seems, is a fact required of the form of human life we in fact participate in. For it seems required for the intelligibility of this form of life, to some of our basic aims and concerns. We can maintain this premise without offering a specific account of the control needed for moral responsibility. This control is, after all, the exercise of the moral competence we ordinarily expect from adult human persons.

### 3.6 Conclusion

I have argued that Strawson's "Freedom and Resentment" fails to deliver on its immodest goal. It cannot decisively defeat incompatibilism. For there are, it seems source incompatibilist intuitions enshrined in our responses to manipulation cases and tragic formative history. Attempts to build on the thematic elements of Strawson's view seem in each case only seemed to turn Strawson's own argument against him. I concluded by offering a suggestion. Those compatibilists who want to offer, as Strawson did, an argument from the facts of morality to compatibilist free will might best distance themselves from elements of Strawson's substantive view.

## Chapter 4

### On an Argument for Realism about Free Will

**Abstract:** Philosophers typically approach questions about free will and moral responsibility in terms of the compatibility question: is free will compatible with determinism? In this paper, I articulate an alternative question. Call it the existential question: is free will real? Does it exist? I argue that realists about free will—both compatibilists and libertarians—have offered a kind of *reductio* argument in favor of their view. But these arguments are neither well understood nor well developed. I offer what I take to be the best versions of this argument, which can remain neutral about the compatibility question. This kind of *reductio* argument places a powerful burden on skeptics of freedom and responsibility. They must show how it is practically conceivable that the normal condition for persons that they are not morally responsible for what they do.

#### 4.1 Introduction

The standard way of approaching the topic of free will proceeds in light of the compatibility question: is free will compatible with the thesis of physical determinism? This feels very natural. For the classic problem, it seems, is a modal one. Is it *possible* that there be a world where at least one person has free will and determinism is true at that world? The logical space neatly cleaves in two: compatibilism and incompatibilism. Naturally, then, the matter of whether or not persons are morally responsible comes after our metaphysics of agency, and not before. (Or, if it does come up, moral responsibility is a way of picking out the sort of agency of which we are in need of a metaphysical account. Substantive upshots about moral responsibility come after we have this account of agency in hand). Indeed, this ordering—metaphysics first, metaethics later—is how I learned about the problem of free will as an undergraduate. It is how I teach the problem to my students.

But a different way of starting to think about free will begins with an *existential* rather than a modal question. Is free will *real*? Does it *exist*?<sup>80</sup> I take realism to be the view that persons are free and morally responsible. Anti-realism (or nihilism) is the opposing view that free will does not exist: no

---

<sup>80</sup> See Kane (1996: 13) for a different way of approaching the question about compatibility and the question about existence.

person has free will. Compatibilists and libertarians are therefore both realists.<sup>81</sup> (You might have noticed that when you teach the free will in terms of the compatibility question, this agreement is often a source of confusion among students!) In fact, one could be a realist in this sense, but an agnostic about the question of whether or not free will is compatible with determinism.<sup>82</sup> But this division also feels natural once we consider a different set of concerns. Both compatibilists and libertarians maintain that persons are accountable to the demands of morality. And by approaching questions about free will by beginning rather than ending with considerations of moral responsibility, this shared commitment becomes more perspicuous. This alternative approach, I will argue, is a boon to realism.

As the alternative approach begins by directly theorizing about moral responsibility, it has recently been associated with the work of P.F. Strawson (2008/1962). For this reason, it is thought to be amenable to compatibilism. But this approach needn't be. Indeed, we might make theoretical progress on the existential question before anything about the compatibility question has even come up. And I think such progress can be made.

In this paper I discuss a kind of *reductio* argument for realism, motivated by commonsense evidence about our shared moral experience. This kind of argument concludes that there is a kind of absurdity in thinking that no one is morally responsible, and so we must have the kind of control needed to be morally responsible. In other words, the argument suggests that it is absurd to think we lack the pertinent kind of control. And even if we disagree about what kind of control the argument

---

<sup>81</sup> The logical space is a bit more complicated. Although compatibilists about free will and determinism are typically realists, there are many different compatibility questions one might ask. Imagine a compatibilist of the sort who says that free will *requires* determinism, and so who thinks that indeterminism is incompatible with free will. Now imagine that they also think indeterminism obtains in the actual world. This person would not be a realist. Generally, thinking in terms of the compatibility question can fail to align with the logical space afforded by the existential question.

<sup>82</sup> For instance, see Al Mele's well known agnosticism in (1995) and (2006).

implicates and whether or not it is compatible with determinism, this control is a plausible candidate for what we mean when we talk about free will.

I claim, however, that the force of such arguments has not been well understood, even by those who offer them. In particular, I develop the kind of argument in terms of the generalization that persons are morally responsible. This claim is, if anything, “a platitude— something only a *philosopher* would dream of denying.” (Lewis 1983: 166, emphasis added).

#### 4.2 Realism, Commonplaces, and Compatibility

The argument I develop and discuss will look like Strawson’s argument in “Freedom and Resentment” to some readers. In that (in)famous paper, he offers what he admits is a “facile” argument that determinism is compatible with moral responsibility precisely because that thesis has no import on whether or not persons can be aptly held morally responsible for their actions. His argument is a provocative *reductio*, that it would be “practically inconceivable” that determinism “would or should” make us abandon the moral attitudes and feelings which constitute our practices of moral accountability (2008/1962: 25). It is important that he introduces his argument by stating that what he has to say mostly consists of commonplaces. (2008/1962: 22). Strawson sides with those who are optimistic about freedom and responsibility (even if his position is radically unlike his optimistic utilitarian contemporaries).<sup>83</sup> And an optimist is, as he puts it, someone who takes the facts as we know them to provide adequate support for our ordinary moral thought and practice, which the pessimists feel imperiled by determinism (2008/1962: 19). Strawson’s argument suggests that the kind of control needed to be accountably held to the demands of morality is compatible with determinism.

Strawson’s main concern, notice, is not at first the compatibility question, but rather an adequate account of the commonplaces of our responsibility practices. Indeed, he admits that he does not even

---

<sup>83</sup> E.g., Schlick (1939) and (Ayer 1954)

know what the thesis of determinism amounts to (2008/1962: 19). And he is apparently not primarily concerned with arguing against anti-realism about freedom and responsibility. Instead, his main aim is to reconcile two kinds of realists, who are having a confused tussle over the compatibility question. And he seems to think commonplaces can settle that dispute.

Strawson's argument is much maligned.<sup>84</sup> Nevertheless, many have thought this sort of argument to be both novel and promising in favor of compatibilism, even if Strawson's version of it is not successful.<sup>85</sup> But his argument is indicative of a broader set of concerns and arguments which have historically motivated what I am calling realism. These arguments have a strikingly similar structure. They are all somewhat transcendental in flavor. And they are all, it seems, variations on a kind of *reductio*. The absurdity reached in these arguments is less-than-strictly logical. Instead, the absurdity has a specifically practical dimension.

Consider one historical example for comparison. Aquinas answers the question of whether or not human persons have free will in the affirmative. He writes, "I answer that, Man has free-will: otherwise counsels, exhortations, commands, prohibitions, rewards, and punishments would be in vain" (I. 83: 1). One might see this as a single sentence *reductio*! It seems impossible that all of our moral goings-on would be in vain! For these goings on are commonplaces. Persons are, of course, morally responsible for what they do. And so, of course, persons have free will. So, Aquinas seems to answer the existential question in the affirmative by starting with the apparent appropriateness of holding one another accountable, by pointing to our responsiveness to moral demands. Indeed, he seems to reason thusly: we are competent with respect to the demands of morality, since it is absurd to imagine that we are not. We therefore must have whatever sorts of abilities are required to be so

---

<sup>84</sup> See Wiggins (1973), Bennett (1980), Watson (1987), Russell (1992), Pereboom (2001) and (2014), and Todd (2016)

<sup>85</sup> See, for instance, R. Jay Wallace's compelling (1994) development of the view.

competent. (At least, someone could be plausibly inspired by Aquinas here to suggest this line of reasoning).

Yet Aquinas is (arguably) a libertarian. Why? Well, the commonplaces that Aquinas points to are what *needs to be explained* by a theory of the special kind of competence—skills, dispositions, and abilities—that makes us apt recipients of counsels, exhortations, commands, prohibitions, rewards, and punishments. Free will is invoked to explain this special moral competence of human persons. Well, what sort of free will can underpin all the ordinary moral goings-on? Aquinas's suggestion is that (glossed in more contemporary terms) a free agent's will must be guided by her intellect. Free agents have a special power to act in response to reasons (I. 83: 2-4). As Eleonore Stump (2003) has argued, Aquinas takes the intellect in question must be the agent's own intellect, and so the agent is the ultimate source of her own actions in a libertarian fashion. This, she takes it, is incompatible with compatibilism, for it entails that free acts must originate from causal chains that are internal to an agent's intellect and not extrinsic to it. And this seems incompatible with the thesis of physical determinism.

Here's another important instance. (Or, at least, an instance inspired by another historical figure). On one plausible reading, Kant's idea of transcendental freedom is meant to ensure that agents can believe without inconsistency that they are morally responsible for their actions even though nature is deterministic. Free actions have to be produced *by the agents themselves*, rather than prior physical causes (Pereboom 2006: 541-542). But since *we must* do what the moral law requires on Kant's view, it follows that we must take ourselves free so that our actions may be properly imputable to us. Indeed, it seems that for Kant that persons are bound to the moral law is constitutive of what it is to be a person. Kant defines actions that are properly imputable to agents as only those acts which are not brought about by prior physical causes (A533/B561; A541/B56). Therefore, we must believe that

we and all free agents are outside of the deterministic causal nexus of nature. Hence, we have transcendental freedom.

Consider a contemporary instance of the phenomena. Peter van Inwagen (1983, 2017, chap. 12, *inter alia*) marshals these kinds of concerns when he explains his own belief in free will. As he puts it (1983: 209): “without free will, we should never be morally responsible for anything; and we are sometimes morally responsible.” But if, as he argues, moral responsibility requires alternative possibilities, and deliberation only makes sense in light of open alternatives, then moral deliberation in particular seems totally and absurdly incomprehensible in a world with only one physically possible future. This might be taken to explain why we have such terribly strong feelings of free choice. The alternative is, one might say, unthinkable.

In each of the four instances discussed, we see a similar structure. It would be absurd to think we do not have free will (or freedom in the pertinent sense), since it would be absurd to think we are not morally responsible. Nevertheless, there are disagreements. Interestingly, these are not exhausted by disagreements about the compatibility question. Unlike Aquinas, for instance, van Inwagen thinks we cannot come up with an adequate theory of the abilities or capacities that underpin our apparent competence as free moral deliberators. (At the very least, he cannot see, as of now, how we are going to manage it). He is a mystifier about free will (2000, 2002). He argues that both determinism and indeterminism are hostile to free will; nevertheless, he remains committed to the reality of freedom and responsibility.<sup>86</sup>

This is striking. It suggests that the kind of arguments for realism we have been all too briefly discussing can proceed *without* a particular metaphysical conception of free will to underpin them.

---

<sup>86</sup> Indeed, he worries (2017, chap. 13) that there is no folk concept “free will”, and that “free will” is a problematic philosophical term. He goes so far as to suggest that much work on free will is simply irrelevant to what he now calls the Culpability Problem, which is a set of apparently inconsistent theses about abilities, culpability, determinism, and indeterminism.

They are arguments firstly for *realism*. The upshots of these arguments are debatable. Evidently, Aquinas and van Inwagen would disagree about what we can learn from these arguments.

Let's return to Strawson for a moment. The careful reader of "Freedom and Resentment" will notice that the phrase "free will" is suspiciously absent from the text. Strawson, like van Inwagen, does not offer a positive theory of free will. Both have come under criticism for this absence.<sup>87</sup> If one starts with the compatibility question in mind, two stranger companions you will never see! But if one starts with the question of realism in mind, then these striking convergences are not only apparent, but perhaps even understandable. Both seem to agree that commonplaces support realism. And neither seems to think commonplaces offer sufficient grounds for any *particular* theory of free will. Both nevertheless think we can learn something about the nature of our freedom, for instance, whether or not it is compatible with the truth of physical determinism.

I will not wade into this debate as to what, exactly, this kind of *reductio* argument reveals about the compatibility question. Compatibilists and libertarians differ, it seems, as to whether or not commonplaces require "somewhat far-reaching assumptions about the self" (Chisholm 1989: 5). But what is clear from the proceeding is that such *reductio* arguments are arguments for realism. And all I wish to do here is get a well worked out schema of these arguments on the table. In giving this schema, my aim is not to faithfully reconstruct any particular version of such arguments. In this case, I doubt that the devil is in the details.

The *reductio* assumption of the argument schema I suggest is this: that no person is morally responsible—independent of any particular skeptical challenge. My reason for doing so is that the sorts of arguments I have discussed above have arisen in different philosophical contexts. They were

---

<sup>87</sup> See Wiggins (1973) and Bennett (1980) for this criticism of Strawson, and Wallace (1994) for an attempt to fill in the absence. For criticism of van Inwagen see Ekstrom (2003). Shabo (2011) defends van Inwagen's mysterianism.

conceived of as responses to different kinds of challenges to realism. This suggests that the arguments are *general* arguments against anti-realism about free will.<sup>88</sup>

### 4.3 The Form of The Argument

Here then, is the form of the argument, in brief: it is a pretheoretical truism, and in everyday usage a platitude, that persons are morally responsible for what they do. Persons are the fitting or appropriate target of moral demands. Why? Persons can culpably meet those demands. This requires a special kind of competence on the part of persons that makes holding them to these moral demands fitting or appropriate. In short, persons are morally responsible agents, who have a capacity for morally responsible agency.<sup>89</sup>

Of course, though, not all persons are morally responsible agents. Some people are abnormal in one way or another. Some persons lack (or otherwise are in a state such that they cannot exercise) the capacities that make them the inappropriate target of moral demands: the very young, the grievously mentally ill, persons with severe enough mental handicaps, and so on. (This has nothing to do with their value *qua* persons, which is as much as anybody else's as far as I'm concerned. And, of course, these persons deserve the full moral consideration owed to persons.) Instead, the abnormality of these persons *qua* moral agent, given the discussion at hand, has to do with their competency as moral agents, and thereby, what we can expect of them.<sup>90</sup> The demands of morality will tell us the sorts of things that would count as such an incapacitation: anyone competent enough to meet them must meet them, anyone who lacks this competency does not. In short, *normal* persons have a special

---

<sup>88</sup> The careful reader of "Freedom and Resentment" will notice, for instance, that Strawson admits to not knowing what the thesis of physical determinism even amounts to. He even asserts that the pertinent question about the appropriateness of responsibility can be answered without even knowing what it is (2008/1962: 25-26).

<sup>89</sup> Strawson's own emphasis on the reactive attitudes is just one way of pointing this fact out. Implicit in our responses to others is this very fact of another person's capacities.

<sup>90</sup> Some who are relevantly incapacitated are moral agents, i.e., agents who can perform right or wrong actions, even if they are not morally responsible agents, i.e., agents who can culpably perform right or wrong actions (cf. McKenna 2012).

kind of competence that makes them the fitting or appropriate target of moral demands, a capacity for morally responsible agency. It's just that not every person is normal in this sense.

Moreover, of course, not everyone who is normal in the relevant sense, vis-à-vis the demands of morality, is morally responsible in every particular instance. Sometimes people cannot be held to morality's standards because certain local conditions obtain: they trip, they lack the financial resources, the physical strength, and so on (cf. Strawson 1992: 136-137). When these local conditions obtain, we can say that even though these normal persons are morally responsible insofar as they are competent moral agents, they aren't in some specific range of cases. Something has interfered with their competence in meeting the demands of morality. These somethings are mitigating exceptions. They aren't exceptions to the claims that *this* person or *that* person or *all* persons are morally responsible. They are exceptions to particular connections between the moral norms and some particular person in question. Some local condition obtains such that *this* person isn't to be held to *those* standards. We could fill this out in many ways. Perhaps it would be *unfair* to hold persons to the standards when these local conditions obtain (cf. Wallace 1994). But the particulars do not matter. What matters is that the demands of morality themselves will tell us why these persons are not to be expected to meet them in these kinds of conditions.

Now, if no normal person is morally responsible. In other words, persons would either be exempted or excused.<sup>91</sup> If everyone were exempted, it must be true that everyone was (in the pertinent sense) abnormal. We would all be incapacitated in the relevant way so as not to be morally responsible agents. We would lack whatever competence it was that made normal people culpable for meeting the demands of morality. No one, it turns out, has the right kind of competency. So, no one is a normal person, in the pertinent sense. But isn't this absurd? For we do, in fact, hold some persons morally

---

<sup>91</sup> I exclude justifications here for simplicity. The argument can be adapted to fit them too, along similar lines to the case of excuses. I take the categories of exemption and excuse to be otherwise exhaustive.

responsible because they seem to be fitting or apt targets of the demands of morality. So, we do, in fact, take some people to be normal persons. Some of us, in fact, have the kind of competency that makes us fit to be held to the demands of morality. Now, how could we be wrong about that? For it would mean that no one would be the sorts of being we take them to be. Ourselves included! If this is true, it doesn't just follow that no individual person is morally responsible, it follows that no one is a normal person.

If, on the other hand, everyone was excused, it must be that in every case some local condition obtained such that we could not be held culpably to the demands of morality. So, even though everyone was (in the pertinent sense) a normal person, no one (in the pertinent sense) ever acted like one. In any token instance of *apparent* competence with respect to the demands of morality, no one ever is actually competent. But, how could we be wrong about that fact? It's not as if, in every instance, *really*, we were pushed, or lacked the finances, or the muscle strength, or... and so on. Every instance of apparent competence would have to have been an illusion, all along. Or maybe we are missing some critical piece of information that would change how we understood these things. But how could we be wrong about each of our own situations, to this degree? It is absurd to think that everyone, including ourselves, has been systematically wrong all along.

So, we have all being going along just fine thinking that persons are subject to the demands of morality in virtue of a special kind of competency. It is absurd to think that no one is, was, or will be the sorts of beings we take ourselves to be as human persons. Why? It seems absurd to deny that everybody lacks (or never successfully exercises) the special competence that makes normal human persons morally responsible agents. Such a denial flies in the face of the facts of ordinary, day-to-day living, where some competent people can undoubtedly (and sometimes unfortunately fail to) meet the demands of morality. And since we have reached absurdity, something must be wrong with the idea that no person is morally responsible. But if there is something wrong with that idea, then there is

something wrong with anti-realism about free will. All this talk of special competence, after all, seems to be a way of picking out what we mean when we say that persons have free will.

This kind of argument seems to impose a strong burden on the anti-realist. The anti-realist must explain why we are wrong about the special competence in light of which we understand the kind “person” in the platitudinous sense.

#### 4.4 Getting Precise About Practical Absurdity

Some readers will have found the foregoing argument deeply intuitive. Others will surely have not and may even find it silly. To make things more perspicuous, here is the argument schema in premise/conclusion form:

1. Assume for *reductio* that no person is morally responsible.
2. If no person is morally responsible, it is either because everyone is exempted, or everyone is excused.
3. It is practically absurd to think that everyone is exempted, for this would mean that no normal persons have the special competence that makes them fitting targets of the demands of morality.
4. Likewise, it is practically absurd to think that everyone is excused, for this would mean that no normal person in any particular instance ever exercised their competence with respect to the demands of morality.
5. So, it is practically absurd to think that no person is morally responsible.
6. At least some persons are morally responsible.
7. Since moral responsibility requires this special competence, and what it takes to have this competence is a plausibly free will, anti-realism about free will is false.

Now, if anything like this is going to work, we will need to say more about the notion of practical absurdity. (We will also have to say more about and why the argument appeals to normality and what this special competence is —more on this in a bit).

There is of course a sense in which it is both conceivable and possible for everyone to be impaired such that no one was morally responsible. Perhaps an evil scientist will release a toxic gas into the atmosphere that turns us all into zombies. There is a sense in which this scenario is not absurd. And the realist should not deny this. She should just assert that there is a special sense of *practical* absurdity manifest in the case. What is this sense?

One possibility is that the practical absurdity generated in the argument just falls out of our deep human commitment to seeing one another as morally responsible persons. Let's say a deep commitment is a commitment that we cannot help but believe and cannot reasonably give up. Perhaps our commitment to induction or our commitment to the existence of other minds fall into this category (as Strawson 1985 suggests). There is something absurd, it seems, to deny these commitments. It is not totally implausible to think that we have a deep commitment to moral responsibility, and so free will. As van Inwagen (2017: 90) asserts the credo: "free will undeniably exists."

There are at least two problems with the deep commitment construal of practical absurdity. The first is that it simply makes the *reductio* argument under discussion irrelevant. We could marshal a more straightforward argument for realism about both free will and responsibility by employing the idea of deep commitments. Consider this straightforward argument:

1. We have some commitments which we are so deeply held that we cannot reject them.
2. Any argument that attempts to show that we can reject them are ultimately idle.
3. Our commitment to free will and moral responsibility is one such deep commitment.
4. So, arguments against free will and moral responsibility are idle.<sup>92</sup>

This argument is surely controversial. But before even getting to that one might ask: if we can just go for an argument like this, why bother with excuses and exemption? Why worry about the potential futility of counsels, exhortations, commands, prohibitions, rewards, and punishments?

Perhaps we should bother with the more complicated *reductio* because the appeal to deep commitments is in itself problematic. Such arguments do not seem to support realism directly, but rather, the conceptual indispensability of some of our commitments. Perhaps conceptual

---

<sup>92</sup> Strawson (1985) seems to offer a version of this argument.

indispensability speaks in favor of realism, but perhaps not.<sup>93</sup> Another worry is that such arguments do little to advance the debate between those who believe in freedom and responsibility and those who do not. Anti-realists about free will just reject premise 3 without an independent argument. Indeed, that many philosophers reject realism about free will seems to be an existence proof that this premise is false.

Instead of the deep commitment construal of practical absurdity, I suggest that the practical absurdity should be cast in terms of “practical inconceivability”, as in Strawson (2008/1962). Let’s start with a conception of inconceivability. We shouldn’t commit ourselves to anything since all we are after is a schema for an argument, but we need something particular to work this. So, let’s say, following Yablo (2004), that something *p* is inconceivable if we cannot imagine any world that we don’t take to falsify *p*.<sup>94</sup> Now I agree with Strawson that it is conceivable that no person is morally responsible (2008/1962: 26). Again, we can perfectly well imagine a world where an evil scientist will release a toxic gas into the atmosphere that turns us all into zombies. Indeed, we should not wade into debates here about whether or not conceivability entails possibility or anything of the like. So, we need to think about how to introduce and be precise about a practicality metric over the space of possible worlds.

How should we get precise about that? For one, zombie world isn’t the sort of world the anti-realist thinks we are in. In the relevant free will and moral responsibility defeating cases (fate, determinism, etc.), we are experiencing presupposition failure. The capacities we presuppose in ordinarily treating each other as morally responsible are in fact not there, in spite of appearances. So,

---

<sup>93</sup> The indispensability of math to our best scientific theories was famously taken by Quine (19xx) and Putnam (19xx) to speak in favor of ontologically committing ourselves in that domain. Some fictionalists, e.g., Yablo (2005) have argued that indispensability is a mark of pretense.

<sup>94</sup> This is equivalent to the following: “for every world we can imagine, I take that world not to verify *p*.” It should not be read as follows: “for every world we can imagine, we do not take that world to verify *p*.” This states a non-conceivability criteria. (Yablo 2005: ft. 60).

let's try another case. We can perfectly well imagine a world where, all along, everyone has been mind-controlled by Martians. Let's stipulate that this robs us of our free will. In this world we were, all along, misled about what was actually happening vis-à-vis our capacities and our exercise of them.<sup>95</sup> Perhaps these Martians are in principle empirically discoverable, but at present the persons at those worlds lack the scientific tools to discover them (van Inwagen 1983: 109).

Now, even if this Martian world is conceivable, I claim that it is practically inconceivable. Why? In this world, the normal case is that we are not free and responsible. But it is this possibility, that the normal condition is one in which persons are not free and responsible, the way things have been, as it were, all along, which really seems unthinkable. When we imagine Martian world, we cannot help but see it, as compared the (apparent) features of the actual world, as one in which everyone is *abnormal*. And this itself is an assertion of a kind of empirical fact: normally, people have the special competence that makes them the appropriate target of the demands of morality. In giving the *reductio*, the realist is asserting that for every practically imaginable world, worlds where our practical aims as such are imaginable at that world, it is not a world where persons are normally non-responsible. So, that persons are normally non-responsible is practically inconceivable.

As Lewis discusses our epistemic situation with respect to the space of possible worlds: "We do not find out by observation what possibilities there are...What we do find out by observation is what possibilities *we* are: which worlds may be ours, which of their inhabitants may be ourselves" (1986: 112). The idea here is that we do know about our practical lives, we know what is normal for human persons by observation. Indeed, we know it by the most intimate observation possible, by living human lives. And so, by such observation, we know what worlds we are cannot conceivably be a part of.

---

<sup>95</sup> There is a difference between being ignorant and being misled when it comes to such matters. Strawson (1992: 138) makes this distinction.

Normality is, of course, a tricky thing. Normalcy seems to have both descriptive and prescriptive features (Bear and Knobe 2016). This is precisely the point. Our ordinary practical aims hinge on our making a distinction between the normal case, where we presuppose that persons are free and responsible, and abnormal cases where we exempt or excuse. The normal and the abnormal case each have both descriptive and prescriptive elements. For there is the descriptive matter of our having (or not having) certain relevant capacities and the normative upshot that the presence (or absence, or disfunction, or masking of, or...and so on) of those capacities makes us the appropriate target (or not the appropriate target) of moral demands. Any world where this isn't the case renders at least some of our core practical aims unimaginable within the context of that world.

The anti-realist will surely protest here. I will return to her protests in section 8. For now, we can just start with the claim that persons are just those beings that normally are morally responsible. When we claim that “persons are morally responsible” we are not making a universally quantified claim. Nevertheless, it is a modal claim about persons and what they are capable of. The claim furthermore admits of exceptions without becoming false. For sometimes persons are not morally responsible, either because they are not normal persons, or because in some particular instance they are not morally responsible for some particular thing.

In the next section, I will explain these features of the claim by appealing to the semantics of *generic generalizations*. This may not be the only way to do so, and I'll briefly address some alternatives. By appealing to generics and generic thinking, I think we can get a clearer sense of what is so practically absurd, construed in terms of practical inconceivability, in thinking that we could be in a world where normally persons are not morally responsible.

#### **4.5 Understanding the Generalization**

There are plenty of true claims about persons like this: “persons feel pain”, “persons have desires”, and “persons are morally responsible.” These claims are akin to claims like “leopards have spots.”

Claims of this form share the three unique features listed above (perhaps among others). First, generic generalizations are not quantificational. Claims of the form “all/some/most leopards have spots” contain information about the number of the kind that has the property, as do claims containing adverbs of quantification like “usually” or “mostly” (Lewis 1983). Claims like “leopards have spots” contain no such information. Secondly, generic generalizations are modal claims. Even if by some bizarre happenstance every actual leopard was albino, the claim “leopards have spots” would still be true; conversely, even if by bizarre happenstance every lawyer in the world was right-handed the generic “doctors are right handed” would be false (cf. Nickel 2016: 17). Part of what it is to be a leopard is to be spotted; lacking spots indicates that something irregular has occurred. On the other hand, there is no requirement that doctors be right handed. Finally, generic generalizations admit of exceptions. Not all leopards have spots, after all. Nevertheless, “leopards have spots” is true. So generic generalizations do not specify a set of necessary and sufficient conditions for kind membership.

A widespread, although not universal, way of thinking about the semantics for generic generalizations grounds their truth in ways of being normal for a kind (Asher and Morreau 1995; Pelletier and Asher 1997; Nickel 2008, 2016).<sup>96</sup> To eschew formalism, the idea is this: a generic generalization about some kind is true in virtue of a way of being normal for that kind.<sup>97</sup> Intuitively,

---

<sup>96</sup> Although I believe normalcy-based approaches are the most promising semantics for generic generalizations, there are other accounts that deserve consideration. I will not canvass all of them here. The argument I offer below is applicable given a wide range of them. For instance, I believe the argument is compatible with a domain-restriction approach to generic generalizations rather than the normalcy-based approach, where one understands generics as quantified over a contextually determined set of relevant individuals (cf. Schubert and Pelletier 1987; Declerk 1991; Chierchia 1995). I am skeptical that one particular account is compatible with the argument: Cohen’s (1996, 1999, 2004) probabilistic approach. This approach has come under criticism by purported counterexample (Leslie 2007, 2008; Nickel 2012), so the success of this kind of account remains to be seen. Strawson’s version of this sort of argument, it seems, did appeal to statistically normalcy, and for this reason, is subject to some serious concerns (see Bennett 1980).

<sup>97</sup> For instance, on Asher, Morreau, and Pelletier’s semantics, roughly, *K’s are F* is true if and only if for each individual *K*, the most normal worlds for that *K* are such that *K* is *F*. Normal worlds are determined by context. On Nickel’s recent (2016) semantics, *K’s are F* is true if and only if for every normal *K*, that *K* is *F*.

ways of being normal allow us to separate out exceptions to the generic that would falsify the generic and exceptions that do not. Since it is, plausibly, normal for leopards to be spotted, the existence of albino leopards does not falsify the generic because they are not normal. Thus, they do not matter for the truth conditions of the generic claim. This remains true even when every actual leopard is in fact albino. Generics are shown false by exception only when the exceptions come from normal members of a kind or class, not from abnormal ones (Nickel 2016: 52-56).

So, let's say that "persons are morally responsible" is a generic generalization, a claim about what constitutes a *normal condition* for persons, as seen in the argument schema above. Two important points follow from this. First, exemptions from responsibility are explicitly about abnormal persons. On the sort of semantics discussed above, these persons are not exceptions to the sorts of generic claims that characterize normal persons. Even though some people are not morally responsible that does not mean that the claim "persons are morally responsible" is false. Even if everyone were abnormal, this would not count against the generic claim that persons are morally responsible. For it would mean that everyone lacked the capacity needed for the special competence persons have in matters moral. This does not entail that persons are not morally responsible.

Second, excuses would not count as exceptions to the generic *even if* everyone was actually always excused in their actions. If everyone were excused, the generic "persons are excused from responsibility" would not therefore become true, for it would be an accidental feature of all persons that they were never responsible for what they did. Persons *can be* excused from responsibility. Nothing about that falsifies the generic claim or makes true something other claim about persons. In fact, the possibility of a world where everyone who is morally responsible is excused in each particular instance *relies* on the generic claim that people are responsible for what they do. For that claim is a modal claim about persons. So, even if in every token instance we were wrong about the local conditions of action, persons would still be morally responsible.

This feature of the truth-theoretic properties of generic generalizations gives us insight into the particular practical absurdity claimed in the argument schema. Universal exculpation is practically inconceivable because we cannot imagine any world that doesn't falsify the claim no persons are morally responsible, in this generic sense. Nothing obviously follows about the *normal conditions for kinds of things we are with respect to the demands of morality* from the conceivability of worlds where responsibility defeating scenarios obtain.

One might capture the core idea about persons here in ways akin to *ceteris paribus* laws of nature in the special sciences. *Ceteris paribus*, persons are morally responsible.<sup>98</sup> Another way to capture the idea would be to develop the argument in a neo-Aristotelian naturalist fashion, appealing to unquantifiable "Aristotelian categoricals" about the human form of life (cf. Foot 2001). This will amount to changing premises 3 and 4 in the schema.

I prefer the generics approach. It is neutral with respect to substantive claims about the nature of human persons. By my lights it also makes more perspicuous the *logical* features of the central premise in the argument, which allow for a clearer reconstruction of the sense in which it is practically absurd that no person is morally responsible. But this is admittedly debatable, and other ways developing the central idea of the type of argument in question are worth exploring.

#### 4.6 What Makes the Generalization True?

When we say that "persons are morally responsible" is a generic generalization, a claim about what constitutes a *normal condition* for persons, we have to specify the underlying truthmaker for the claim. Why is it the case that, normally, persons are morally responsible? As articulated in the argument schema, it is because, normally, persons have a kind of special competence with matters moral. They have a capacity (or some set of capacities) in virtue of which they have a special kind of agency, morally

---

<sup>98</sup> Thanks to Terry Horgan for this suggestion and helpful discussion on the possibility space.

responsible agency, which distinguishes them from moral patients and even moral agents (who can act morally but are not accountable for what they do).<sup>99</sup>

What kind of capacity do normal persons have so as to be fitting targets of moral demands? The argument suggests two desiderata. First, the capacity should be capable of underwriting the ordinary facts of life, that some persons are morally accountable for what they do. Any proposed capacity that cannot do this work should be ruled out. Second, the capacity should not be uncharacteristic of persons. By stipulation, normal persons are those with the right kind of competence to be held to moral standards. Therefore, the capacity should not be anomalous. I will consider two plausible candidates.

One venerable tradition in the theory of agency suggests that, in broad strokes, the relevant capacity involves the power to recognize pertinent considerations in favor of a particular choice and the power to respond aptly to those considerations. Typically, these considerations are taken to be reasons. Call these our reasons-responsive capacities. There is plenty of disagreement as to the details.<sup>100</sup> One might be tempted to associate this tradition with compatibilism, but central aspects of this view show up in versions of libertarianism too.<sup>101</sup> A different tradition suggests that, in broad strokes, the relevant capacity involves the power to understand and evaluate the emotions, motivations, and desires of other agents, and to be moved to respond to these forces. There is of

---

<sup>99</sup> Strawson's particular argument in "Freedom and Resentment" seems to lack this ingredient. This is the source of much of the criticism of that article. See Wiggins (1973), Bennett (1980), Watson (1987), Russell (1992), and McKenna (1998).

<sup>100</sup> Some suggest that these reasons-responsive powers be located in mechanisms of action (Fischer and Ravizza 1998), whereas others suggest that these powers be of the agent (McKenna 2013). Some suggest that this power is a one way-power, offering us guidance over our actions without requiring of us that we be able to do otherwise (Fischer and Ravizza 1998, Sartorio 2016). Others suggest that the power is a two-way power, exercisable in more than one way at the moment of action (Wolf 1990, Nelkin 2011, Vihvelin 2013).

<sup>101</sup> O'Connor (2009: 213) thinks that reasons *structure* our libertarian agent-causal powers. Nozick (1981: 294-316) suggests that we have a power to *weigh* reasons for action free of prior determination. Many incompatibilists suggest that there is indeterminism in the deliberative process (e.g., Kane 1996, Mele 2008, Balaguer 2010).

course disagreement about which forces are relevant.<sup>102</sup> There is disagreement about what the capacity is too.<sup>103</sup> We might nevertheless gloss the needed capacity as one allowing for the sorts of interpersonal relationships that adult human persons have. We might then go on to think of the demands of morality as made appropriate by and within the bounds of these relationships. Again, this view does not is neutral with respect to the compatibility question, for it might be that what is required of this capacity is really libertarian free will.

Why are these candidate capacities plausible? They are arguably (and believed by many) to underpin the ordinary practices, beliefs, and feelings appealed to in the argument schema. Moreover, they are not anomalous, but paradigmatic capacities of persons. And I should note that these are not competing suggestions. It could be that what underwrites the generic claim that persons are morally responsible is in fact a complex of many such capacities, which normally rise and fall together. They are the sorts of abilities and capacities that philosophers interested in freedom and responsibility take to be at issue in the free will debate. There are, in principle, compatibilist and libertarian versions of each.

#### 4.7 Particular Versions of the Argument

I will now offer one particular version of the *reductio* to illustrate how to fill in the schema. Let's say that persons are apparently morally responsible in virtue of their reasons-responsive capacities. Now, assume that no person is morally responsible. (It was all only apparent). Everyone is therefore either exempted or excused.

If they are exempted, it means that everyone is incapacitated by way of abnormality: they lack the pertinent capacity needed to be the fitting target of the demands of morality. In other words,

---

<sup>102</sup> It might be quality of will (Strawson 2008/1962), or the kinds the judgment sensitive attitudes they hold (Scanlon 2008), or perhaps their character (Watson 1996). Or maybe we are responsive to all three in some fashion (Shoemaker 2015).

<sup>103</sup> Russel (2004) suggests it is a kind of moral sense.

everyone lacks the capacity to recognize and respond aptly to a sufficient range of reasons for action. But some of us are reasons-responsive so as to be aptly held to the demands of morality, aren't we? Some of us are normal persons in this sense. Now, how could we be wrong about that fact? For it would mean that many people who *seem* to be receptive to a sufficiently wide range of reasons, and responding to an appropriate subset of these reasons, are not in fact doing so. Ourselves included! This is practically inconceivable.

If, on the other hand, everyone was excused, it must be that in every case some local condition obtained such that we could not be held culpably to the demands of morality. So, even though many of us are normal *reasons-responsive* agents, no one ever actually acted like one. In any token instance of *apparent* competence with respect to the demands of morality, no one ever is actually competent. In other words, every time it *looks* like someone responded aptly to reasons, no one actually ever did. But how could we be wrong about each of our own situations, to this degree? It is absurd to think that we have been wrong all along either about our powers of reasons-responsiveness or how well we use them. It is practically inconceivable.

Notice that we could plug in all sorts of different capacities into the argument structure and get the same basic result. Take the claim that "persons are morally responsible" to be a generic generalization underwritten by some plausible capacity (e.g., whatever capacities are needed for adult interpersonal relationships). If anti-realism about free will is true, then the evidence for our having any particular responsibility-relevant capacity and the evidence about the local conditions within we exercise of these capacities, namely our ordinary experience of ourselves and others as persons, is somehow insufficient justification for believing that persons are morally responsible. And there is something rather absurd about that.

#### 4.8 Assessing the *Reductio*

Earlier, I said that the *reductio* seems to impose a strong burden on the anti-realist. It seems like the anti-realist must explain why we are wrong about the special competence in light of which we understand the kind “person” in the ordinary sense. I can now rephrase. The burden on the anti-realist is to offer a plausible story of how the normal condition for the kind “person” could fail to include special competencies like reasons-responsiveness. She has to make that possibility seem practically conceivable.

There are at least four very reasonable responses to this kind of argument for realism about freedom and responsibility. I consider each in turn.

First, a tried and true anti-realist tactic is to suggest that libertarian views of these competencies are unintelligible or incoherent, whereas compatibilist views of them are insufficient for genuine free will and moral responsibility. I will set this response aside entirely. For this reply has to do with substantive views of the competencies and question, and the argument schema, and indeed, many instances of the argument-type in question, say nothing about whether or not the abilities or capacities needed to be so competent are compatible with determinism. Certainly, if no realist view on offer is successful on its own terms, that shifts the burden back onto realists to develop their theories. It does not, however, alleviate the burden faced by anti-realists in light of the *reductio*.

The second problem has to do with practical conceivability. Consider, for instance, Pereboom’s careful work (2001, 2014) in arguing for free will skepticism. Let’s first define *basic* desert as desert merited simply by the performance (or omission) of an action. Call the sort of control that would make us deserve anything in the basic sense free will. Pereboom argues that we lack this kind of control, and hence we lack free will, and so are not morally responsible in the basic desert sense. (He does so by way of manipulation arguments which I will not rehearse here). But Pereboom further argues that much of what we care about in human life—many of our cherished practical aims— would

largely survive a world purged of moral responsibility practices (in the basic desert sense). The list of survivors is pretty exhaustive. Let me consider a paradigm instance. Personal relationships, Pereboom argues, could be reconstituted by emotions of sadness, regret, and love, which do not implicate basic desert, in the absence of blaming and praising practices. Pereboom, in effect, argues that there really is a practically conceivable normal condition for persons in the absence of freedom and responsibility. He goes on to argue that that there are ways which this new normal would enhance human flourishing. Some of our most basic human aims like love and human fellowship might be better realized. Simply put: the realist argument I offer here relies on a failure of moral imagination.

Now, a complete picture of a world in the absence of basic desert, of moral responsibility as we understand it, would also include a different picture of morality. Pereboom (2017) has a well worked out, axiology-only replacement for morality as we know it, (which I take to include deontic components). Along with this, he endorses a non-basic substitute for moral responsibility involving moral protest.

This objection directly meets the burden I suggest these *reductio* arguments force on the anti-realist. The question, then, is whether or not this substitute is practically conceivable, as the skeptic suggests.

The realist should deny that it is. Why? For when we imagine the skeptic's world, with her conception of "moral responsibility" and of "persons", we are not imagining a world where *persons* exist anymore. As Fischer and Ravizza (1998: 1) put it in the very first sentence of their seminal *Responsibility and Control*: "an important difference between persons and other creatures is that only persons can be morally responsible for what they do." Since the skeptic's world involves a different conception of the demands of morality. And so, it would offer a different conception of persons. Persons as a kind are characterized by their special competence with the demands of morality. That is to say, the beings of Pereboom's world are not candidates for being ourselves—the worlds they inhabit are not candidates for which world among all the possible worlds is in fact the actual world.

Perhaps this is too quick. Couldn't we, through revisionary action, *become* the better angels of our nature? Isn't it possible that we are really in a Pereboom-world, and we simply haven't realized it yet? I am not convinced, but there is something uplifting about that thought. But I also feel deeply unsettled. For there is something fine and good in *being responsible* and *acting responsibly* and *taking responsibility*. And all this would be lost, at least in the forms we currently know. If we were to radically revise our concept of persons, perhaps the basic thought that persons have dignity would be lost too.

Maybe this rejection of the optimistic anti-realist scenario is confounded. Our understanding of persons as morally responsible is shaped by our current social-moral practices, after all, and so we should expect ourselves to be biased in favor of our existing normative-conceptual scheme (Milam 2018). This leads the second worry, which has to do with the *reductio's* appeal to generic thinking.

Generic thinking is linked to prejudice and stereotyping. Following Sarah Jane Leslie (2017), I take it that problematic generics are related to our tendency to essentialize kinds. In this context, to essentialize a kind is to form a (tacit) belief that a hidden, stable property explains and grounds the common features of the kind (Leslie 2017: 405). (We seem to think things like "dog-ness" explains the apparent features of dogs, for instance). The problem, then, is that we have a tendency to essentialize *social* kinds in negative ways, and then to generalize: a person might be liable to believe the pernicious generic claims like "Muslims are terrorists" or "women are submissive" when they essentialize the social kind terms "Muslims" and "women".

How do we know when a generic is pernicious? Let's consider Sally Haslanger's (2011) discussion of pernicious generics and ideology. The generic claim "women are submissive" can be shown false by showing that it is not the case that women are submissive, even if all or most women are submissive. Why? It is not part of the nature of women (i.e., it is not normal for the kind "women" or part of the essence of being a woman) that they be submissive. What seems to make this generic true involves our dispositions to interpret the material world in an essentializing way. The conditions

which makes the generic seem true are generated by these very dispositions. (Haslanger calls this the “loopiness” of social structures). Rejecting claims like “women are submissive” thus achieves “a first step in ideology critique”, by helping to “make evident the interdependence of schemas and resources, of the material world and our interpretation of it.” (Haslanger 2011: 199).

The question, then, is whether or not “persons are morally responsible” only seems true because we have corresponding essentializing dispositions to interpret persons as something that they are not, and that this has pernicious import. In this context, the anti-realist will appear in the guise of an emancipatory ideology critic. She can say that to assert “persons are morally responsible” is to assert something like “women are submissive”. The realist *thinks* she is asserting a kind of empirical fact about the nature of persons, about the sorts of things which characterize them. But she is not. Instead she is reifying a destructive social practice that subjects persons to negative feelings, sanctions, and punishment. Perhaps she reifies this practice because she believes persons have something called “free will”. (We could easily reframe Pereboom’s project in this light). Put this way, I find the anti-realists worry very compelling.

Notice that this debate between the anti-realist-as-critic and the realist has a normative and a metaphysical dimension: it is *pernicious* to think persons are morally responsible, *because* they are not. On the normative question, I think the realist stands on firm ground. Why? Well, how do we fight pernicious generics, rooted in problematic essentializing? Leslie suggests that we eschew the very labels that make us prone to pernicious generalizing. For instance, we should drop the label *Muslim* and instead describe someone “as *a person who follows Islam*, emphasizing that person is the relevant kind sortal and that following Islam is a particular property that the individual happens to possess” (Leslie 2017: 420). In other words, adopting *person* as the relevant kind term seems to be a way to undermine pernicious generic thinking. It is an inclusive kind category. It would therefore be surprising if the kind *person* itself was morally problematic.

On the metaphysical question, the realist has at least two options. (Perhaps there are others). First, she could deny that our special moral competence is a matter of interpretation or social construction. She could insist that there is an independent and intrinsic power which makes it appropriate to hold persons to the demands of morality. And this power is part of what it is to be a person. She could provide arguments to the effect that our ordinary presuppositions about persons and responsibility are therefore justified in a robustly realist way. Alternatively, a realist could accept that freedom and responsibility are, at least in part, socially constituted (cf. Vargas 2018). If so, then the realist has two tasks. She has to defend the acceptability of taking persons to be free and responsible to be at least in part constituted by our disposition to interpret persons as free and responsible. Then she must argue that this situation is not pernicious, but instead, justified.<sup>104</sup>

The final objection is a dilemma. Either the argument proves too much, or it proves too little. If the argument succeeds, it not only shows that anti-realism is false, but also fictionalism. It rules out that talk of freedom and responsibility is mere pretense. And we should be suspicious of any argument that proves so much so quickly. But if it doesn't rule out both anti-realism and fictionalism, then it proves too little. For then the argument only shows that we have good reason, given our practical aims, to *treat* ourselves as free and responsible.<sup>105</sup> In other words, the argument does not have the kind of mind-directedness I intend it to have. Couldn't we just be wrong about whether or not persons are morally responsible? (cf. Vargas 2018: 96-98).

I think this objection rests on a mistake. Typically, realism about a domain is the conjunction of two theses. The first is that the objects within the domain are real. The second is that these objects exist independent of anyone's beliefs, attitudes, and so on. Notice that the *reductio* only operates along the existential dimension of realism so construed, not the independence dimension. The objection

---

<sup>104</sup> Vargas himself adopts a form of revisionist realism; only *some* of how we go about things now is justified.

<sup>105</sup> Thanks to David Brink for bringing this objection to my attention.

assumes that the argument involves both the existence and the mind-independence claim; hence it either proves too much or too little. But the argument does not involve both claims. This leaves it open for views that accept the existence claim but not the mind independence claim to endorse the *reductio* argument, as I have offered here. These possible positions include constructivist, constitutivist, and response-dependence views of freedom and responsibility, perhaps among others. This might sound funny. To some, this motley bunch aren't *really* realist views. For the purposes of beginning to theorize about freedom and responsibly in terms of the existential question rather than the compatibility question, these views are realist enough for me. Indeed, neutrality on this question is warranted; I have left open whether or not free agency is in part socially constituted and so whether or not it is wholly mind-independent. We can decide later what implications follow from—and how much to care about—the possibility of mind-dependence in this debate.

#### 4.9 Conclusion

I have offered a discussion of a kind of argument for realism about freedom and responsibility. It critically supposes that, normally, persons are morally responsible. Indeed, the argument hinges on this being a platitude. After all this disputation, someone just might say to me that platitudes are all well and good most of the time. But when *we philosophers* do philosophy, there is a terrible metaphysical possibility to grapple with: maybe there are no persons in the ordinary sense because no “person” is a free and responsible agent. This possibility won't keep most people up at night. Perhaps it shouldn't keep us philosophers up at night either.

## Chapter 5

### Compatibilism and Responsibility-Based Realism about Free Will

**Abstract:** Many philosophers are realists about free will. They believe that free will exists. One sort of argument for realism goes like this. It seems like if persons didn't have free will, then they would never be morally responsible. And we sometimes are morally responsible. Some libertarians and compatibilists about free will—realists who disagree about whether or not free will is compatible with the thesis of physical determinism—avail themselves of this kind of argument. In fact, many take this kind of argument to directly answer the question of whether or not free will is compatible with determinism, skipping over realism entirely. This is especially true of arguments that characterize our being responsible in terms of the aptness of our moral responsibility practices. In this paper, I take up the question of what, if anything, this kind of argument says about the compatibility of free will and determinism.

#### 5.1 Introduction

Many philosophers are realists about free will. They believe that free will exists.

Why be a realist? Well, you might think that the arguments against the reality of free will and moral responsibility fail. But you might have something positive to say. For instance, Peter van Inwagen (1983: 209) offers this little argument: “Without free will, we should never be morally responsible for anything; and we are sometimes morally responsible.” Some will find this little argument sort of silly, but I do not.<sup>106</sup> This is as good of a reason as any to be a realist about free will.

Reasoning from our being morally responsible to our being free is nowadays primarily associated with compatibilism—the view that free will is compatible with the thesis of physical determinism. This is ironic, given van Inwagen's straightforward endorsement of this kind of reasoning. Many following P.F. Strawson (2008/1962) claim that when we inspect our ordinary moral practices to discern the appropriateness conditions of holding one another responsible, we see that

---

<sup>106</sup> Maybe it's not straightforwardly silly. You might just deny that our practices of moral responsibility are directly relevant to the free will problem. You might think that the free will problem is a problem in metaphysics. Maybe it has very important moral upshots. Nevertheless, the drawing of conclusion from moral responsibility about free will is just beside the point. There is a genuine methodological dispute here, one which I cannot address in this space. My aim is to offer a picture of a sort of realist methodology and to draw conclusions about free will and determinism from it.

these conditions are compatible with determinism (e.g., Wallace 1994, Russell 2017, *inter alia*). Why? Some sort of control over one's actions is implicated in these appropriateness conditions. This kind of control is plausibly had in virtue of some ability or set of abilities. And plausibly, whatever ability or abilities that afford us kind and degree of control necessary for moral responsibility just constitute free will.<sup>107</sup> The idea is to reject the view that the aptness of our moral responsibility practices hangs in the balance, and instead, to treat these practices as a given, a starting place for a normative investigation into moral agency (Wallace 1994: 98).

Some libertarians have, in criticizing this Strawsonian turn, argued that this way of framing things simply points out how much would be lost, would be made nonsensical, would be rendered pointless, if determinism were true (e.g., Wiggins 1973). Perhaps a close look at our practices of holding one another responsible actually favors the case for libertarianism. Perhaps they enshrine a fair opportunity to avoid wrongdoing which is not compatible with physical determinism (e.g., Franklin 2018, cf. 2010).

Van Inwagen has worried that free will is not compatible with either physical determinism or indeterminism; perhaps this argument for realism says nothing about the *nature* of the freedom which grounds moral responsibility (2014, *inter alia*). Nevertheless, it is worth noting that in recent work, even he tries to locate the relevant kind of ability in the free will debate with references to our moral practices (2014: 221-225).

It is difficult, then, to talk about this kind of argument for realism about free will, the sort grounded in the apparent fact that we are morally responsible, without *immediately* bringing up a separate question, the question, that is, of whether or not free will is compatible with the thesis of

---

<sup>107</sup> This is my preferred way of framing issues about free will, although I recognize it's not wholly uncontroversial. I am taking a cue from Pereboom (2001: xxii), Mele (2006: 17), McKenna (2008: 187), Sartorio (2016: 7-8).

physical determinism. Not all realists are decided on this question.<sup>108</sup> Nevertheless, most realists about free will are also disputants in the free will debate. (The fact that the debate about compatibility is aptly referable to as *the free will debate* proves how much of the literature is devoted to the compatibility question). It is a noteworthy fact that the foregoing realists about free will that we have been discussing—let’s call them *responsibility-based* realists—appeal to the very same reasons that justify their realism in support of their preferred answer the compatibility question. Strawson and the philosophers fighting under his banner take responsibility realism to show that compatibilism is true; detractors can take the very same to show that incompatibilism is true. Indeed, for many, the realism about free will is an afterthought. For incompatibilists who are skeptics about free will, the line of reasoning can run in reverse: upon examining our practices of moral responsibility, we see that they implicate powers or capacities which we lack. So, we don’t have free will, and therefore are not morally responsible after all. But for realists about free will, the considerations surrounding our practices of moral responsibility are often taken to *directly* answer the compatibility question.

In this paper, I take up the question of what, if anything, these kinds of responsibility-based considerations in support realism say about the compatibility question. I will argue that they in fact favor compatibilism. Why? An incompatibilist theory of free will requires not only evidence in favor of a theory of agency, but additional evidence in favor of physical indeterminism. Our practices of responsibility speak *against* adopting evidential standards for justified belief in freedom and responsibility which require evidence of the falsity of determinism.

## 5.2 Responsibility-Based Arguments

Although I understand these responsibility-based arguments to be first and foremost arguments for realism, they are often taken to directly support positions in the free will debate.

---

<sup>108</sup> Al Mele (1995, 2006) is a well-known agnostic on this issue.

For a typical example of this, let's consider how R. Jay Wallace (1994) proceeds. His project is a direct successor to P.F. Strawson's (2008/1962) case for compatibilism, with a Kantian flair to it.

Like Strawson, Wallace accepts that to hold someone responsible is to have towards this person a reactive attitude like resentment or indignation. Second, Wallace argues that reactive attitudes have beliefs about quality of will, involving judgment about moral obligations, as their objects. We are resentful, for instance, when someone willfully violates a moral obligation that they have towards us. Finally, Wallace endorses the following grounding thesis: that to be morally responsible is just a matter of whether or not one is the appropriate target of blame. And with these theses in place, Wallace's argument can be stated quickly. As McKenna and Coates (2015, §D) put it: "Determinism would not show that no one ever violates moral obligations, nor would it show that everyone is incapacitated to understand or comply with the demands involved in moral obligations."

In other words, Wallace's account suggests that we do have a kind of control, compatible with determinism, in virtue of which we can be responsible for our actions: the capacity to understand and comply with the demands of moral obligations. And Wallace supplies a normative principle for why this kind of control is sufficient. It is only fair to demand of agents that they comply with morality if they have this capacity. Since Wallace supplies a capacity in virtue of which we are free and responsible, his account seems well poised to block worries about incompatibilism about free will. Excuses that look like they should support incompatibilism, he argues, *only* operate in special circumstances. For example, claiming that "she couldn't have done otherwise!" only works, he suggests, when the agent is unable to understand and comply with the demands of morality. It does not show, for instance, that real moral responsibility requires that there be alternative possibilities which would not exist if determinism obtained.

The general form of the argument, then, is to reflect on when it is appropriate to hold a person responsible, and from this, glean a set of appropriateness conditions. From these appropriateness conditions, we can get an answer to the question: is free will compatible with determinism?

Here, surely, the anti-realist can raise a powerful challenge. This argumentative strategy simply starts from the *apparent* fact that we are morally responsible. To echo van Inwagen (1983: 209): “we cannot but view our belief in moral responsibility as a justified belief.” The point here is cast in terms of *belief* rather than actuality. This is problematic. For it could turn out that we are all simply the puppets of Martians who are controlling our minds. Even if our Martian puppeteers give us no reason to think no one ever violates moral obligations, it would nevertheless be deeply *unsettling* to find out we were the Martians puppets. We would lack the sort of freedom we *thought* we had by having, well, free will! This, in turn, casts doubt on whether or not we really are justified in believing we are responsible. Wallace’s argumentative strategy simply won’t do. As Susan Wolf (1981) put it, the only way to assure ourselves against this sort of pessimistic possibility is to show that we are *actually* free. Indeed, as Vargas (2011: 96) nicely articulates the problem with the responsibility-based approach:

“surely we could be wrong about whether or not we are morally responsible creatures in the same way people were wrong about water being basic and indivisible, in the way people have been wrong to think of women as property, and in the way in which it was a mistake to deny people the right to vote because of their skin color.”

We can’t therefore help ourselves to the premise that persons are responsible in order to show that they are also free.

Some readers will find this kind of problem for the responsibility-based strategy decisive.<sup>109</sup> I disagree. Let me consider two strategies. First, we could adopt a kind of relativism that would allow us to preserve the crucial premise. This was Strawson’s (1985) solution to this issue. He suggests we relativize between our “involved” or “participant” attitude towards the world and our “detached” or

---

<sup>109</sup> Thanks to Manuel Vargas for in person discussion on the methodological issues at play here.

“objective” attitude (1985: 36). From the one view, we experience the world in ways tied up with our natural commitments—commitments to external bodies, the power of induction, the responsibility of other persons, even morality itself. From the other, we see these things as illusions or inventions and no more. But Strawson suggests that of our natural, moral way of seeing the world, that “our natural disposition to such attitudes and judgements is naturally secured against arguments suggesting that they are not in principled unwarranted or unjustified” (1985: 38-39). Questions of warrant or justification only make sense in the general framework of either the “involved” or “detached” way of seeing things; there is no way to externally justify either mode of engagement by reason (1985: 41).

Funnily enough, a relativizing suggestion was also endorsed by C.A. Campbell in his (1957: 170) in defense of libertarianism: “There is no reason whatever why a belief that we find ourselves obliged to hold *qua* practical beings should be required to give way before a belief which we find ourselves obliged to hold *qua* theoretical beings; or for that matter, *vice versa*.” As he saw the dialectical, all that a determinist can prove is a deep conflict between the nature of the world and the nature of ourselves as we know it to be: “an antinomy in the very heart of the self” (1957: 171).<sup>110</sup>

Once again, the convergence between compatibilists and libertarians is striking—Strawson may have had Campbell in mind as one of his dialectical opponents—suggesting a common argument in favor of realism that is neutral with respect to the compatibility question. Indeed, Strawson points out that a similar dialectic played out between Hume and his anti-skeptical opponent, Reid. Each endorsed the view that we have practical commitments which are unassailable but deeply disagreed about the nature of these commitments (1985: 33).

This kind of relativism might strike some as deeply unpalatable, so here is another suggestion. First, we should grant the objectors that we could be wrong about moral responsibility. Maybe there

---

<sup>110</sup> For an interesting defense of anti-realist incompatibilism against this charge, see Nichols (2015).

are structural aspects of the abilities and capacities that make us responsible that are discoverable but are not yet discovered. Maybe there are normative properties of responsible agents we have failed to recognize. And I'd be willing to bet that we are wrong about the extension of responsibility in hard cases, even though I think we've made significant progress on problems of extensionality. Vargas (2011) is not wrong to raise worries about the responsibility-based approach. But my suggestion doesn't need to admit that we can't be deeply wrong about the details, since we can argue for the crucial premise that we are in fact responsible by *reductio* and get at the details of realism about free will along the way.

Elsewhere, I have defended the following form of argument, which aims to take show that we are, in fact, morally responsible, and so, in fact, free:

1. Assume for *reductio* that no person is morally responsible.
2. If no person is morally responsible, it is either because everyone is exempted, or everyone is excused.
3. It is practically absurd to think that everyone is exempted, for this would mean that no normal persons have the special competence that makes them fitting targets of the demands of morality.
4. Likewise, it is practically absurd to think that everyone is excused, for this would mean that no normal person in any particular instance ever exercised their competence with respect to the demands of morality.
5. So, it is practically absurd to think that no person is morally responsible.
6. At least some persons are morally responsible.
7. Since moral responsibility requires this special competence, and what it takes to have this competence is a plausibly free will, anti-realism about free will is false.

It is not my aim to defend this argument from objections here, and it needs a good bit of unpacking. Nevertheless, it will be a helpful guidepost for thinking in general about responsibility-based reasons for realism. It is completely neutral with respect to the compatibility question since, as I will explain below, it is neutral about what abilities or capacities constitute free will. In this case, special moral competence is just a plausible way of picking out what we mean when we say that persons have free will. The basic idea is, like Wallace's (1994) strategy, to get at the nature of morally responsible agency indirectly by picking out the appropriateness conditions for holding an agent morally responsible.

Unlike this strategy, however, we start from the assumption that we are in fact wrong about what the responsibility-based approach typically takes for granted and show that it is absurd. *Practically* absurd, anyway.

Well, then, what the heck is “practical absurdity”? I’m following Strawson (2008/1962) here, who casts his argument in terms of “practical inconceivability”; he argues that although that it is not totally inconceivable that no person is morally responsible, it is practically inconceivable. (2008/1962: 26). Let’s say that something *p* is inconceivable if we cannot imagine any world that we don’t take to falsify *p* (cf. Yablo 2005). Let’s say that practical inconceivability, then, is inconceivability with respect to the practical domain, the imaginability of worlds where *p* concerns practical matters.

The argument thus claims that anti-realism about responsibility is absurd not in some strictly logical sense—we can after all imagine worlds that falsify realism, which verify that anti-realism is a logical possibility; imagine a world where everyone has lost their minds. A world where every person is really being mind-controlled by Martians is absurd, not insofar as it is not imaginable, but insofar as we cannot imagine there being *normal* persons in that world. (Holding fixed that Martian mind-control is, in fact, responsibility-undermining and that normally, persons are morally responsible).

To make this assertion about the normal case for persons is, in effect, to assert an empirical fact about ourselves, about the kinds of things that we are. Of course, we are not always morally responsible in particular cases. But to be a person is to, normally, be morally responsible. As van Inwagen says: “That we are convinced that we know something does not, of course, prove that we do know it or even that it is true. But it *is* true that we are morally responsible, isn’t it? And we *do* know it to be true, don’t we?” (1983: 209). In other words, when we think about which possibilities *we are*, any possible world that is a candidate to be the actual one has to be a normal-with-respect-to-persons world, given the facts as we know them.

Let's think about the claim "persons are morally responsible" as a generic generalization rather than a universally quantified statement. (There are, again, persons who are not morally responsible). I take it that a generic generalization about some kind is true in virtue of a way of being normal for that kind (Asher and Morreau 1995; Pelletier and Asher 1997; Nickel 2008, 2016).<sup>111</sup> This particular generalization is plausibly made true in virtue of some ability or capacity, which grounds the competence normal persons have with respect to the demands of morality. Many take the relevant capacity involves the power to recognize and respond aptly to reasons.<sup>112</sup> If you are inclined to the practiced-based compatibilism suggested above, you might think that the relevant ability or capacity involves the power to understand and be moved by the internal states of agents, like attitudes or concerns, which feature heavily in our responsibility practices.<sup>113</sup> Notice that both kinds of abilities or capacities are good candidates for what free will consists in, since, plausibly, free will just is whatever kind of control is needed to be aptly held to the demands of morality. Moreover, it is very important that these sorts of abilities or capacities are ordinarily just taken for granted. We presuppose that, in general, persons have them.<sup>114</sup>

With this in hand, let's formulate a more specific version of the *reductio* argument. Say, for instance, what makes it true that normally persons are responsible is that persons are able to aptly

---

<sup>111</sup> Asher, Morreau, and Pelletier say that *K's are F* is true if and only if for each individual *K*, the most normal worlds for that *K* are such that *K* is *F*. Normal worlds are determined by context. Nickel (2016) says that *K's are F* is true if and only if for every normal *K*, that *K* is *F*. For context-based alternatives to this approach, see: (cf. Schubert and Pelletier 1987; Declerk 1991; Chierchia 1995). For probabilistic alternatives see: Cohen's (1996, 1999, 2004) probabilistic approach. I cannot adopt this latter approach. Statistically normalcy cannot do the sort of work I need an appeal to normality to do (see Bennett 1980).

<sup>112</sup> See Fischer and Ravizza (1998), McKenna (2013), Sartorio (2016), Wolf (1990), Nelkin (2011), Vihvelin (2013) for some compatibilist views. For some libertarian view, see O'Connor (2000), Nozick (1981), Kane (1996), Mele (2008), and Balaguer (2010).

<sup>113</sup> See, for instance: Strawson (2008/1962), Watson (1987), Scanlon (2008), and Shoemaker (2015).

<sup>114</sup> Of course, not all persons do, it's just that the normal case is that persons do. And this normalcy claim doesn't have any deeper moral upshots. Persons who lack these abilities or capacities are still persons and deserving of treatment in accord with their dignity. It may be that how respect for their dignity manifests in our treatment of them differs from other persons. But, of course, none of us are perfectly normal, and so the ways in which we are to be treated, at least sometimes, will be different too.

respond to reasons. If anti-realism is true, then no person is in fact a normal person because no person ever aptly responds to reasons. Thus, the evidence for our having this particular responsibility-relevant capacity, along with the evidence we have about the local conditions within we exercise of this capacity, is somehow misleading. But can hold fixed *that they are normal persons* in the pertinent sense while giving up on this evidence, by taking it as misleading?<sup>115</sup> No. It's absurd to think that every person could fail to be person in the pertinent sense.

Although I will proceed with my preferred argument for realism in mind, I take it that this basic idea about our evidence manifests in various responsibility-based arguments. The anti-realist asks us to hold fixed that some agents who lack free will and so are not morally responsible are also persons in the pertinent sense. This means taking our ordinary evidence to the contrary as misleading. The responsibility-based realist points out that taking this evidence to be misleading is practically absurd; there are no worlds where there are normal persons and where all persons lack reasons-responsive capacities, for example. And so anti-realism about free will is to be rejected, because reasons-responsiveness, for example, is a perfectly good candidate for free will, whatever control is needed to be morally responsible.

Thus, I think the relativizing suggestion made by Strawson and Campbell somewhat misconstrues the tension between our practical and theoretical commitments. It is not that there is deep conflict between these commitments, but rather, that we can consider the space of possibilities with or without taking into account what we know from within our involved viewpoint.

But if I am right that there is a general responsibility-based argument for realism, why do realists *disagree* so intensely about the question of free will and determinism?

---

<sup>115</sup> Early adopters of Strawson's responsibility framework were motivated by precisely this point; Fischer and Ravizza (1998) begin with a discussion of the distinction between persons and non-persons in which Strawson's work is cited.

### 5.3 The Easy Case for Compatibilism...

The *reductio* for realism about freedom and responsibility suggests that we have good commonplace evidence in favor of our being morally responsible (and so free). The argument seems to offer us an easy case for compatibilism about free will and the thesis of physical determinism. This might help explain why many compatibilists are attracted to Strawsonian views.

To see why, let's look at a very simple argument for anti-realism based on physical determinism:

1. Physical determinism obtains.
2. Physical determinism renders agents unfree and so non-responsible. In other words, physical determinism is exculpatory.
3. Therefore, anti-realism about freedom and responsibility is true.

Now, the anti-realist has to provide us with reasons to think if determinism is true, then we are misled by our ordinary evidence that, normally, persons are morally responsible. Otherwise we should not accept premise 2. But how can this evidence be misleading *because of* determinism? Is it a reason to think we lack some further and crucial evidence about our capacities or the particular exercise of them? It is difficult to see why we should admit determinism as a reasonable ground for such doubt. And perhaps we should say that determinism does not provide us with any such reason. It doesn't obviously give us a reason to think that persons normally lack the special competence that makes them fitting targets of the demands of morality. (You can see how this is, in essence, a generalized version of Wallace's argument, abstracting away from his Strawsonian and Kantian commitments).

We have to tread carefully here. The point is not that we could never be wrong about our being morally responsible. There is a difference between being ignorant and being misled when it comes to such matters.<sup>116</sup> Maybe some evil demon is making us all hallucinate and so deceiving us

---

<sup>116</sup> Strawson (1992: 138) makes this distinction.

about our own mental lives. This salient skeptical possibility may provide reason to raise our epistemic standards for what we ought to count as knowledge. If I know that there is a coffee mug in front of me, then I know that there is no such evil demon. Since I cannot know that there is no such demon, I mustn't know about the mug! But determinism is not the same sort of hypothesis (or at least a compatibilist should contend so). An evil demon scenario remains a salient possibility for us precisely because any evidence we have to rule it out could be misleading.

On the contrary, in order for determinism to play an exculpatory role it must count as grounds for doubting that our ordinary evidence about the moral competency of persons settles that they are in fact morally competent. Given the nature of the evidence for the claim that “persons are morally responsible” and what is required to reasonably believe that some particular person is an exception to it, it is difficult to see how determinism could reasonably ground such doubt. For determinism (so far) does not seem to say anything about the morally pertinent capacities or the local conditions in which we exercise them. Determinism is just the thesis that there is one physically possible future. *Prima facie*, it says nothing about the capacities of agents, or whether or not they have the financial resources, or the muscle strength, or the...and so on... to exercise their responsibility-relevant capacities in particular situations (cf. Strawson 1992: 136-137). We could be ignorant about some deep story as to how all of this responsibility-relevant stuff happens. But we are not misled in thinking the way we ordinarily do. And so, *prima facie*, realism offers a positive answer to the compatibility question. After we accept realism, then, we should accept compatibilism.

#### **5.4 ...and Why It's Wrong**

Some readers will have grown impatient with this line of reasoning. And for good reason! It is obviously problematic. For it seems that there are existing pleas and calls for exemption in our moral responsibility practices that might count universally if determinism were true. This suggests that the capacities needed to underwrite the generic claim “persons are morally responsible”, the abilities or

capacities in virtue of which we are fitting targets of the demands of morality, are not compatible with physical determinism. The sort of realism recommended by the argument, this protest suggests, offers a negative answer to the compatibility question. Free will is not compatible with determinism, since whatever it takes to be morally responsible is not compatible with determinism.

The libertarian realist will argue as follows. Determinism is the thesis that that the past and the laws of nature entail one unique future. If so, then “being unable to do otherwise” would count as a universal excuse if determinism were true. Since there is one unique future, there are no alternative possibilities. Given this, no one could have done otherwise than they in fact did or will do otherwise then they in fact will do. Franklin (2018), for instance, argues that that the demands of morality (as revealed through the aptness of our reactive attitudes) are sensitive to a condition of fair opportunity to avoid wrongdoing. If determinism were true, then no one would have the opportunity to avoid wrongdoing (since there is only one physically possible future). How could we blame someone who could not have avoided wrongdoing?

We accept other kinds of excuses and exemptions. For instance, if experienced horrendous child abuse and became an abuser themselves, we might find our blame very much attenuated (cf. Watson 1987). Why? We might reason that this person’s history explains their future actions. From there, we might start to reason that, *really*, anyone’s history explains their future actions. Doesn’t it?

This line of reasoning looks surprisingly similar to the structure of the so-called manipulation arguments (e.g., Mele 2006, Pereboom 2001 and 2014). Manipulation undermines responsibility. It does so, it seems, because a manipulated agent is not the source of his or her actions. Likewise, if determinism is true, agents would not be the source of their own actions.

Thinking about pleas related to history and manipulation seems to express what might be called an ultimacy condition that free and responsible action requires that one be the ultimate source or *arche* of one’s own actions (cf. Kane 1996: 33-37).

So, some excuses and exemptions, some forms of pleas, seem to apply universally if determinism is true. It is important here that one only needs a *token* instance of a plea to count in every. Even if we couldn't dispel our natural tendency to hold others responsible to the demands of morality, we could learn that each token instance of this tendency expressed is ultimately unwarranted (cf. Bennett 1980). Russell (2017) illustrates this with an example: even if we could never dispel the natural fact that we feel fear, we could discover out that each token instance of fear is not warranted.

Hard determinists and libertarians can both avail themselves of this sort of thinking. The hard incompatibilist will be happy with the result that, it turns out, commonplace moral thought opens up the possibility that no one is morally responsible. But a libertarian will shudder at this suggestion. She will think that if determinism were true, everyone would lack the special competence that makes persons *persons*. Something must be done! For she and the compatibilist are both realists about freedom and responsibly, after all.

The responsibility-based libertarian may now appear to be the hero of our story. Let us again take a look at the argument-schema, only focusing on the case of excuses (the reader can fill in something similar for exemptions). If determinism were to count as grounds for universal excuse, then in a deterministic world no one would exercise the kind of competence that normal persons have. Of course, though, most of us are normal persons who exercise this kind of competence. Given that we have not ruled out the possibility that in every token instance we are gravely wrong, as the generalization strategy shows, we should reject the claim in the anti-realist's argument that determinism is true, not the claim that determinism exculpates. For determinism *actually renders us non-responsible*, because in a deterministic world everyone is in fact in conditions unsuitable for the having or the exercise of the agentic capacities that make us morally responsible. Realism is vindicated; incompatibilism is true. The practical absurdity of anti-realism about free will based on determinism follows from determinism, not from thoughts about universal exculpation.

## 5.5 Compatibilism Revisited

Earlier, when we were considering compatibilist realism, I wrote that it is difficult to see how determinism could reasonably ground doubt of the evidence given by ordinary practical life. It is no reason to think we are misled about the capacities of agents or the local conditions in which these capacities operate. The generalization strategy seems to show that I was wrong about that. It shows that our ordinary moral thought is sensitive to metaphysical considerations of agency, and so determinism. This seems to show that something more than the ordinary evidence underwrites the claim that persons are morally responsible. This something more provides evidence that determinism is false (at least with respect to our basic actions). Below, I argue that this strategy is not as promising as it first appears.

Why is it not promising? The important feature of the *reductio* is simply that mundane, generic, commonplace features of human persons are sufficient evidence that normally persons are morally responsible. The claim should be considered more or less platitudinous. The whole idea behind what I have been calling responsibility-based realism is that ordinary facts about acting for reasons, capacities for adult interpersonal relations, and so on, the sort that pretheoretically play a role in our responsibility practices, provide sufficient justification for being realists about free will. They are, in fact, good candidates for what free will consists in.

But if the libertarian is right that determinism is in fact exculpatory, then it seems like it would take *additional* evidence beyond ordinary considerations to reasonably show that persons are in fact morally responsible. Why? We would need evidence for the falsity of determinism. To be clear, I am not suggesting here that we talk in terms of justified belief in responsibility. The point is to ask *in virtue of what evidence* do we *know* that persons are responsible and so free? The libertarian suggests that there is some evidence that we have which shows that determinism is false. But it is this standard of evidence, however, that we ought to reject on pain of (not strictly logical) absurdity. On such a

standard, it is practically conceivable that we are not, in fact, acting in apt response to reasons, or engaged in interpersonal relationships, and so on. Determinism could have been true all along. It's an open physical hypothesis. It is precisely this possibility, that we are not normally responsible, that the *reductio* argument for realism suggests is absurd, as in, not practically conceivable. So, funnily enough, the libertarian has ended up undermining her own position.

This point about evidential standards does not show that libertarian is mistaken in some strict sense. When we deal with claims like “persons are morally responsible” we are dealing with, to use Aristotle’s turn of phrase, the realm of things “which are not necessarily so, nor always, but usually” (Metaphysics 6.1026b). Indeed, nothing further I have to say rules out the fact that determinism could supply us with excuses or exemptions. My complaint is that it does not establish that we *should* generalize our commonplace pleas if determinism obtains.<sup>117</sup> As I have articulated the responsibility-based *reductio* for realism, the central premises are not a matter of straightforward universally quantified claims. They are not open to token counterexample. Rather, they are a matter of what is normal for the kinds beings that we seem to be, namely, persons. And this we can glean by thinking about what is practically conceivable for beings like us, i.e., what worlds we would count as normal for beings like us. Therefore, whether or not we should generalize our pleas critically depends on which context-dependent evidential standards we ought to accept regarding our use of the concept “person”.

Put otherwise: the *reductio* is an argument to the effect that we ought to prioritize everyday evidence in the matter of who is and who is not a normal person. We are already prioritizing everyday evidence in giving the *reductio* anyways. The semantics of generic generalizations usefully elucidates the features of such standards. The argument thus shows that assuming standards which allow

---

<sup>117</sup> Recall Strawson’s question in “Freedom and Resentment” (2008/1962). He asks: would or should determinism make us abandon our moral feelings and attitudes? The reasons I give for an answer an answer of “no” to the should question differ from Strawson’s.

determinism to count as a relevant evidential consideration in thinking about freedom and responsibility leads to absurdity. Standards of this type suggest that we are not justified about things we can safely presuppose. We should therefore not adopt these standards.

Perhaps one was taken by libertarian's reasoning when cast in terms of being the source of your own actions. Indeed, I am inclined to think that ordinary moral thought enshrines deep intuitions about the sources of agency, and so find the libertarian's case here particularly compelling. Intuitively, *we* develop our essential moral competencies over time. An intervention on these developing competencies can undermine the extent to which they are our own. It is not crazy to imagine a person for whom the whole development has been hijacked since birth through indoctrination or abuse. It is often argued that determinism is analogously like manipulation.<sup>118</sup> Both, perhaps, undermine this crucial self-shaping activity. Does the compatibilist's point about evidential standards *really* undermine this kind of thinking?

I think so. At least, I am hopeful about it. The details would need to be filled out. The thought would go something like this: if we lacked these self-shaped capacities, and if indeed we need them to be aptly held to the demands of morality, then we would not be competent in the ways in which many of us know that we are. Most of us have not been robbed of our moral personalities. Responsibility-based realism about free will suggests, then, that for many of us, we have already undergone (or are still undergoing!) whatever historical processes make us the sources of our moral selves. This speaks in favor thinking that whatever it takes to be self-made or self-formed, it does not require the sort of metaphysical underpinning that has sometimes been thought incoherent at the extreme, for the ordinary self-shaping would have to count as evidence against physical determinism. For the reasons suggested above, this seems like a mistake (cf. Arpaly 2006: 122-123).

---

<sup>118</sup> For well worked out versions, see Pereboom (2001) and (2014), and (2008) for discussion.

## 5.6 Begging the Question

The libertarian may still insist that since pleas generalize if determinism is true, we cannot reject her epistemic standards, as doing so would simply amount to denying that there is a problem of free will and determinism. We are simply begging the question against the anti-realist who thinks that determinism is true and so we lack free will on the grounds that our shared evidence is misleading. In accepting the compatibilist position, I am recommending, we would be denying the epistemic standards by which we can have a common ground for debate.<sup>119</sup> The only dialectically acceptable position for the responsibility-based realist to accept, given the compatibilist rejoinder just articulated, is libertarianism.

Although this looks like a compelling libertarian rejoinder, it is problematic. Get up out of the armchair for a moment. Everyone ordinarily finds it obvious that they and others are persons, and that persons are normally morally responsible, and so they and others are normally morally responsible. Why is this obvious? Relative to our ordinary evidential standards, it is indeed obvious because normal persons are reasons-responsive, capable of adult interpersonal relationships, and so on. In thinking that this sort of evidence is not sufficient for justifiably believing that “persons are morally responsible”, we philosophers are effectively committed to claiming that ordinary people (and even ourselves, most of the time) are operating with intolerably lax evidential standards. But of course, pretty much everybody, all the time, employs these lax evidential standards when it comes to believing that persons are responsible. So, by-and-large, everyone is running around with intolerably lax standards about what they can justifiably believe about themselves and others *qua* persons as a generic category.

---

<sup>119</sup> The same goes for anti-realist hard incompatibilism, where the anti-realist thinks free will is incompatible with both determinism and indeterminism.

*Really?* It seems to me that the ordinary epistemic criteria are a better guide to the operative semantic standards governing correct use of the generic concept “person” than anything we can cook up from the armchair. For this reason, ordinary judgements concerning normalcy with respect to that concept seem more apt than judgements made in light of the compatibility question. It seems like a heavy dialectical burden to argue otherwise. Yet that is what libertarians—and indeed, all incompatibilists must do. In effect, the incompatibilist needs an error-theory about our most basic conceptual competencies. But error-theory about our ability to discern who is and who is not a normal person—and what is normal and what is not for a person—is not obviously forthcoming. This presents us with a strong *prima facie* case in favor of responsibility-based compatibilism. The incompatibilist’s need for such a thoroughgoing error theory shows incompatibilism is deeply revisionary with respect to our shared sense of what it is to be a person. But the revisionism does not end here. The reason why the incompatibilism is a revisionary about persons is because it must appeal to evidential standards that render our shared evidence of normal personhood hostage to determinism. You might say that this is revisionary with respect to our evidential standards too.

Here’s the bottom line: it is not question-begging against the anti-realist to point out the theoretical costs of entering into a shared domain of disagreement with her.

Perhaps some libertarians will be willing to stop defending commonsense and adopt a kind of revisionism in light of this problem. As a rule, revisionists about freedom and responsibility think the costs of revisionism are worth the theoretical gain. Nevertheless, they often fail to see the need for an error-theory about our ordinary conceptual competence, which by my lights renders such revisionism suspect.

### **5.7 Argumentative Burdens and Agentive Phenomenology**

The responsibility-based libertarian may see a way around this problem, though. She may (wisely) construe her project as a way of working out our shared commitments about the concept of persons.

If we were to stop and think about the forms of excuse and exemption we employ, everyone would see the *real* structure of their commitments. These commitments involve a requirement that agents not be determined prior to action. This would not be a revisionary project.

The libertarian can argue as follows: since we have good evidence from ordinary life that we are free and responsible, determinism must not obtain. So, the compatibilist is not wrong about the evidence for realism about freedom and responsibility. She is just wrong about what this evidence is evidence for. This misunderstanding underpins the charge of revisionism. Thus, the libertarian can deny the compatibilist-friendly evidential standards for justified belief in freedom and responsibility while staying true to the spirit of the *reductio*. Libertarians sympathetic to this line of argument tend to think that the phenomenology of agency is the source of this ordinary evidence (e.g., Campbell 1957, O'Connor 1995, and Swinburne 2013). When we introspect, they claim, we get good evidence about the nature of our powers.

Setting aside worries about what we can glean from introspection, I think it is perfectly acceptable for libertarians to draw on phenomenological resources in thinking about a responsibility-based defense of libertarian realism. We ordinarily presuppose that what it is to be a person is to be free and responsible. The *reductio* for realism under discussion, and in general any responsibility-based approach to realism, will get at this presupposition second-personally. In general, such approaches ask us to consider when we would be willing to *suspend* blame (or praise) in light of new information about the agent or the conditions under which she acted. What our practices of excusing, for instance, show us is that we typically presuppose that a person is free and responsible *unless* we have defeating evidence.

Indeed, our practices of blame seem to bake in the search for such defeaters since, plausibly, the aim of blame is to produce an *accounting* of what happened. If I find out that what I understood to be a punch to the face was in fact a very unfortunate and unintentional fist-first fall, I cease to hold

you accountable for your action. There might be lots of different explanations for what grounds my cessation. Maybe my blaming attitudes are no longer fitting or appropriate because I now know you didn't have bad attitudes towards me. Maybe they are no longer fitting or appropriate because I see you weren't in control of your fall—you didn't choose or try or decide to do it. Maybe both. Or maybe something else.

A phenomenological approach to realism can get at this presupposition first-personally. It asks us to consider what it is like to perform a basic action: to try, to decide, to choose, and so on. It then asks us to consider what it would take to think that the feelings of trying, deciding, choosing are somehow illusory. In doing so, it points out that we ordinarily presuppose that we are in control of ourselves in virtue of the performance basic actions. It is, it seems *up to us* that we perform them.

You might see where this is going. These approaches can converge insofar as they both seem to provide us with evidence about the pertinent abilities or powers which make us the appropriate targets of the demands of morality. We ask what it would take for our feelings or practices to be apt, which they seem to be. If the introspective case is strong that determinism would rule out that our choices are up to us in the right way, then we would have evidence that these powers are not compatible with determinism. Recall how Campbell (1957) saw the dialectic. The determinist has to show how the libertarian's phenomenological analysis was wrong, and the libertarian has to disprove the theoretical arguments of the determinist.

It will be useful here to work with a concrete view of agentic phenomenology in mind when assessing this libertarian rejoinder. Perhaps, as Horgan (2015: 36-37) puts it, our agentic experience involves *core optionality*. Perhaps we experience, in a non-discursive way similar to perceptual or kinesthetic awareness, that we have actionable options, that we could have done otherwise, and it is up to us that we could have done so. If this experience was in fact good evidence about our actual powers of choice, then it's no surprise that libertarians might take it as evidence that determinism

must be false. Ordinary agentic phenomenology is full of experienced options, which would be illusory options if determinism were true. And they don't seem illusory.

Since we've set aside suspicion of phenomenological methodology, notice that we can check the outputs of the responsibility-based considerations by introspection, in our own case. When would we think it would be inappropriate for another person to hold us responsible? When is it inappropriate for us to feel guilty for acting thus-and-so? Well, one such instance might be when we cannot comply with what morality demands of us. We might, for one reason or another, lack the option to comply by our own lights.<sup>120</sup> It seems to me, then, that the two approaches can be made interdependent. It seems perfectly acceptable, then, for a libertarian to point to her own experience of morally responsible agency in thinking about how to understand the case for responsibility-based libertarianism.<sup>121</sup>

Recently, some compatibilists have argued that our agentic phenomenology actually has compatibilist success conditions (e.g., Horgan 2011, 2015 and Deery 2015a, 2015b). For instance, perhaps our experience of ourselves as the source of action is compatible with a mental or brain state being the source of action, since this experience is just *absence* of the experience of our actions being

---

<sup>120</sup> Again, it is worth noting that a longstanding trend in libertarian thinking about free will locates freedom *precisely* in the experience of complying with the demands of morality, (e.g., Campbell 1957). This is of course not the only way to get introspective evidence about free will. We can exercise our powers of control in morally neutral situations.

<sup>121</sup> P.F. Strawson himself pointed this very interdependence out in his late work *Analysis and Metaphysics* (1992, 137-138):

“Our proneness to moral attitudes and feelings is a natural fact, just as the sense of freedom is a natural fact. I have remarked that they are linked, and it is time to say more about the link. In speaking of the sense of freedom, I connected it closely with the sense of self. Our desires, decisions, actions are not in general felt as alien, as things that simply happen in or to us, like a pain or a blow. They are we. Our awareness of them is awareness of ourselves. I remarked that we attribute to others this same sense of freedom and this same sense of self. We see others as other selves, and are aware that they so see each other. But this is not a matter of a conclusion drawn by analogical reasoning. In a variety of ways, inextricably bound up with the facts of mutual human involvement and interaction, we feel towards each other as to other selves; and this variety is just the variety of moral and personal reactive attitudes and emotions which we experience towards others and which have their correlates in attitudes and emotions directed towards ourselves”

caused by such states, rather than the experience of our actions *not* being state-caused (cf. Horgan 2015, Horgan and Timmons 2011).

Since we have reason to think that such phenomenological approaches and responsibility-based approaches can be advanced in tandem, we should think about this compatibilist rejoinder from the responsibility-based perspective. Indeed, considering both the first and second-personal perspectives will be illuminating.

I have argued above that libertarian realists have to not only provide evidence that we are responsible but also that determinism is false. We should therefore examine how can the phenomenological evidence from our agentic experience count *against* determinism. Just like it is hard to see how the ordinary, second-personal evidence of responsibility as gleaned from our responsibility practices can be undermined by the truth of determinism, it is difficult to see how experiential evidence of freedom could count against determinism too.

What does my free agency phenomenology say about physics and laws of nature? Nothing obvious. Let me explain. A libertarian might point to our experience of excuses by saying that, if we find introspectively that we lack the option to comply with the demands of morality, we should not be held responsible. This seems to yield straightforward incompatibilist results. For if determinism were true, we would lack the option to comply in every case. There would be no options in the pertinent sense! So, determinism must be false. The compatibilist should, in turn, argue that our ordinary thinking about options does not yield incompatibilist results. When we experience a lack of options, it is (plausibly) because some additional experience is superimposed on our basic agentic phenomenology (cf. Horgan 2015: 37). Some of these feelings might be internal, like the experience of an intense desire to do what one ought not to, or the felt lack of muscle strength; others might be more complex, like what it feels like to not have enough the financial means to help a friend in need (cf. Strawson 1992:137). None of these responsibility-undermining experiences are obviously like the

experience of being physically determined. For all we know, physical determinism actually obtains. There isn't a relevant experience to be superimposed because our agentic phenomenology is not directly sensitive to physical determinism. We have no contrast class here to compare experiences.

Indeed, our understanding of responsibility-undermining experiences does not seem to operate *at the same level of description* as physical determinism. It would be surprising if these ordinary feelings did operate at that level. That would be to confuse two very different modal notions (cf. Wallace 1994: 217, List 2019: 97-107): the notion of what it is physically possible for an agent to do given the past and the laws of nature and what it is physically possible for a person to do given her constitution, her capacities, and her situation (glossed in some suitably ordinary way as it relates to our practices of moral responsibility). Why think our introspective experience would have success conditions in terms of the first modal notion rather than the second? By my lights, we don't have very good reason to. Ordinary experiences do not involve cognitive content about the deep laws of physics. Maybe they involve basic causal insights, but it's not obvious how those would "bubble up" to an understanding of the world as physically determined. There is therefore a strong *prima facie* reason to think that, whatever the success conditions are for our agentic phenomenology, at least when it comes to experiences of morally responsible agency, these conditions will be compatible with determinism.

## 5.8 The Deeper Dispute

Here is the upshot of this back-and-forth dialectic. The deeper dispute between compatibilists and incompatibilists on this responsibility-based way of proceeding is really about the epistemic standards for justified belief in responsibility, and so freedom, as a normal condition of persons.

And here is a simple way to capture this deeper dispute about the epistemic standards for justified belief in moral responsibility, which I have been characterizing dialectically. Recall that when I discussed the easy case for compatibilism, I differentiated, on the behalf of compatibilists, the case where determinism obtained the case where a skeptical epistemic scenario obtained. Maybe some evil

demon is making us all hallucinate and so deceiving us about our own mental lives. An evil demon scenario remains a salient possibility for us because any evidence we have to rule it out could be misleading. But determinism is an open physical hypothesis. It does not show our ordinary experience to be misleading. For all we know, determinism (or something near enough to it) does in fact obtain. Perhaps we are ignorant about the physical facts. But the compatibilist denies on pain of practical absurdity that this shows that we are misled about the underlying competencies that make us morally responsible agents. This is to deny premise 2 in the very simple anti-realist argument based on determinism:

1. Physical determinism obtains.
2. Physical determinism renders agents unfree and so non-responsible. In other words, physical determinism is exculpatory.
3. Therefore, anti-realism about freedom and responsibility is true.

A libertarian realist will characterize the case of physical determinism and the case of epistemic skepticism quite differently. So, she will reject premise 1.

For instance, van Inwagen (1983:210-214) sees these two cases—determinism and the presence of an evil demon— as equivalent. If we have good reason to think we are morally responsible, then we have good reasons to think we have free will. If free will implies the falsity of determinism, then we have good reason to think determinism is false. But determinism is the sort of thing that is outside of the reach of our local epistemic capacities. It's the sort of thing decided by empirical science, or if not so decidable, the sort of thing we ought to suspend judgement about. Likewise, if we have good reason to think that most of our claims to know are apt, and that the aptness of knowledge claims implies that there is no deceiving evil demon, then we have good reason to think there is no deceiving evil demon. But whether or not there is a deceiving evil demon is a question outside of our epistemic ken, the sort of thing we ought to suspend judgement about.

Van Inwagen sees two forms of response here. We could either bite the bullet and accept that we do have good reasons for thinking we are free and also not in a skeptical scenario, or deny a very

plausible inference rule: If one has good reason to believe P, and P entails Q, then one has good reason to believe Q. He favors the first response. Indeed, van Inwagen sees no other way around this problem for a libertarian. He or she cannot treat the two cases differently. That would be treating essentially the same argument in two different ways arbitrarily (1983: 214).

The compatibilist, then, can respond to these two cases differently without risking arbitrariness. She denies that the ordinary evidence for our being morally responsible is also evidence of the falsity of determinism. That is to say, she denies that van Inwagen's plausible inference rule applies in this case. So, she denies that determinism defeats our evidence that we are, in fact, sometimes morally responsible. Determinism wouldn't show that our ordinary evidence is misleading. By contrast, in the evil demon scenario, the possibility of the demon undercuts our claim to know about anything, and if we did know something, it would show that there is no demon.

Since the libertarian thinks that this evidence does entail the falsity of determinism, and that determinism could defeat our ordinary evidence, then the two cases are exactly alike. So, all things being equal, realists ought to adopt compatibilism. As I have argued, the compatibilist is right to think that there is a difference between the cases.

## 5.9 Conclusion

I have argued that the responsibility-based considerations that support realism about free will also support compatibilism about free will. Our practices of moral responsibility speak against adopting evidential standards for justified belief in freedom and responsibility which require evidence of the falsity of determinism. Perhaps there are reasons to be skeptical that our practices of responsibility have anything to say about the metaphysics of free will. That is a debate for another time. By my lights, the fact that we are morally responsible is as good a reason as any to believe in free will. It is also, it turns out, a good reason to be a compatibilist.

## References

- Algoe, S.B. & Haidt, J. 2009, "Witnessing excellence in action: the "other-praising" emotions of elevation, gratitude, and admiration", *The Journal of Positive Psychology*, vol. 4, no. 2, pp. 105–127.
- Aquinas. T. *The Summa Theologiae of St. Thomas Aquinas* Second and Revised Edition, 1920  
Literally translated by Fathers of the English Dominican Province Online Edition Copyright  
© 2017 by Kevin Knight. <http://www.newadvent.org/summa/index.html>
- Aristotle. *Metaphysics*. ed. W.D. Ross. Oxford: Clarendon Press. 1924.
- Arpaly, N. 2006. *Merit, Meaning, and Human Bondage: An Essay on Free Will*. Princeton: Princeton University Press.
- Asher, N. and M. Morreau. 1995. "What Some Generic Sentences Mean", in *The Generic Book*, Eds. G. Carlson and F.J. Pelletier Chicago: Chicago University Press: 300–339.
- Ayer, A. J. 1954. "Freedom and Necessity." *Philosophical Essays*. New York: St. Martin's Press: 3-20.
- Balaguer, M. 2010. *Free Will as an Open Scientific Problem*, Cambridge, MA: The MIT Press.
- Bear, A and J. Knobe, 2017 . "Normality: Part Descriptive, Part Prescriptive." *Cognition* 167: 25-37.
- Beglin, D. "Responsibility, Libertarians, and the "Facts as We Know Them": A Concern-Based Construal of Strawson's Reversal \*." *Ethics* 128, no. 3 (2018): 612-25.
- Bell, M. 2013, *Hard Feelings: The Moral Psychology of Contempt*. Oxford University Press, New York.
- Bennett, J. 1980. "Accountability." in *Philosophical Subjects: Essays Presented to P.F. Strawson* ed. Zak van Straaten Oxford: Clarendon: 14-47.
- Bennett, J. 2008, "Accountability (II)" in *Free Will and Reactive Attitudes: Perspectives on P.F. Strawson's "Freedom and Resentment"*, eds. M. McKenna & P. Russell, Ashgate, Burlington, VT; Farnham, England, pp. 47.
- Campbell, C.A. 1957. *On Selfhood and Godhood: The Gifford Lectures Delivered at the University of St. Andrews During Sessions 1953-54 and 1954-55*. London: Allen & Unwin.
- Chisholm, R. 1964. "Human Freedom and the Self". Reprinted in Chisholm, Roderick. 1989. *On Metaphysics*. Univ. of Minnesota Press.
- Clarke, R. 2009. "Dispositions, Abilities to Act, and Free Will: The New Dispositionalism." *Mind*, Vol. 118: 470: 323-51.
- Clarke, R. 2015. "Abilities to Act", *Philosophy Compass*, vol. 10. 12:893-904.
- Cohen, A. 1999. "Generics, Frequency Adverbs and Probability", *Linguistics and Philosophy*, 22(3): 221–253.
- Cohen, A. 1996. *Think Generic: The Meaning and Use of Generic Sentences*, Ph.D. dissertation, Carnegie

Mellon University.

- Cohen, A. 2004. "Generics and Mental Representation", *Linguistics and Philosophy*, 27(5): 529–556.
- D'Arms, J, and D. Jacobson. 2003. "The Significance of Recalcitrant Emotion (or, Anti-quasijudgmentalism)." *Philosophy*: 127-145.
- Darwall, S.L. 2006, *The Second-Person Standpoint: Morality, Respect, and Accountability*, Harvard University Press, Cambridge, Mass.
- Davidson, Donald. 1973. "Freedom to Act". In Ted Honderich (ed.), *Essays on Freedom of Action*. Routledge.
- Deery, O. 2015a. "The Fall from Eden: Why Libertarianism Isn't Justified by Experience". *Australasian Journal of Philosophy* 93 (2):3 19-334.
- Deery, O. 2015b. "Is Agentive Experience Compatible with Determinism?" *Philosophical Explorations* 18 (1):2-19.
- Ekstrom, L. 2003. "Free Will, Chance, and Mystery," *Philosophical Studies* 113: 153-180.
- Elkman, P. and W. Friesen, 1986, "A New Pan-Cultural Facial Expression of Emotion." *Motivation and Emotion*, vol. 10, no. 2, pp. 159–168.
- Epps, J. and P.C. Kendall, 1995, "Hostile Attributional Bias in Adults." *Cognitive Therapy and Research* 19, no. 2, pp. 159-178.
- Fischer, J.M. 2011. "The Zygote Argument Remixed." *Analysis* 71: 267-72.
- Fischer, J.M., and Ravizza, M. 1998. *Responsibility and Control: An Essay on Moral Responsibility*. Cambridge: Cambridge University Press.
- Fischer, John Martin. 1994. *The Metaphysics of Free Will*. Oxford: Blackwell Publishers.
- Foot, P. *Natural Goodness*. Oxford: Oxford University Press, 2001.
- Foot, Philippa. 1978. *Virtues and Vices and Other Essays in Moral Philosophy*. Oxford University Press.
- Frankfurt, Harry. 1969. "Alternate Possibilities and Moral Responsibility." *Journal of Philosophy* 66: 829-39.
- Frankfurt, Harry. 2002. "Reply to John Martin Fischer." In Sarah Buss and Lee Overton, eds., 2002. *Contours of Agency: Essays on Themes from Harry Frankfurt*. Cambridge, Mass: MIT Press: 27-31.
- Franklin, C.E. 2018. *Minimal Event-Causal Libertarianism*. Oxford University Press.
- Franklin, C.E. 2010. *Strawsonian Libertarianism: A Theory of Free Will and Moral Responsibility*, ProQuest Dissertations Publishing.

- Fredrickson, B.L. 2001, "The Role of Positive Emotions in Positive Psychology: The Broaden-and-Build Theory of Positive Emotions", *American Psychologist*, vol. 56, no. 3, pp. 218-226.
- Ginet, C. 1997. "Freedom, Responsibility, and Agency." *Journal of Ethics* 1: 85-98.
- Giubilini, A. 2015, "What in the World is Moral Disgust?", *Australasian Journal of Philosophy*, vol. 95, no. 2, pp. 227-242
- Greene, J & Haidt, J. 2002. "How Does Moral Judgment Work." *Trends in Cognitive Sciences* 6 (12):517-523.
- Haidt, J. & Morris, J.P. 2009, "Finding the Self in Self-Transcendent Emotions", *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 19, pp. 7687-7688.
- Haidt, J. 2003a, "Elevation and the positive psychology of morality." in *Flourishing: Positive psychology and the life well-lived.*, eds. K. CLM & J. Haidt, American Psychological Association, , pp. 275–289.
- Haidt, J. 2003b, "The moral emotions" in *Handbook of affective sciences*, eds. R.J. Davidson, K.R. Scherer & H.H. Goldsmith, Oxford University Press, Oxford, pp. 852-870.
- Hazebroek, J.F., K. Howells, and A Day, 2001, "Cognitive Appraisals Associated with High Trait Anger." *Personality and Individual Differences*, vol. 30, no. 1, pp. 31-45.
- Horgan, T. 2011. "The Phenomenology of Agency and Freedom: Lessons from Introspection and Lessons from Its Limits." *Humana Mente* 15 (Jan. 27, 2011): 77–97.
- Horgan, T. 2015. "Injecting the Phenomenology of Agency into the Free Will Debate". *Oxford Studies in Agency and Responsibility*, vol 3. Ed. David Shoemaker. Oxford University Press: Oxford. 34-61.
- Horgan, T., & Timmons, M. (2011). Introspection and the Phenomenology of Free Will: Problems and prospects. *Journal of Consciousness Studies*, 18(1), 180-205.
- Kane, R. 1996. *The Significance of Free Will*, New York, Oxford University Press.
- Kant, I. 1998. Guyer, P. & Wood, A., Eds. *Critique of Pure Reason*, Cambridge: Cambridge University Press.
- Kant, I. 1998, "Religion Within The Boundaries of Mere Reason" in *Religion Within The Boundaries of Mere Reason: And Other Writings*. Ed. Allen Wood and George di Giovanni. Cambridge: Cambridge University Press.
- Kelly, D. 2012, *Yuck: The Nature and Moral Significance of Disgust*, MIT Press, Cambridge, Mass.
- Lear, J. 2006. *Radical Hope: Ethics in the Face of Cultural Devastation*. Harvard University Press.

- Lehrer, Keith. 1968. "Can's Without 'If's.'" *Analysis* 24: 159-60.
- Leslie, S.J. 2007, "Generics and the Structure of the Mind", *Philosophical Perspectives*, 21(1): 375–403.
- Leslie, S.J. 2008, "Generics: Cognition and Acquisition", *Philosophical Review*, 117(1): 1–47.
- Lewis, D. 1973a. *Counterfactuals*. Cambridge: Harvard University Press.
- Lewis, D. 1973b. "Causation," *Journal of Philosophy*, 70: 556–567.
- Lewis, D. 1997. "Finkish Dispositions." *The Philosophical Quarterly*. vol. 47. 187: 143-158.
- Lewis, D. 1983 "Languages and Language". in *Philosophical Papers, Volume 1*. Oxford: Oxford University Press: 163-188.
- Lewis, M. 2008, "Self-Conscious Emotions: Embarrassment, Pride, Shame, and Guilt" in *Handbook of emotions*, eds. M. Lewis, J.M. Haviland-Jones & L.F. Barrett, 3rd Edition edn, Guilford Press, New York, NY.
- List, C. 2019. *Why Free Will is Real*. Cambridge, Mass: Harvard University Press.
- Macnamara, C. 2015, "Blame, Communication and Morally Responsible Agency" in *The Nature of Moral Responsibility*, eds. R. Clarke, M. McKenna & A. Smith, Oxford University Press, New York.
- Manley, D. & Wasserman, R. 2007. "A Gradable Approach to Dispositions." *The Philosophical Quarterly*, vol. 57. 226: 68-75.
- Martin, C.B. 1994. "Dispositions and Conditionals." *The Philosophical Quarterly*. vol. 44. 174: 1-8.
- McDougall, C. 2007, *The Hidden Cost of Heroism*. NBC News. Rodale Inc, Emmaus.  
<http://www.nbcnews.com/id/21902983/ns/health-behavior/t/hidden-cost-heroism/#.V9Bp07Ut2Uc>
- McKenna, M 2005. "Where Frankfurt and Strawson Meet." *Midwest Studies in Philosophy* 29: 163-80.
- McKenna, M 2008. "Ultimacy & Sweet Jane." In Nick Trakakis and Daniel Cohen, eds., *Essays on Free Will and Moral Responsibility*. Cambridge Scholars Publishing: 186-208.
- McKenna, M. 2008a, "Ultimacy & Sweet Jane." In Nick Trakakis and Daniel Cohen, eds., 2008. *Essays on Free Will and Moral Responsibility*. Cambridge Scholars Publishing, pp. 186-208.
- McKenna, M. 2008b, "The Limits of Evil and the Role of Moral Address: A Defense of Strawsonian Compatibilism" in *Free will and reactive attitudes: perspectives on P.F. Strawson's "Freedom and resentment"*, eds. M. McKenna & P. Russell, Ashgate, Burlington, VT; Farnham, England, pp. 201.
- McKenna, M. 2008c. "A Hard-line Reply to Pereboom's Four-case Argument" *Philosophy and Phenomenological Research* 77 (1): 142-59.

- McKenna, M. 2013. "Reasons-Responsiveness, Agents, and Mechanisms." In David Shoemaker, ed., 2013. *Oxford Studies in Agency and Responsibility*, vol. 1. Oxford: Oxford University Press: 151-84.
- McKenna, M. 2012. *Conversation and Responsibility*, Oxford University Press: Oxford.
- McKenna, M. 2019. "Watsonian Compatibilism." In *Oxford Studies in Agency and Responsibility*, Justin Coates and Neal Tognazzini, eds. Vol. 5.
- McKenna, M. and J.D. Coates, , "Compatibilism: State of the Art", *The Stanford Encyclopedia of Philosophy* (Winter 2018 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/win2018/entries/compatibilism/supplement.html>.
- Mele A.R., 2006 *Free will and luck*, Oxford University Press: Oxford.
- Mele. A. 1995. *Autonomous Agents*. New York: Oxford University Press.
- Mele. A. 2005. "A Critique of Pereboom's Four-Case Argument for Incompatibilism." *Analysis* 65: 75-80.
- Mele. A. 2006. *Free Will and Luck*. New York: Oxford University Press.
- Moore, G. E. 1912. *Ethics*. Oxford: Oxford University Press.
- Nelkin, D. 2011. *Making Sense of Freedom and Responsibility*. Oxford University Press.
- Nelkin, D. 2013. "Reply to Critics." *Philosophical Studies* 163: 123-131.
- Nichols, S. 2004. *Sentimental Rules: On the Natural Foundations of Moral Judgment*. Oxford University Press.
- Nickel, B. "Dutchmen are Good Sailors: Generics and Gradability", in *Genericity* eds. Mari, A., C. Beyssade, and F.D. Prete, Oxford: Oxford University Press: 390-405.
- Nickel, B. 2008. "Generics and the Way of Normality", *Linguistics and Philosophy*, 31(6): 629-648.
- Nickel, B. 2016, *Between Logic and the World: An Integrated Theory of Generics*, Oxford: Oxford University Press.
- Nozick, R. 1981. *Philosophical Explanations*. Cambridge, Mass: Harvard University Press.
- Nussbaum, M. 1987. "Non-Relative Virtues: An Aristotelian Approach". *Midwest Studies in Philosophy*. 13 (1):32-53.
- Nussbaum, M. 1986. *The Fragility of Goodness: Luck and Ethics in Greek Tragedy and Philosophy*. Cambridge University Press.

- Nussbaum, M. 2004; 2009, *Hiding from Humanity: Disgust, Shame, and the Law*, Princeton University Press, Princeton, N.J.
- O'Connor, 1995. "Agent Causation", in *Agents, Causes, and Events: Essays on Indeterminism and Free Will*, ed. T. O'Connor, New York: Oxford University Press: 173-200.
- O'Connor, T. 2000. *Persons and Causes: The Metaphysics of Free Will*. Oxford; New York: Oxford University Press.
- Pelletier, F.J. and N. Asher, 1997, "Generics and Defaults", in *Handbook of Logic and Language*, J. van Benthem and A. Meulen (eds.), Cambridge, MA: MIT Press, pp. 1125–1179.
- Pereboom D. 2014 "Kant on Transcendental Freedom". *Philosophy and Phenomenological Research* 73 (3):537-567.
- Pereboom D. 2008. "A Hard-line Reply to the Multiple-Case Manipulation Argument." *Philosophy*
- Pereboom D. 2001. *Living Without Free Will*. Cambridge, UK: Cambridge University Press.
- Pereboom D. 2014a. *Free Will, Agency, and Meaning in Life*. New York: Oxford University Press.
- Pereboom D. 2014b, "Kant on Transcendental Freedom". *Philosophy and Phenomenological Research* 73 (3):537-567.
- Rosenblatt, R. 1982, *The Man In The Water*, Time Incorporated, New York.
- Rozin, P. & Fallon, A.E. 1987, "A Perspective on Disgust", *Psychological review*, vol. 94, no. 1, pp. 23-41.
- Rozin, P., Haidt, J. & McCauley, C.R. 2008, "Disgust" in *Handbook of Emotions*, eds. M. Lewis, J.M. Haviland-Jones & L.F. Barrett, 3rd Edition edn, Guilford Press, New York, NY, pp. 757-776.
- Rozin, P., Markwith, M. & Nemeroff, C. 1992, "Magical Contagion Beliefs and Fear of Aids", *Journal of Applied Social Psychology*, vol. 22, no. 14, pp. 1081-1092.
- Rozin, P., Millman, L. & Nemeroff, C. 1986, "Operation of the Laws of Sympathetic Magic in Disgust and Other Domains", *Journal of personality and social psychology*, vol. 50, no. 4, pp. 703-712.
- Russell, P. 1992. "Strawson's Way of Naturalizing Responsibility." *Ethics* 102: 287-302.
- Russell, P. 2004. "Responsibility and the Condition of Moral Sense." *Philosophical Topics* 32: 287-306.
- Russell, P. 2017. *Limits of Free Will*. New York, NY: Oxford University Press.
- Sartorio, C. 2016. *Causation and Free Will*. Oxford: Oxford University Press.
- Sartorio, C. 2005. "Causes as Difference-Makers." *Philosophical Studies*, 123(1), pp.71–96.
- Scanlon, T.M. 2008. *Moral Dimensions: Permissibility, Meaning, Blame*. Cambridge, Mass: Belknap Harvard Press.

- Schlick, M. 1939. "When is a Man Responsible?" In Schlick, M. , *Problems of Ethics*. Prentice-Hall
- Schubert, L.K. and F.J. Pelletier. 1987, "Problems in Representing the Logical Form of Generics, Bare Plurals, and Mass Terms", in *New Directions in Semantics*, E. Lepore eds., London: Academic Press: 385–451.
- Shabo, S. 2011. Why Free Will Remains a Mystery. *Pacific Philosophical Quarterly* 92 (1): 105-125
- Shabo, S. 2012. "Where Love and Resentment Meet: Strawson's Interpersonal Defense of Compatibilism." *Philosophical Review* 121: 95-124.
- Shoemaker, D. 2007, "Moral Address, Moral Responsibility, and the Boundaries of the Moral Community", *Ethics*, vol. 118, no. 1, pp. 70-108.
- Shoemaker, D. 2015, *Responsibility from the Margins*. Oxford, UK: Oxford University Press.
- Smilansky, S. 2000, *Free Will and Illusion*, Clarendon Press: Oxford.
- Smith, A. 2012 'Moral Blame and Moral Protest'. In Justin Coates and Neal Tognazinni (eds.), *Blame: Its Nature and Norms*. Oxford: Oxford University Press, 27–48.
- Smith, A. 2015. "Responsibility as Answerability." *Inquiry*, 58 (2): 99-126
- Strawson, G. 1986, *Freedom and Belief*, Oxford: Clarendon Press.
- Strawson, P. F. 2008/1962, "Freedom and Resentment" in *Free Will and Reactive Attitudes: Perspectives on P.F. Strawson's "Freedom and resentment"*, eds. M. McKenna & P. Russell, Ashgate, Burlington, VT; Farnham, England, pp. 19.
- Strawson, P. F. 1980. "Reply to Ayer and Bennett." In Zak van Straaten, ed., 1980. *Philosophical Subjects: Essays Presented to P.F. Strawson*. Oxford: Clarendon: 260-66.
- Strawson, P. F. 1985. *Skepticism and Naturalism: Some Varieties*. New York: Columbia University Press.
- Strawson, P. F. 1992. *Analysis and Metaphysics: An Introduction to Philosophy*. Oxford, UK: Oxford University Press.
- Strawson, P.F., Strawson, G. & Montague, M. 2011, *Philosophical writings*, Oxford University Press, Oxford.
- Strohmingner, N. 2014, "Disgust Talked About", *Philosophy Compass*, vol. 9, no. 7, pp. 478-493.
- Stroud, B. 1968. "Transcendental Arguments". *The Journal of Philosophy*, 65(9), 241-256.
- Stump, E. 1988. "Sanctification, hardening of the heart, and Frankfurt's concept of free will". *Journal of Philosophy* 85 (8):395-420

- Stump, E. 2003. *Aquinas*. Routledge
- Swinburne, R. 2013. *Mind, Brain, and Free Will*. Oxford: Oxford University Press.
- Todd, P. 2016, "Strawson, Moral Responsibility, and the "Order of Explanation": An Intervention", *Ethics*, vol. 127, no. 1, pp. 208-240.
- van Inwagen, P. 1983. *An Essay on Free Will*. Oxford: Clarendon Press.
- van Inwagen, P. 2000. "Free Will Remains a Mystery," *Philosophical Perspectives* 14: 1-19.
- van Inwagen, P. 2002. "The Mystery of Metaphysical Freedom," in Robert Kane ed., *Free Will*, Oxford: Blackwell Readings in Philosophy, pp. 189-195.
- van Inwagen, P. 2017. *Thinking about Free Will*. Cambridge: Cambridge University Press
- van Riel, Raphael and Van Gulick, Robert. "Scientific Reduction", *The Stanford Encyclopedia of Philosophy* (Winter 2016 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/win2016/entries/scientific-reduction/>>.
- Vargas, M. 2018. "The Social Constitution of Agency and Responsibility: Oppression, Politics, and Moral Ecology." In *Social Dimensions of Moral Responsibility*. Oxford University Press, Chapter 5.
- Vargas, M. 2013. *Building Better Beings*. Oxford, UK: Oxford University Press.
- Väyrynen, P. "Thick Ethical Concepts", *The Stanford Encyclopedia of Philosophy* (Fall 2017 Edition), Edward N. Zalta (ed.), URL = <<https://plato.stanford.edu/archives/fall2017/entries/thick-ethical-concepts/>>.
- Velleman, J.D. 1992. "What Happens When Someone Acts?" *Mind* 101:462-81.
- Vihvelin, K. 2013. *Causes, Laws, & Free Will: Why Determinism Doesn't Matter*. New York: Oxford University Press.
- Wallace, R.H. 2019. "Responsibility and the Limits of Good and Evil," *Philosophical Studies*, 176 (10): 2705-2727.
- Wallace, R.J. 1994, *Responsibility and The Moral Sentiments*, Harvard University Press, Cambridge, Mass.
- Watson, G. 1987 "Responsibility and the Limits of Evil: Variations on a Strawsonian Theme" in *Free will and reactive attitudes: perspectives on P.F. Strawson's "Freedom and resentment"*, eds. M. McKenna & P. Russell, Ashgate, Burlington, VT; Farnham, England, pp. 115.
- Watson, G. 1996. "Two Faces of Responsibility." *Philosophical Topics* 24 (2): 227-48.
- Watson, G. 2004. *Agency and Answerability*. New York: Oxford University Press.

- Watson, G. 2014. "Peter Strawson on Responsibility and Sociality." In David Shoemaker and Neal A. Tognazzini, eds., 2014, *Oxford Studies in Agency and Responsibility, vol. 2: 'Freedom and Resentment' at 50*. Oxford: Oxford University Press: 15-32.
- Wenzel, A. and C. Lystad, 2005, "Interpretation Biases in Angry and Anxious Individuals." *Behaviour Research and Therapy* vol. 43, no. 8, pp. 1045-1054.
- Wiggins, D. 1973. "Towards a Reasonable Libertarianism." In Ted Honderich, ed., 1973. *Essays on Freedom and Action*. London: Routledge & Kegan Paul: 31-62
- Wolf, S. 1981, "The Importance of Free Will", *Mind* vol. 90, pp. 386-405.
- Wolf, S. 1990. *Freedom within Reason*. Oxford: Oxford University Press.
- Yablo, S. 1993. "Is Conceivability a Guide to Possibility?" *Philosophy and Phenomenological Research* 53.1 (1993): 1-42.
- Zagzebski, L. 2003. "Emotion and moral judgment." *Philosophy and Phenomenological Research*, 66 (1):104–124.