THE MINIMAL APPROVAL VIEW OF ATTRIBUTIONAL-RESPONSIBILITY

by

August G. Gorman

A Dissertation Presented to the
FACULTY OF THE USC GRADUATE SCHOOL
UNIVERSITY OF SOUTHERN CALIFORNIA

In partial fulfillment of the requirements of the degree of
DOCTOR OF PHILOSOPHY
(PHILOSOPHY)

August 2018

to Sam, for believing in my dreams, for believing in us, and for believing that those two things were compatible.

Acknowledgments

I do not come from a family of academics. I don't think my mother understood why, after breaking my leg on the free trip to Aruba she'd won and had taken me with her on after my second year of grad school, I was content to have her wheel me out to read Susan Wolf's *Freedom Within Reason* on the beach. Nor do I think she understood my excitement when I read a passage in that book in which Wolf briefly floats the idea that desires cannot speak for us as agents because "the prospect of satisfying such desires may not be preferable to the prospect of eliminating these desires in other ways."[1] (I thought I had discovered right then and there the answer to the question of attributable agency. I still think so, but it turns out that fleshing out the account was a bit more complicated, and occupied me plenty for the next three years.) My parents have a running joke about which of the two of them I got my "smarts" from, the joke being that my interest in abstract conceptual matters seems to come completely out of left field. I don't know the degree to which they realize just how much of my ability to carry out a project like this *was* shaped by them: how I learned to have a critical eye and a dogged persistence from my dad, how I learned to seek out only what actually interests me and to forge a path that is unequivocally my own from my mom. For that, and for their unwavering emotional and financial support over the course of this journey, I am truly grateful.

I didn't have the slightest idea what philosophy was when I arrived at college at the University of Mary Washington, but for whatever reason had a hunch that I might like it. I owe a tremendous debt to the professor who taught my first philosophy class, Nina Mikhalevsky, who made it seem like the world was endlessly full of puzzles, the pursuit of any of which might be worth devoting a lifetime to solving. My grad school aspirations were cemented by Sam Emswiler, who broke her personal ban on ever recommending that a student go on to grad school in philosophy to suggest that despite the odds, in my case I absolutely had to. I was given the autonomy to pursue a fairly idiosyncratic research project in thinking

[1] Wolf (1990): 31.

about how first-personal accounts of the lived experience of Tourette syndrome line up with Harry Frankfurt's discussion of agential alienation under the guise of a senior independent study with my ever-supportive undergrad advisor, Jason Matzke. For that I am especially grateful, as much of that research informs my work even today.

I wouldn't trade my experiences at the University of Mary Washington for the world, but there was a learning curve between undergrad and grad school that at many times felt insurmountably steep. Having come from a relatively unknown public school in Fredericksburg, Virginia, the first hurdle was merely to actually get in. I count myself incredibly lucky to have come across an application online for the summer seminar Bob Pasnau runs for students from relatively unknown colleges with philosophy grad school aspirations. Through working with Bob, who I must thank for seeing early signs of potential in my work, I honed a writing sample and gained a letter of recommendation that I am sure I have, in great part, to thank for my acceptances.

After visiting six other schools, I was sold on USC after a thirty minute Skype call with Mark Schroeder. From working with me on issues of accessibility my first year, to coordinating my job market letters with a very quick turnaround time, I am thankful for Mark's continued support and commitment to making sure no USC grad student ever slips through the cracks. Thanks are due to all of the professors I worked with at USC, but especially to Kadri Vihvelin, Elyn Saks, and Janet Levin, who each influenced the early stages of development of this dissertation in different ways. Kadri Vihvelin wrote a comment on a draft of mine that read "weakness of will is unfreedom?!!!"Answering that worry grew into Chapter 3 of this dissertation. Working with Elyn Saks, whose impressive body of work builds off of her own experiences with mental health, inspired me to have the courage to explicitly write about my own disorder, misophonia, in this dissertation, and I think my discussion is greatly enriched for it. Janet Levin served as a member of my qualifying committee and gave very helpful comments on early drafts of several chapters.

Of course my biggest debt of gratitude at USC is to my committee. Thanks to John Monterosso for serving as my external committee member, and for encouraging me to clarify the methodological underpinnings of my project. Thanks to Steve Finlay for playing the role of my greatest critic and for giving me incredibly detailed comments on several chapters. Knowing that this dissertation would be read by at least one person who is deeply suspicious of the very idea of a Deep Self view has been instrumental in broadening the audience I aim to speak to in my work. Thanks are due to Dave Shoemaker, first, for writing *Responsibility from the Margins*, a book I read closely and excitedly in a reading group my third year that showed me what an exemplar of the kind of work I wanted to do might look like. Dave's work continually shows a commitment, first and foremost, to paying careful attention to what our responsibility practices imply for how we treat the real and imperfect people in our lives, one which I have tried to emulate in my own work. Thanks also for going out of your way to support my work and serve as a member of my committee even amidst your research and teaching obligations at your own university.

Thanks to Gary Watson for several semesters of independent studies and countless long meetings, helping me both to chart the lay of the land in the free will and moral responsibility literature and to find my own voice within it. Thank you for helping me learn how to emulate you in prioritizing wisdom over cleverness, and for being a sounding board for the earliest incarnations of the ideas this dissertation builds upon.

I owe a larger debt of gratitude to my advisor, Jon Quong, than I can adequately put into words. On the advice of other graduate students I approached Jon at the end of my third year with what I knew to be an outlandish request. I asked him to be my advisor despite the fact that I:

1. hadn't so much as introduced myself to him in the three years I'd been at USC,
2. didn't work in his area, and
3. was moving, and so would have to work entirely remotely via Skype.

For whatever reason, he agreed instantly, and I could not be more thankful that he did. Over the years he has proven himself to be nothing short of the Platonic ideal of an advisor: critical, kind, thorough, encouraging, and trusting. He has read and given me extremely helpful feedback on more drafts of this material than I can count. I am not sure whether our working styles just happen to line up perfectly or if he has an uncanny ability to intuit exactly what would best suit each of his advisees, but a through a combination of faith in my ability to set my own goals and deadlines and encouragement that my projects were viable even in the face of criticism, he helped me find my confidence at a time when it was sorely needed. This dissertation simply would not have come into existence without his guidance and encouragement.

I am thankful also for the community of grad students at USC. In particular, I don't know where I would be without heart-to-heart conversations over the years with Aness Webster, Kenneth Silver, and Ara Astourian or without having Rima Basu, Renee Jorgensen Bolinger, and Maegan Fairchild in the cohorts above me to give me advice. To Caleb Perl, Abelard Podgorski, Nick Laskowski, Alex Dietz, Joe Horton, Nathan Robert Howard, Jonathan Charles Wright, and anyone else whose money I took at our weekly poker game, thank you for fostering a sense of camaraderie and for being a part of my favorite thing in LA. I am especially grateful for Tanya Kostochka, for making sure I didn't stay holed up in my apartment doing work 24-7, for going through the whole grad school experience as thick as thieves with me from day one.

Outside of USC, I am truly grateful for the incredibly kind philosophers in the agency and responsibility subfield who instantly welcomed me into their ranks and made me feel like I had something important to contribute to the field. Without their encouragement I doubt I would have had the nerve to take on a project this ambitious. Special thanks to Michael McKenna, Tamler Sommers, Phil Swenson, Matt Talbert, Justin Coates, Neal Tognazzini, Gunnar Björnson, Ben Mitchell-Yellin, and Ben Matheson for all the conversations at conferences, often over beers, that in one way or another informed my perspective on this material.

As for my chosen family from the East Coast, I never could have done this without your support. Thanks to Eugene Chislenko for bringing me in to guest lecture on this material at Temple, and for helping fund my last summer of writing by letting me be your research assistant. Thanks to Geena Cain for being the least excited person about me going off to grad school in LA, but supporting me every step of the way nonetheless. Thanks to Eben Hilpert for helping me learn how to take small breaks and relax. And thanks to Elsie Bartlett for helping me get through those first couple years and for making me a home-cooked dinner every single night of my most difficult semester.

Kari David Collins is the person I narrate my life to in real-time, and so over the past few years has heard more than any layperson would ever reasonably want to know about the day-to-day grind of writing a dissertation on attributional-responsibility. I have so much to thank you for, but especially for always being my listening ear and for every single trophy emoji that kept me going. (And for playing the Galinda to my Elphaba on head-clearing sing-alongs to the *Wicked* cast album on our cross-country road-trips.)

Finally, thank you to my partner, Sam Cirulis. It's been quite a journey, to say the least. Thank you for trusting that we could make a long-distance relationship work despite meeting me right before I flew off to LA for a PhD program. Thank you for flying out to see me even that one time it meant getting stranded in Dallas due to Spirit Airlines, thank you for bringing joy into my life during the toughest times along this journey, and thank you for helping me land my postdoc by pretending to be every possible archetype of job interviewer while I was practicing. There have been many days when you single-handedly kept me going, and I hope you know that I intend to spend the rest of our lives together returning the favor.

-*A.G.G.*

Table of Contents

Chapter 1: *A Common Necessary Condition for Attributional-Responsibility*

1. Introduction

Imagine that your friend Corey has promised to come over and support you after the death of your pet. You sit waiting for him to appear at the time he said he'd be there, but time continues to tick by and he does not show up. As it becomes clear that he's not going to show up, you might blame Corey. In response to the perceived slight, you might start to feel angry with him, come to judge him as being insensitive, or even begin to question your friendship. But now, suppose you find out that Corey suffers from OCD and his absence is explained by the fact that while he desperately wanted to come support you, he felt the need to instead act on a compulsive desire to stay at home repeatedly performing rituals, turning the lights in each room of his house on and off several times.

It seems that you have reason to suspend your reactions to your friend in light of finding out this information. You shouldn't blame Corey because it seems he wasn't, after all, being insensitive about your situation, since his absence was caused by his compulsive behavior. But why do conditions like Corey's OCD exempt agents from moral responsibility in these sorts of situations?

Philosophers often ask questions like this in the context of a skeptical worry about how it could be the case that anyone is ever responsible for what they do at all. One commonplace thought is that when agents' behaviors are due to compulsion their "brains make them do it." But, in a literal sense, it might seem that our brains make all of us do everything that we do.

Given a commitment to the veridicality of this important intuitive distinction between ordinary responsible action and compulsive behaviors, there are two central kinds of approaches to these worries in the moral responsibility literature. One approach is to hold that what we really mean is that agents like Corey don't have sufficient control over their compulsive actions; being morally responsible is largely a matter of control. Another approach is that what we really mean is that when agents like Corey act compulsively they don't act in accordance with what they *really* want to be doing; moral responsibility is a matter of agents having a certain profile of mental states that contribute to their actions.

In this dissertation I advance an account of moral responsibility that aims to solve several of the biggest problems facing current accounts that take this latter approach, which are often known as Deep Self accounts. In the first part of this chapter I explain and motivate several key aspects of Deep Self accounts. While the view I advance diverges in crucial ways from traditional Deep Self views, it nevertheless shares many of their broad theoretical commitments. My discussion here serves both to shed light on the general theoretical orientation that underpins my project and to highlight the specific contribution that my view makes to the dialectic.

Each traditional Deep Self view relies on identifying a type of mental state that is invariably internal. In the second part of this chapter I argue that internality is best understood on each view in terms of the agent approving of being motivated in the way that she is to some degree. This helps locate a commonality among Deep Self views: they all seem to hold that approving of one's action is a necessary condition for attributional-responsibility. In Chapter 2, I build on this foundation to develop a new view of attributional-responsibility.

2. Actual Sequence Compatibilism

2.1 Dialectical Motivation for Actual Sequence Compatibilism

How can we justify our system of moral responsibility practices given a scientific picture of our world in which it seems likely that our choices are explicable by means of causal chains that stretch back to events that happened before we were even born? Against a backdrop of scientific understanding according to which we recognize that our brains make us do everything that we do, and that these events are reducible to neurochemical reactions, how can we preserve the distinctions we want to make between agents who are blameworthy for their actions and those who are not due to conditions like Tourette syndrome or OCD?

One response to these initial concerns is to claim that, in fact, we cannot justify our practices and intuitive distinctions. The most popular argument for this conclusion runs as follows:

1. It's only appropriate to blame agents who act freely.
2. Acting freely requires the presence of alternative possibilities for action.
3. Having alternative possibilities involves having metaphysically robust options available at the moment of choice.
4. Our scientific picture of the world is correct, and it crowds out the space for metaphysically robust options to ever exist at the moment of choice.[2]
5.Therefore, it is never appropriate to blame anyone. [3]

[2] Another route to Moral Responsibility Skepticism is to take no stand on whether or not we live in a deterministic world, and instead show that the kind of free will that could ground moral responsibility is incompatible with both determinism and indeterminism. See, for example, Pereboom (2005).

[3] While this argument is very popular, it is possible to be an Incompatibilist about moral responsibility and determinism without thinking that alternative possibilities are required for free will. For example, some Incompatibilists think that the kind of free will required for moral responsibility requires that an agent herself be the ultimate *source* of the chain of events leading to her actions in a way that is not possible given the truth of determinism. For an overview of several different views in this family, see Tognazzini (2011).

While Moral Responsibility Skeptics are happy to accept this conclusion, each of the premises of this argument has been disputed somewhere in the literature.

Libertarians generally dispute Premise 4. They believe that for an agent to have the requisite kind of alternative possibilities when she acts her action must be non-deterministically caused such that at the moment of action there is another action she could have undertaken (and that afterwards it will be true to say of her that she "could have done otherwise"). While this sounds like a mysterious sort of power to ascribe to agents, Libertarians have advanced several different pictures of the production of action that attempt to mitigate these worries.[4] Libertarian views are also subject to luck objections.[5] If an agent's prior plans, values, commitments, etc. cannot fully determine her course of action since, according to the Libertarian, the choice must remain open at the very moment of action, there can be nothing about the agent's mental states that determines whether she will undertake one action or the other. This agent actually seems to lack an important form of control over what she does, and so, it is often argued, this cannot be the sense of alternative possibilities required for the kind of free will required for moral responsibility.

For a long time the most popular strategy for those who wanted to preserve both the scientific picture of action production and justification of our practices of responsibility was to target Premise 3. Classical Compatibilists accepted that moral responsibility requires free will and that free will requires the ability to do otherwise, but rejected that this is incompatible with determinism. Instead, they attempted to posit less metaphysically robust senses of the ability to do otherwise that were both compatible with causal determinism and convincingly able to ground ascriptions of free will and moral responsibility. Initially, these compatibil-

[4] See Clarke (2003) for an overview.

[5] See Mele (1999, 2006).

ists put forth conditional analyses of the ability to do otherwise that took the following form: If the agent had chosen/wanted to/decided to do otherwise, she would have done otherwise. In this way, they hoped to secure the conditions for free will while retaining the understanding that our actions are controlled by fixed chains of mental events.

These analyses are thought to have definitively failed, however, since they appear to predict that agents have the ability to do otherwise in cases in which we think they clearly don't. For example, imagine a girl who has such a psychological aversion to picking up blonde puppies such that she could never become such that she would pick up a blonde puppy. Now suppose that when offered a choice of a blonde or black puppy, she picks up the black puppy. The conditional analysis says of this case that if she had wanted to pick up the blonde puppy she would have picked up the blonde puppy and so it returns the verdict that she does have alternative possibilities in the Classical Compatibilist sense. But since she is unable to become psychologically such that she would ever pick up the blonde puppy, this seems clearly incorrect. In what sense does she have the ability to pick up the blonde puppy if she could never be or become such that she would actually do so?[6]

More recently, some Compatibilists, sometimes called New Dispositionalists, have tried to revive the spirit of this project by positing that the ability to do otherwise is grounded in more complex agential dispositions, or bundles of dispositions.[7] For example, following David Lewis, Kadri Vihvelin suggests that rather than simple conditionals, the dispositions at issue should be taken to involve intrinsic properties that are the causal ba-

[6] For more on this example, see McKenna and Coates (2015). This problem is, arguably, just an application of a more general problem for analyzing dispositions in terms of simple conditionals, known as the problem of finked dispositions. See Vihvelin (2004) for discussion.

[7] This label first appears in Clarke (2009). Prominent defenses include Smith (2003); Fara (2008); and Vihvelin (2004, 2013).

ses of those dispositions. She thinks that in order to assess claims about whether an agent has alternative possibilities, we must consider various counterfactual scenarios in which the causal base of the pertinent disposition operates unimpaired. In the spirit of Classical Compatibilism, Vihvelin argues that the relevant sorts of abilities are dispositions to make choices on the basis of reasons.[8] While these sorts of accounts solve some of the problems of the more simplistic conditional analyses, they continue to face a less resolvable challenge from defenders of Premise 3.

Defenders of Premise 3 argue that the New Dispositionalists' accounts do not get us to the kind of alternative possibilities required for free will, since what's valuable about us having alternative possibilities at the moment of choice is that they make it so that our choices are "up to us." And, as they argue, it's just not clear that the New Dispositionalists' notion of ability to do otherwise captures this sense. Randolph Clarke puts the point as follows:

> An agent with an interesting bundle of dispositions and in friendly surroundings might have a rich array of narrow and wide abilities to do things that she doesn't in fact do. That an agent might have such abilities even if determinism is true is an important fact. But it will take further argument to show that having such dispositions and being in such surroundings suffices for its being up to you, on some occasions, whether one or another of these dispositions is manifested, and hence whether you do this or that.[9]

[8] Vihvelin (2004) suggests that the collection of dispositions might include, among others, dispositions to form and revise beliefs in response to evidence and argument, to form intentions in response to desires and form beliefs about how to achieve those desires, to engage in practical reasoning in response to one's intention to make a rational decision about what to do, and to believe that by engaging in practical reasoning one will succeed in making such a decision.

[9] Clarke (Forthcoming): 26-27.

Here, New Dispositionalists find themselves in a battle of intuitions; they allege that having the sorts of dispositions they put forth as an analysis of the ability to do otherwise *just is* what it means for our actions to be "up to us."[10] It is unclear how the clash of this particular set of intuitions could be resolved, and this leaves the debate at a standstill. Some have even taken this clash of intuitions as reason to conclude that our notions of free will and moral responsibility are confused concepts we should try to do without in order to make progress.[11]

I adopt what I take to be a more promising strategy for those who want to preserve both our moral responsibility practices as we know them and the scientific picture of action: Actual Sequence Compatibilism. Actual Sequence Compatibilism gives up on the Classical Compatibilist attempt to redefine the sense of alternative possibilities that is required for freedom and responsibility and instead resists the notion that the conditions for moral responsibility have anything to do with control or alternative possibilities at all, thus targeting Premise 2 (and sometimes also Premise 1 in the process).[12] The idea is that the criterion for an agent to

[10] It is sometimes alleged that the incompatibilist intuition about the relevance of alternative possibilities is a common sense intuition, while others argue that compatibilist notions of alternative possibilities are the more predominant common sense intuition. There is an extensive but inconclusive experimental philosophy literature on this topic. See Björnsson and Pereboom (Forthcoming) for an overview.

[11] See Vargas (2011) for an overview of Revisionist approaches to free will and moral responsibility.

[12] In this dissertation I reject premise 2 explicitly and remain agnostic about premise 1. I propose conditions for one form of moral responsibility, and say relatively little about free will. I leave the question of whether or not my account is best thought of as a non-freedom-requiring account of responsibility ("Semicompatibilism"), or as an account of the kind of freedom required for responsibility to the reader (as the answer seems to me to be determined in large part by what one wants their conception of "free will" to do), although I suspect the former may be the better way of understanding the project.

count as having acted (freely and) responsibly can be found within the actual sequence that leads to her action. In arguing that exemptions from moral responsibility can be explained without reference to alternative possibilities, the Actual Sequence Compatibilist boldly forges past the stalemate and frees up the opportunity to propose quite different kinds of criteria for responsibility. This gives the Actual Sequence Compatibilist a significant dialectical advantage in the debate.

2.2 Support for Actual Sequence Compatibilism

Dialectical advantages aside though, are there any reasons to believe in the truth of Actual Sequence Compatibilism? There is, admittedly, at least something initially strange about the thought that whether or not an agent could have done otherwise has no bearing on whether or not she acts freely and responsibly.

This strangeness, though, is mitigated by the fact that there is something equally strange about views that do not focus on the actual causal sequence that leads to the agent's action. As Carolina Sartorio emphasizes, it would seem quite inappropriate for an agent to attempt to absolve herself of responsibility for some action by pointing to factors that were not in any way explanatory of why she acted in such a way. Sartorio puts this point succinctly: "if a factor is completely irrelevant to why you acted, it seems that it cannot be used to excuse your behavior."[13] Since there are intuitions that tell against the relevance of alternative possibilities and the irrelevance of alternative possibilities, it's clear the debate cannot be settled by appeal to initial intuition alone.

In various places throughout the body of his work, Harry Frankfurt provides support for the view that alternative possibilities are irrelevant to moral responsibility. Most famously, Frankfurt offers the case of Jones and Black.[14] He has us imagine that Jones is an agent who is going to perform

[13] Sartorio (2016): 2. See also Mele (2006).

[14] Frankfurt (1969).

an action, φ-ing, for which we would all agree he would intuitively be clearly morally responsible, were he to go through with it. Unbeknownst to him, however, Black is waiting in the wings and has placed a chip in Jones' brain that will not activate unless Jones is *not* independently going to go through with φ-ing, in which case it will cause him to φ anyway. As it happens, Jones decides to go through with φ-ing on the basis of his own deliberation and Black never needs to, and so does not, interfere. Jones lacks alternative possibilities in any reasonable sense; no matter what, he was going to φ at *t*. And yet he still appears to be just as morally responsible for his action as he would have been if Black had never been waiting in the wings at all.

Cases that follow this general format have been termed "Frankfurt cases," and a very large literature has emerged over the years that questions the adequacy of the thought experiment for proving that the presence of alternative possibilities is not required for moral responsibility.[15] For example, some question whether or not it is methodologically appropriate to make the assumption that Black knows what Jones will do before he does it, and others question whether or not the action Black would cause would be identical to Jones' actual act.

Whether or not any particular Frankfurt case is successful at decisively establishing a foolproof example of an agent who has no alternative possibilities but is nevertheless responsible for her action, there is a broader lesson that these cases do help to illustrate: alternative possibilities appear to play no *explanatory* role in action.[16] In Frankfurt's words:

> The fact that a person could not have avoided doing something is a sufficient condition of his having done it. But…this fact may play no

[15] See Fischer (2010) for an overview in which he compares the state of the literature here to the state of the literature surrounding Gettier cases.

[16] See also McKenna (2008) for a similar take on the relevance of Frankfurt-cases.

> role whatsoever in the explanation of why he did it. It may not figure at all among the circumstances that actually brought it about that he did what he did, so that his action is to be accounted for on another basis entirely. Even though the person was unable to do otherwise, that is to say, it may not be the case that he acted as he did *because* he could not have done otherwise.[17]

Frankfurt cases remind us that it seems much less important to figure out whether or not there are possible worlds in which Jones does not φ than it does to figure out why he actually φs.

This lesson can also be drawn from Frankfurt's cases of volitional necessity. Frankfurt draws attention to cases in which people feel that their actions are necessitated by constraints on their wills in positive ways. For example, agents committing acts of extreme love or loyalty may have this feature, like a woman jumping in front of a bullet to save a friend or a man running back into a house on fire to save his child. Doing otherwise in these cases appears to be simply unthinkable for such agents, yet they certainly seem to be morally responsible. As Frankfurt remarks, when Martin Luther made his famous declaration "Here I stand; I can do no other," we do not usually take Luther's seeming lack of control over his course of action to undercut his claim over his action, but rather, if anything, to intensify it.[18] Whether or not Luther's claim is taken literally, the rhetorical force comes from the natural thought that feeling yourself to have no other option often serves to intensify your sense of ownership of your action, not to undercut it. As with Frankfurt cases, the larger point is not that the exact sense of alternative possibilities targeted by Classical Compatibilists, Libertarians, or Skeptics is the exact same as what agents lack in volitional necessity cases, but rather that the degree to which someone is able to do

[17] Frankfurt (1988): 8.

[18] Frankfurt (1988): 87.

otherwise has no simple correlation to how free they are to act in any sense that we seem to care about when we make judgments of moral responsibility.

A final set of cases offered by Frankfurt helps to get to the heart of the debate: his willing and unwilling addict cases. The unwilling addict desperately wishes not to be compelled to take drugs but continues taking them, despite himself, due to his addiction. The willing addict, on the other hand, absolutely loves taking drugs, and though he would not be able to resist if he ever tried to, he would never want to resist in the first place since he fully endorses what he does. Those who are focused on alternative possibilities and control will not differentiate between the responsibility of willing and unwilling addicts; so long as an addict cannot sufficiently control her action, she is not responsible, they'll allege. But Frankfurt thinks the willing addict is responsible for taking drugs, despite his lack of control. This intuition can be strengthened by supposing that the willing addict has no awareness of his addiction and wholeheartedly chooses to take drugs for reasons that have nothing to do with their irresistibility.

Chandra Sripada adds further support by offering a structurally analogous case that abstracts away from the particularities of addiction. His example is of a "willing exploiter" who strongly desires to watch exploitative kinds of pornography.

> … suppose [his] desires and the actions they issue in are deeply expressive of his self. This man has a narcissistic kind of self-love at his core. He is attracted to the idea that he is in a position of dominance over others and the exploitiveness of the pornographic material is thus exactly what he finds so deeply gratifying. The person thus stands strongly in favor of his desires to view exploitive images and wouldn't change a thing. In envisioning this case, we are to keep all other relevant aspects of the Willing Addict case the same. In particular, the attitudes of this person's self, via their role in deliberation and the

> formation of practical judgments, provide motivational support in favor of his viewing the images. Additionally, the desires to view the images are sufficiently powerful in their own right that, though he doesn't and wouldn't ever try to resist these desires, were he to try, he would fail.[19]

The willing exploiter certainly seems to be morally responsible for his viewing of the pornographic images despite the fact that he wouldn't ever be able to resist his desires to do so if he were to try to, and so control, again, seems to be fairly irrelevant to our willingness to ascribe responsibility.

3. Deep Self Theory

These cases also point towards a new kind of criterion that has nothing to do with alternative possibilities or control. The willing exploiter seems responsible for what he does because it's something he embraces, and he does so precisely because he embraces it. This insight is key to the development of a new kind of criterion for morally responsible agency.

Recall that the Actual Sequence model claims that alternative possibilities are not relevant to moral responsibility and that we should instead focus on the actual sequence of events that leads up to the agent's action. But this falls short of a substantive theory of moral responsibility, since it does not yet tell us *which* aspects of the action's causal sequence are relevant to determining whether or not the agent is responsible. As the willing exploiter case illustrates, one natural way of filling in the story is by looking towards whether an agent does what she in fact "stands in favor of" doing. When agents stand in favor of the action they undertake, it seems reasonable to hold that they express something about what those agents are like.

[19] Sripada (2017): 802-803.

Views that feature these sorts of criteria for moral responsibility are variously referred to as "Self-Disclosure," "True Self," "Real Self," or "Deep Self" views. The most common title currently at use in the literature is "Deep Self," and Deep Self views easily represent the most influential strand of Actual Sequence Compatibilism.[20] On a Deep Self view, what matters for moral responsibility is that the agent acts in accordance with what she really wants to do, where "really wanting" always involves some further mental state beyond merely having a first-order desire. The necessity of locating some further mental state comes from the fact that an agent can act on one of her first-order desires without thereby standing behind it.

Cases of compulsion illustrate this point well. Compulsion seems best described as involving an agent being moved to φ *against her will* by a rogue first-order desire or urge to φ that overpowers her identification with some other course of action. Insofar as agents should not be held morally responsible for their compulsive actions, there is reason to locate the criterion for moral responsibility in the presence of some further kind of mental state. And so Deep Self theorists each make some sort of demarcation within agential psychology that explains how only some subset of an agent's motivational states can produce actions for which an agent can, in principle, be praiseworthy or blameworthy. An agent can only be held

[20] One source of potential confusion here is that views that posit that *control* over self-disclosing agential capacities as a condition on moral responsibility are occasionally referred to as Deep Self views despite not being Actual Sequence Views. Agnieszka Jaworska refers to these views as "Broad Identificationist Views" and Actual Sequence Deep Self Views as "Narrow Identificationist Views" (Jaworska [2017]). It is unclear, however, what precisely demarcates these Broad Identificationist Views from Classical Compatibilist views that posit that the relevant sense of alternative possibilities to ground moral responsibility has to do with counterfactual conditionals like "if one had chosen/valued/endorsed acting differently she would have acted differently" (or their New Dispositionalist equivalents). Here I will instead adopt the more popular taxonomy that considers Deep Self views to be a proper subset of Actual Sequence Views.

morally responsible when her action is motivated in the right sort of way because only then is her action is produced in such a way that it can "speak for" her as an agent.

Deep Self theorists generally proceed by proposing a special *kind* of mental state that motivational states must align with in order for an agent to be morally responsible. Many different Deep Self views have been put forward, but the views that have been most prominent in the literature are those that privilege either valuing, planning, caring, endorsing, or some combination of these states. I will refer to these as "deep self mental states."[21] These special mental states are said to "mesh" or "align" with the agent's motivational states such that the motivational states "flow from" the values, plans, cares, or endorsements.

While overly metaphorical language is often used to describe this relationship, there has been recent interest amongst Deep Self theorists in getting clearer about what the relationship of "meshing" might amount to. Many theorists seem to think about the relation as being causal: an agent is attributionally responsible for φ-ing if the motivational state that causes the agent to φ is itself caused in part by the agent's deep self mental states. It's worth noting that the deep self mental state need not make the differ-

[21] Views that put forth other candidate deep self mental states include Susan Wolf's "sane Deep Self view" on which deep self mental states must meet further "sanity" requirements (Wolf [1987]); David Velleman's view, on which deep self mental states are desires to act in accordance with reasons (Velleman [1992]); and coherentist views on which deep self mental states are those that bear special relationships to the agent's other mental states either by being relatively unopposed by other states (Arpaly and Schroeder[1999]) or by being narratively coherent (Matheson [2018]). While departing from traditional Deep Self views in significant ways, other views sometimes said to "strike deep self themes," appear in Scanlon (1998), Arpaly (2003), Smith (2005, 2008), Sher (2009), and Buss (2012). (Sripada [2016] offers the latter list with the caveat that these views may fail to count as "Deep Self" views on many uses of the phrase. In Chapter 5 I follow Talbert [2016] in referring to the views of Scanlon, Smith, and Sher as "New Attributionist" views.)

ence between the agent acting as she does and her refraining, since deep self mental states might also function by causally overdetermining the agent's course of action. One problem with broadly causal views, first noted by Neil Levy, is that it seems possible that an effective first-order desire and resultant action could be caused by a deep self mental state while intuitively the deep self mental state is not expressed through the action.[22] Chandra Sripada gives the following example, which uses the proposed deep self mental state of caring:

> Suppose Jimmy's son has gone missing in Afghanistan. He cares for his son so much that he ruminates continuously, and this in turn gives him a severe headache for which he must take an aspirin. Standard theories of causation would say that Jimmy's caring for his son causes his taking an aspirin— very roughly there is a chain of causal dependence that links the two. Jimmy's taking the aspirin, however, does not express his caring for his son.[23]

An alternate understanding of the expression relation is what Sripada and Shoemaker call a "content harmony" relation.[24] On this view, an agent is attributionally-responsible for her action only if the motivational state that she acts on is congruent with the content of the deep self mental state in some sense. On Sripada's understanding, the congruence amounts to the motivational state's being characteristically disposed to be produced by the deep self mental state. For example, if I judge my mother's health to be of value to me, I might be characteristically disposed to be motivated to take her to the doctor. A content harmony requirement might be added to a causal requirement, or it may be thought to be a competing explanation

[22] Levy (2011).

[23] Sripada (2006): 1216.

[24] See Sripada (2006), Shoemaker (2012, 2015b).

for what the expression relation is. For example, one possible view is that an agent is attributionally-responsible for φ-ing iff she endorses her desire to φ, whether or not that endorsement itself has any causal bearing on her φ-ing.

3.1 The *"Deep"* in "Deep Self"

Given this description of the commitments of Deep Self views I have outlined, one might wonder what cause there is for adopting the language of "*deep* self." For all I have said, these views merely target some special subset of mental states and propose that these states, rather than others, due to the fact that they ensure agential identification, grant an agent's actions the ability to speak for that agent. But what sort of additional commitments are taken on by adopting the language of the "deep self" and what role do they play in the view?

Answers to this question by leading Deep Self theorists are extremely varied. For example, David Shoemaker writes that

> the 'deep' in 'deep self' simply refers to the psychic element's place in an agential structure as the ultimate psychological *source* of various 'surface' attitudes subject to its governance.[25]

So, for example, cares are deeper than ordinary first-order desires since, for example, caring about your family is the *source* of a desire to take your daughter to soccer practice. The meaning of "deep" here does not imply any sort of strong metaphysical commitment to the Self. On the other end of the spectrum, Chandra Sripada thinks talk of deep selves commits him to the existence of "fundamental conative states that robustly and globally shape action," the existence of which he takes to be a substantive claim about actual human psychology.[26] [27] Deep selves, for Sripada, are presum-

[25] Shoemaker (2015b): 43.

[26] Sripada (Unpublished Manuscript): 15.

ably 'deep' because on his view they play a crucial role in helping to explain a wide variety of agential phenomena including but not limited to: moral responsibility, normative reasons for action, happiness, and weakness of will. A similar but, in theory, distinct idea is the thought that all of an agent's mental states of the kind that are proposed to play the role of deep self mental states together form some sort of whole which either constitutes or provides us with some particularly important insight into the agent's Self.

I take these latter two conceptions to be contingent features of the set of views generally recognized to belong to the Deep Self family of views. Each view does need some story to tell about what privileges actions that relate to deep self mental states such that they are the ones on the basis of which we are permitted to hold an agent responsible. However, the versions of this story on which the deep self mental states together play a foundational role in the core of an agent's conative personality or are together constitutive of the agent's Self only represent a couple of the options for fleshing out this story, among many other possibilities.

Further complicating these issues, as Lippert-Rasmussen points out, people tend to conflate two different connotations of the phrase "deep self." On one conception, deep self mental states have special *authority* for the agent, and on another, deep self mental states have more to do with *authenticity*. As he puts it, on authenticity conceptions of the Deep Self, a person's Deep Self

[27] While it does not seem to me that any such broad sweeping claims about human psychology are required for proponents of Deep Self, Sripada thinks there is much less cause for empirically-driven skepticism about the existence of such deep selves than what many philosophers have been led to believe. According to Sripada, while certain segments of social psychology have been very influential as a source of data for philosophers, data from neuroscience, human behavioral genetics, and personality psychology is all fairly friendly to the idea of robust deep self psychology.

> … is the person's deepest and most genuine commitments and desires…deep, idiosyncratic longings and repressed desires are strong candidates [for deep self states] on [this] account.[28] (20).

This conception of the deep self is, to my mind, not relevant to questions of moral responsibility. Confusion in the literature between the two senses of "Deep Self" is presumably part of what leads Nomy Arpaly to her particularly damning accusation of Deep Self views. She has us imagine a woman, Lynn, who discovers she is a lesbian but would much rather have not come to such a discovery and does not want to be motivated by such desires. Arpaly continues,

> If Lynn were to go to her favorite college professor for help, she would likely be told that she should try to accept herself for who she is, refrain from attempts to suppress her true self, and so on. If, on the other hand, she were to read the moral psychology literature and believe its claims, she would probably conclude that she was right and her homosexual desires are not truly her own. For 'Lynn' and 'homosexual desires', we could substitute 'Victorian lady' and 'any sexual desire', 'nice Jewish boy' and 'hostility toward parents', 'severe perfectionist' and 'desire to get some rest', 'the young E.T.A. Hoffman' and 'desire to be a writer', or any of various characters from various novels and their adulterous loves. In all these cases, the agent who dismisses these desires as reasons for action and treats them as "outlaw desires" is likely to feel that they are not really his.[29]

[28] Lippert-Rasmussen (2003): 20.

[29] Arpaly (2003), 16.

But it need not be any part of a Deep Self view to hold that Lynne's lesbian desires are not an authentic part of who she is or that she should resist them. A Deep Self view should merely say that if she were to engage in a sexual act with a woman without in some sense valuing/ endorsing/ planning on/ caring about doing so, her action would be compulsive or lacking in agential authorization in such a way that would undermine her being an apt candidate for moral responsibility. It is perfectly consistent to additionally hold that Lynne ought to embrace her lesbian desires as being an authentic part of her identity. Deep Self theorists ought to be clearer in rejecting the relevance of the authenticity conception of the Deep Self and instead understand deep self mental states as those that have the authority to speak for the agent; the mesh of deep self mental states with effective motivation needn't be understood to be anything over and above a condition for ownership over one's action in the sense relevant for moral responsibility.

It is, in a way, unfortunate that the "Deep Self" name is the one that has stuck, as it tends to evoke thoughts of a quite ambitious project to locate a central, fundamental, all-important seat of agency within the sea of an agent's mental states. The aims of a Deep Self theorist in practice are generally a good deal more modest, (although as I've highlighted, they vary quite a bit). But because of these confusions, it is difficult to say of any particular view, including the one I advance in the rest of this dissertation, whether or not it ought to count as a Deep Self view.

4. Attributional-Responsibility

Deep Self theorists have historically been attracted to a particular kind of notion of moral responsibility: attributional-responsibility. In "Two Faces of Responsibility" Gary Watson first proposed that different parties to the responsibility debate seemed to be implicitly committed to different ideas about what sorts of responses were justified on the basis of the requirements they proposed.[30] He identified two aspects of responsibility: attributability and accountability. The recent literature has expanded to count answerability as a potentially distinct face of responsibility. This chart very roughly explains the general differences in the concepts in a way that is meant to be broad and inclusive:

If an agent meets the responsibility requirements, then we can…

Attributability	judge, perceive, or otherwise react to her wrongdoing as expressing a personal fault
Answerability	demand justification of behavior that is *prima facie* wrong
Accountability	confront her on the basis of her wrongdoing, often with the aim of demanding recompense or sanctions of some form

There are few points of agreement in the literature about how we should understand the relationship between these notions of responsibility. Some see the three as competing accounts, some think they represent wholly distinct facets of responsibility, others call for their unification, and others see one or more facet as a necessary condition on another.[31]

Rather than debate these points at this level of abstraction, my strategy instead will be to develop what I hope will be an attractive substantive theory of attributional-responsibility. On my view, attributional-responsibility is its own full-fledged form of moral responsibility, and accountability-responsibility is its own entirely distinct form. So one may be accountability-responsible without being attributionally-responsible and

[30] Watson (1996).

[31] See Watson (1996), Fischer and Tognazzini (2011), Strabbing (2011), Shoemaker (2011, 2013, 2015), Smith (2012), Talbert (2012), King (2014), Wolf (2015), and Zheng (2016) for discussion.

vice versa. In this dissertation I gradually advance claims that narrow in on this particular picture, although the earlier parts of the dissertation are meant to be compatible with other views of the overall picture. In Chapters 2 and 3, I defend my own picture of the requirements for attributability, which is, in principle, compatible with other understandings of how these responsibility concepts work in relation to one another. This includes the view that attributability's main purpose is to serve as a necessary condition for accountability-responsibility. In Chapter 4, I advance an account of blame's content that lays the foundation for understanding attributional-responsibility as its own full-fledged form of responsibility. In Chapter 5, I argue for the adoption of accountability-responsibility as a wholly distinct form of responsibility. The structure is such that at any point, the reader may get off the boat while still being able to, in theory, accept the claims of the previous chapters.

In this chapter, I consider various Deep Self views as views of attributability in order to consider them as contenders to the view of attributability I develop. On a Deep Self view of attributional-responsibility, an action's being appropriately related to deep self mental states is what allows us to move from an evaluation of the moral quality of an action to an evaluation of the moral quality of the agent on its basis. The main goal, as I see it, is to find a filter that separates the sorts of acts that an agent cannot truthfully claim come from 'outside themselves' from the ones that stem from mere neurological noise, because this is the class of acts for which agents may be blamed on the basis of what their actions say about them as agents, making them appropriate target of aretaic assessments.

Even given an understanding of responsibility on which attributional-responsibility constitutes its own full-fledged form of responsibility, there are still additional conditions beyond the agential requirements put forth in Deep Self theories that must hold for an agent to be blameworthy. Further conditions, including the moral status of the act, and perhaps epistemic conditions on the agent[32], are needed in order to know *what* any given action says about the agent.

When an agent stands in the proper agential relationship to her action such that it opens her up in principle to being morally responsible for the act (or what it reveals about her), I will describe such an agent as being "attributionally-responsible" for her action. It is consistent with my understanding of the term that an agent may be attributionally-responsible for a morally neutral action. So when I speak of a Deep Self view giving sufficient conditions for attributional-responsibility, I do not mean to say these are meant to be sufficient conditions for an agent being blameworthy for any particular action.

5. B-Tradition versus H-Tradition

With this background in place, we are now in a position to examine the first central question that this dissertation aims to answer: what are the deep self mental states that should be required for an agent to count as attributionally-responsible, and how should we mediate disputes between alternative accounts?

In advancing my positive view in this dissertation, I will make the methodological assumption that being responsible is metaphysically prior to holding responsible. That is, in order to find out when it is appropriate to hold someone morally responsible, we first need to know whether or not the person meets the specifiable metaphysical conditions for actually *being* responsible. David Shoemaker calls this the B-Tradition, which stands in contrast to the H-Tradition, according to which holding responsible takes priority.[33]

Adherents of the H-Tradition tend to hold the view, which has been quite pervasive in recent years, that moral responsibility ought to be ana-

[32] I say more about just what I take these additional constraints to amount to in Chapter 4, §2.

[33] Shoemaker (2015b): 20.

lyzed in terms of the responses (usually conceived of as reactive attitudes) that are *fitting* in holding each other responsible. Fittingness is often taken to be a *sui generis* primitive relation. Defenders of the B-Tradition rely on intuitions about whether or not certain responses are appropriate as well, but take these only to provide a defeasible epistemic guide to the conditions of responsibility.[34] Adherents of the B-Tradition take seriously the need to additionally locate some further explanation as to *why* the particular conditions that an agent must meet in order for it to be appropriate to respond in certain ways must hold rather than some other conditions.

Providing an argument for the metaphysical priority of being responsible over holding responsible is outside the scope of this dissertation, but I do think it makes methodological sense to proceed as though the B-Tradition is true until we have exhausted its possibilities. Gideon Rosen provides a strongly worded defense of this method:

> The Fittingness view is a theory of last resort. We should adopt it only if we have tried and failed to analyze appropriateness or to assimilate it to a relation studied elsewhere under another name….A theory of responsibility aims to articulate the conditions under which blame is appropriate, and then to explain why those conditions are as they are. And the trouble is that the Fittingness View would furnish grounds for abject pessimism about this project.[35]

Defenders of H-Theory might nevertheless echo P.F. Strawson's influential decree that the project of finding metaphysical conditions of responsibility has already failed or is bound to fail.[36] I hope, however, to offer an

[34] See also McKenna (2012), which advances the view that neither being nor holding responsible is more fundamental than the other.

[35] Rosen (2015): 71.

[36] Strawson (1962).

attractive picture of the conditions for responsibility grounded in the metaphysics of agency that might be taken as evidence to the contrary.

That said, while my arguments for my view will in some places assume the B-Tradition, the view I advance is, with some modifications, itself compatible with the H-Tradition.

6. Internality's Role in Deep Self Views

How do Deep Self theorists mediate disputes between alternative accounts of the relevant deep self mental states? Although not every Deep Self theorist is explicit about how they answer this question, most argue that the tokens of only one particular type of mental state or another are invariably "internal."[37] A mental state is internal iff the agent is identified with the state in such a way that it cannot legitimately be taken to be a mere occurrence that does not belong to the agent since it is an "alien" force.[38] I follow Agnieszka Jaworska here in distinguishing internality in this ontological sense from subjective active identification that is based on whether the agent perceives aspects of her psychology as being her own. These senses are perhaps not wholly unrelated, however, as non-self-deceptive subjective identification might be able to provide us with defeasible evidence of internality. While there is a possible view on which the ontological category of internal mental states with which an agent can rightly be identified amounts to nothing more than the states with which the agent takes herself to be identified with, such a view would require an argument.

The concept of an internal state is usually given a gloss as the kind of state from which an agent cannot be alienated. Whether explicitly or implicitly, something like this idea seems to play some role in explaining the

[37] Internality may play a less central role in how these disputes are mediated when the subject is approached from the H-Tradition.

[38] Jaworska (2007): 531.

proposed authority of the particular kind of deep self mental states on every major Deep Self view.

On the endorsing view, as put forth by Harry Frankfurt, the relevant deep self mental states are higher-order volitions.[39] On this view, agents have the ability to influence their actions via the formation of second-order desires, which are desires about what the agent wants to desire to do. Second-order volitions are desires not just about which desires an agent endorses having but about which one of these desires the agent wants to actually act on at a given moment in time. So an agent's action is attributable iff it is caused by a desire to φ that meshes with the agent's further desire to act on the desire to φ. Frankfurt seems to understand the expression relationship that needs to hold between second-order volitions and first-order desires in semi-causal terms. Either the first-order desire is not by itself sufficient to motivate the agent and she needs the 'push' conferred to it from her second-order volition, or her second-order volition to act on a desire to φ accompanies a desire to φ that is already sufficient to motivate her to action, and so her endorsement amounts to over-determining or at least "okay-ing" the fact that she will be led to action by such a desire.

An agent's second-order volitions, for Frankfurt, have the authority to speak for the agent because they are the output of an endorsement process, the goal of which is to confer the status of internality on first-order desires. In forming a desire to act on one of her first-order desires, an agent identifies herself with her first-order desire because, for Frankfurt, the process of endorsement is the process of identification and a state is internal iff the agent identifies with it.[40]

[39] Frankfurt (1971).

[40] One much-discussed serious problem for Frankfurt is that it seems arbitrary that second-order desires, rather than say third or fourth-order desires have special agential authority. The way Frankfurt thinks of the role of internality in the theory is part of what generates the problem. If first-order states are granted the authority to speak for the

On the valuing view, the conception of agential architecture is quite different.[41] On this model, agential psychology is divided between valuing and mere desiring parts, and each has its own ability to motivate the agent. The relevant deep self mental states are evaluative, although importantly they do not consist merely in the pure cognitive judgment that some course of action is best, but rather in the agent's setting ends for herself. An agent is attributionally-responsible for her action iff what she does is controlled by her evaluative system, which prescribes the overall best course of action. She is not attributionally-responsible when what she does is controlled by mere desires that do not flow from what she truly values. Supporters of the valuing view argue that valuing states are invariably internal by explicitly pointing to the following evidence that no agent can truly be alienated from her values: when an agent comes to repudiate one of her values it is always from the perspective of some contrary value, and so the initial valuing state will fails to exist for her *as a value* for her. In this way, an agent's own values are guaranteed to always be internal since she cannot be alienated from them.

agent due to the fact that a second-order process can confer such authority, we might think that the second-order states involved in the process need to get *their* authority from a similar sort of even higher-order process. In order to solve this problem, Frankfurt later concluded that the sequence must terminate in some sort of special kind of state or process of identification, like a decision, that is invariably internal. See Frankfurt (1987, 1992). These views face a larger worry, however, in that guaranteeing internality through a special sort of identification process they fail to be reductionist naturalistic views. Since I take part of the motivation for identifying deep self mental states to be to give a reductionist story of the conditions for attributional-responsibility, in this dissertation I will largely draw on Frankfurt's earlier second-order volition view. (While I do not address it head on, many of my comments on the commitmental aspects of caring views apply to Frankfurt's even later view, on which identification amounts to passive commitment [Frankfurt (2006)]).

[41] See Watson (1975), Mitchell-Yellin (2014, 2015).

On the planning view, as proposed by Michael Bratman, the proposed deep self mental states are personal self-governing policies about how one will act in various circumstances.[42] In many cases these plans are set by what the agent values, but in cases of normative silence in which agents take there to be no distinct best course, they commit to personal policies that apply also to similar situations in the future. An agent's action is attributable to her iff it is in line with her plans, and in this way plans confer internality on motivational states that are instrumental or realizer desires of these plans. Bratman proposes that plans are invariably internal for agents like us because they partially constitute our diachronic agency. We as agents cannot be alienated from our plans, not just because we set them, but because they tie us together as coherent agents over time.

Proponents of the caring view offer a notably different picture on which "identification is, for the most part, a passive process, garnering its authority for self-determination from one's nexus of cares."[43] For Shoemaker (2003) these caring states are conceptual frames for clusters of emotional dispositions that respond to the whims and woes of one's cared for object. For example, if an agent cares about the Phillies, she will experience anxiety over a potential loss, joy at a win, and despair if they don't make it to the playoffs. Sripada goes one step further and proposes that cares are *sui generis* kinds of mental states with distinctive functionally specifiable motivational, commitmental, cognitive, and emotional profiles.[44] In addition to a suite of emotional responses, if an agent cares about X she will be intrinsically motivated to perform actions that promote the achievement of X, be disposed to form judgments that cast X in a favorable light, and will want to go on caring about X. An agent is attributionally-responsible for her action iff, during the operation of the action-directed

[42] See Bratman (2003).

[43] Shoemaker (2003).

[44] Sripada (2016).

psychological mechanisms that are involved in the etiology of the action, her care exerts motivational influences (of sufficient strength) in favor of acting as she does. Like plans on the planning view, cares are proposed to play a crucial role in constituting diachronic agency via their commitmental aspect, which in turn helps to explain the fact that an agent cannot be alienated from her cares. Since the motivational strength of cares stems from the very thing that constitutes the agent as an agent over time, she cannot be alienated from cares: they make up who she is as an agent.

One further option on offer for those who are worried that the other accounts fail to establish a necessary condition on moral responsibility is to hold an ecumenical, or disjunctive view. David Shoemaker currently defends a view on which an agent either has to act in accordance with her cares or with her values (or both) in order for her to be attributionally-responsible for her action.[45] If the fact that they are both thought to be invariably internal is what makes cares and values good candidates to act as deep self mental states, then the role of internality in an ecumenical view is clear.[46] In principle, any combination of deep self mental state types could be combined to form an ecumenical view just as long as the ways in which the states are taken to confer internality on first-order motivational states are compatible sorts of explanations and both ways of conferring internality on resultant actions can coexist.

7. Against Skepticism about Internality

I have shown how every major Deep Self view either explicitly or implicitly relies on a concept of internality to explain why the actions that appropriately mesh with deep self mental states have the authority to speak for agents that they do. But these views rely on support from intuitive under-

[45] Shoemaker (2015a, 2015b).

[46] Although Shoemaker's own theoretical orientation, especially his endorsement of the H-Tradition, on the face of it seems to de-emphasize the role of internality in Deep Self theory.

standings of internal states as basically those from which one cannot be alienated. But what would it mean to be alienated in the relevant sense invoked by each of these views?

If we understand internality in terms of alienation, and alienation in terms of the idea that one's mental state does not belong to oneself in some way without further explication, this feeds directly into a skeptical worry. The skeptical worry goes something like this: one reason we have for thinking that it is possible for one's desire to not belong to oneself in some sense is that people report experiencing a feeling that one's desire somehow does not belong to oneself. But, if such reports are all we have to go on, this may be quite shaky grounding for a theory of attributional-responsibility since we can easily provide error theories that explain why people report that their desires are alien.

Terence Penelhum expresses this line of criticism particularly forcefully. He says, regarding an agent's expression of the fact that his motivating desire does not truly belong to him, that it is a

> form of moral trickery…[that] involves an extension of the notion of non-identification with one's own desires and behavior from the level of harmless and even mildly illuminating metaphor to that of gross literal false-hood. To say harmlessly that one is governed by a desire that is not one's own is to utter a metaphor the literal translation of which is that one is governed by a desire that one does not want to be governed by. To say that the desire is not one's own and mean this literally is to say something obviously false: for the desire is operative and therefore exists, and is not someone else's. This obvious falsehood can be given the appearance of respectability with the aid of philosophical theories about the division of the soul's faculties; and it is a falsehood we are sometimes willing to swallow about others as well as

> about ourselves, as in the Gallic concept of the *crime passionel*. But we all know better.[47]

Arpaly and Schroeder's diagnosis is somewhat different:

> When people find that they have not been as rational, sane, prudent or moral as expected, they may experience…the cause of their misbehavior as an alien intrusion[48]….In a culture such as our own, glorifying decisiveness, self-control and 'follow-through,' and with a tendency to medicalize failures of such traits, many agents will instinctively reject evidence of themselves as straightforwardly akratic, as having simply chosen poorly when they knew better. Instead, in some (and, it seems, a growing number) of circumstances, they experience their failure as apparently incomprehensible, an ugly intrusion upon their lives, and the psychological cause of this failure seems an unpleasant intruder.[49]

However, as both quoted passages suggest, internality skeptics have a shared concern that such feelings of alienation are illusory, fabricated, or the result of self-deception. Skeptics tend to base their accusations on two factors. First, this way of speaking could easily be used as a fancy way to excuse. Second, it seems obvious that all of a person's desires are her own, since they don't belong to anyone else.

Frankfurt has a response to both of these worries, however.[50] He starts by noticing that it is not obvious, except in a fairly trivial sense, that all of

[47] Penelhum (1971): 670.

[48] See also Buss (2012). Buss takes it that when some people speak of alienation, what they really mean is that they act with a lack of a willing attitude. But, as she puts it, "just as autonomous agency is compatible with stupidity and thoughtlessness, so too it is compatible with ambivalence, regret, disappointment, frustration, and self-criticism."

[49] Arpaly and Schroeder (1999): 383.

[50] Frankfurt (1988): 61-62.

our desires belong to us since they do not belong to anyone else. We only attribute some of the events in the history of a person's body to that person in a strict sense; some of them are mere happenings, such as getting lurched forward on a bus or experiencing a bodily twitch. Just as it would be unfair to say that because such behavior is not attributable to anyone else, it must be attributable to the agent, so too is it unfair to make this inference in regards to desires. Of course it does not decisively prove that one may be alienated from one's own desire, but the evidence in favor of motivational alienation is not dissimilar to the kinds of evidence we have of bodily-alienation. Allowing that a person can disclaim certain motivations as external is only as much of an opportunity for moral evasion as allowing that a person can disclaim certain movements of her body as external is. And yet, we do not regularly take this as reason to be skeptical of bodily twitches.

The difference here could be explained by the fact that motivational externality may be fairly rare, such that some people almost never experience their motivating desires to do things like drink a beer or check their ovens as being external.[51] Noting this fact can help give us a good explanation for why many feel that speaking of external desires would be tantamount to making up an excuse for one's action. If those people who do not experience externality were to speak about any of their desires in such a way, they *would* be merely making an excuse for their behavior. The tendency to extrapolate from personal experience in this regard is very common, and bears similarity to public reaction to many psychological conditions before (and sometimes even after) they are validated by sci-

[51] This idea may not be particularly compatible with all Traditional Deep Self views, though, since failures to meet the high bars they set for self-governing action are seemingly commonplace. I return to this point again in Chapter 2, §5, once the positive view I advance is on the table. The view I advance involves quite minimal conditions for non-alienation and so is particularly well suited to give this response to the internality skeptic.

ence. Those with clinical depression are thought to be merely lazy, and non-verbal autistic people are thought to be simply being difficult by people who do not experience similar psychologies themselves. The ways in which these accusations err seems only explicable by realizing that they come about in part because the accusers extrapolate from what would be going on in their own psychologies if they were to display similar behavior. We should be wary to not let *this* kind of "intuition" color philosophical thinking about internality.

While I think this shifts the burden of proof onto internality skeptics, it would be helpful for Deep Self theorists to have a more substantive, less metaphorical account of internality. In the rest of this chapter, I aim to provide one.

8. Internality is Approval

Consider the following case:

> Exhausted Elsie: Elsie is extremely tired, but is desperately trying to stay up to continue an important conversation with her friend. Despite maximal effort to remain awake, she lays down because she knows she is about to fall asleep. Elsie is not culpable for the cause of her exhaustion, let's suppose—it has just gotten very late. Is Elsie attributionally-responsible for lying down to fall asleep, such that in principle it would be appropriate to blame her on the basis of traits displayed by her behavior?

According to every Deep Self view, Elsie is not attributionally-responsible. What Deep Self views have in common in the way they intuitively explain why an agent like Elsie is not responsible is that they show how even though an agent like Elsie might be *motivated* to lay down, she does not *approve* of doing so. This lack of approval, I contend, is what licenses us to say that her motivation acts counter to her, or 'alienates' her from her action. In Chapter 2 I give an account of what I take approving to consist in, but for now I just mean to point to an intuitive sense of approving: when considering the options for what to do at *t* there is at some level something that the agent likes about or finds worthwhile about the prospect of φ-ing at *t*.[52]

If this intuitive notion of approving to some extent of an action rather than merely being motivated to perform it is what makes the difference between cases in which we are willing to grant that an agent's action is caused by an process that bears the mark of internality and ones in which we are not, then we have located a common feature of any plausible candidate deep self mental state. Whether the deep self mental states are proposed to be endorsements, valuings, plans, or cares, or some disjunction of these, they succeed in guaranteeing agents' resultant actions will be internal by guaranteeing that the agent will approve of her action. If I am right about this, this means we can locate a common necessary condition for attributional-responsibility shared by each major Deep Self theory.

9. Approval as a Common Necessary Condition for Attributional-Responsibility

9.1 Approving is Necessary on the Endorsing View

It is perhaps easiest to see how approving of one's course of action is a necessary condition on attributional-responsibility on Frankfurt's endorsement view. Second-order volitions are meant to secure the fact that the agent is not only motivated to act in the way that she does but that she is personally invested in that particular course of action. This aspect of the endorsement view comes out particularly clearly in Frankfurt's discussion of the contrast between wontons and full-fledged agents who form second-order volitions.

[52] This should *not* be taken to imply that the agent necessarily takes it to be the best or even a good course of action, just that she likes it.

A wanton, in Frankfurt's sense, is someone who lacks second-order volitions and so fails to take an interest in her will whatsoever.[53] The wanton lets her strongest motivational states win out and move her to action irrespective of any opinion she might have about the matter. A wanton, according to Frankfurt, lacks the capacity for self-reflective concern and thus acts out of "mindless indifference to the enterprise of evaluating their own desires and motives."[54] Why, on Frankfurt's view, are we meant to think that the wanton's actions aren't attributable to her in the relevant sense? For the wanton,

> …it makes no difference to him whether his craving or his aversion gets the upper hand. He has no stake in the conflict between them and so…he can neither win nor lose the struggle in which he is engaged.[55]

This means that when an agent is attributionally-responsible for her action it is at least partially because she "has a stake" in the outcome of the conflict among the economy of her desires. Having a stake in the conflict between first-order desires competing to become an effective desire seems to amount to having an opinion on the outcome. In other words, the agent needs to approve of being motivated to act in the way that she does.

9.2 Approving is Necessary on the Valuing View

Approving is necessary for attributional-responsibility on the valuing view as well, although a mistaken picture of the contrast between valuing and desiring (in terms of interpretation of valuing deep self views) may make this idea seem somewhat obscure. There is a picture of human agency that pits what an agent wants to do against what she thinks would be best to do, conceiving of the two things as wholly separate. On this view it

[53] Frankfurt (1988): 16.

[54] Frankfurt (1988): 19.

[55] Frankfurt (1988): 89.

is nice when an agent is motivated to do what she thinks is best, but this is either accidental or caused by the agent bringing her motivations in line with what is best; it is not that there is any motivational force to her judgment that a certain course of action is best. Given this sort of picture, it would be hard to see how approving of one's course of action would be a necessary condition on attributional-responsibility on the valuing view. Valuing, however, is often held to have some more intimate connection with motivation. And once this is granted, it is easier to see the connection with approval.

This picture can be further specified in a number of different ways. For example, on one view put forth by David Lewis, valuing X consists in desiring to desire X. Value, for Lewis, just is what a person would be disposed to desire to desire in certain ideal circumstances.[56] If an agent acts in accordance with her values, and valuing is given Lewis's analysis, the connection to the agent's approving of her course of action is clear: the agent who values φ-ing has a stake in wanting to be moved to φ. It's interesting to note that Lewis seems to take the intuitive connection between valuing and approving to be strong enough to support an account where valuing essentially *just is* a certain kind of approving.

While adherents of the valuing view need not be Lewisians about valuing,[57] in order to make the view that valuing is connected up in the right sort of way with agency in the sense that could reasonably ground attributional-responsibility, they do posit some sort of strong connection between valuing and motivation via the fact that an agent approves of doing what she takes to be the best thing to do.

[56] Lewis (1989).

[57] Given that coupling the Lewisian account of valuing with the valuing Deep Self view would make the account of agency hierarchical, proponents of the view, like Gary Watson, who criticize the hierarchical nature of the endorsing view might even have special reason *not* to adopt it.

To act on one's valuing state in the sense that defenders of the valuing view conceive of it is never to merely act in accordance with what one coincidentally believes to be good. Rather, valuing is thought to have something to do with agency, and thus to have an essential connection to motivation. Gary Watson's characterization of valuing makes this connection clear:

> Now, to be sure, since to value is also to want, one's valuational and motivational systems must to a large extent overlap. If, in appropriate circumstances, one were never inclined to action by some alleged evaluation, the claim that that was indeed one's evaluation would be disconfirmed. Thus one's valuational system must have some (considerable) grip upon one's motivational system.

So the notion of evaluation here is in an important way personal; an agent's values issue from a faculty that has a "grip" on her motivations. If the thought "it's the right thing to do" is meant to have a grip on motivation, it must be because the second thought, "and I approve of doing the right thing," is also present in some form. Whether the second thought is a matter of the meaning of rightness, a truth about human nature, or a standing disposition that happens to be present in agents like us (or something else), the fact that the agent approves of acting as she does because it is right seems baked into the story.

Watson explains that the sort of motivational power exerted by valuing is special because we are concerned to bring about the satisfaction of desired ends for some reason that goes beyond the fact that acting alleviates the suffering of having the unsatisfied desire. For an agent to value φ-ing is for her not just to desire to φ but to set φ-ing as an end for herself. And so an agent must not only be motivated to φ, but also actually approve of φ-ing for some reason. As Watson puts it,

> Now, it must be admitted, any desire may provide the basis for reason insofar as non-satisfaction of the desire causes suffering and hinders the pursuit of ends of the agent. But it is important to notice that the reason generated in this way by a desire is a reason for *getting rid* of the desire, and one may get rid of a desire either by satisfying it or by eliminating it in some other manner (by tranquilizers, or cold showers). Hence this kind of reason differs importantly from the reasons based upon the evaluation of the activities or states of affairs in question. For, in the former case, attaining the object of desire is simply a means of eliminating discomfort or agitation, whereas in the latter case that attainment is the end itself. Normally, in the pursuit of the objects of our wants we are not attempting chiefly to relieve ourselves. We aim to satisfy, not just eliminate, desire. [58]

And so, on the valuing view, valuing is necessary for attributional-responsibility precisely because it guarantees that the agent's effective desire becomes effective because she approves of her course of action. And so, proponents of the valuing view, too, should hold that approving is a necessary condition on attributional-responsibility.

9.3 Approving is Necessary on the Planning View

According to the planning view, an agent is attributionally-responsible iff she acts in accordance with her policy about how to act in such a situation.[59] Her policy-setting may be governed by her values in many cases,

[58] Watson (1975): 210-211.

[59] In the article I draw from here, Bratman specifically brackets off questions of responsibility, focusing his discussion on identification alone: "…I want to see if we can, instead, describe without independent appeal to judgments of responsibility—a fairly unified phenomenon that is plausibly seen as the target of such talk of identification" (Bratman [1996]: 2) His picture might just as easily be considered as a contending deep self account of attributional-responsibility, however, and, with this caveat, I will proceed as though it were put forth as one.

and in those cases the same considerations I raised in the last section should lead to the conclusion that she must to some degree approve of her course of action. But in cases that are normatively underdetermined, she forms or acts on a previously determined policy that she just *decides* to treat as reason-giving. If agents in these cases just follow personal policies that are not governed by anything as strong as all-things-considered judgments about what would be best, it might be far from clear that agents who act in accordance with these policies need approve of their actions.

However, following Velleman, Bratman acknowledges the possibility of a case in which an agent forms a plan in such a detached way that the action she takes when she fails to act in accordance with it would still be attributable to her. This provides a reason to supplement the story about what must obtain in these sorts of cases for the agent to be attributionally-responsible. Bratman supplements his account by adding that the agent who φs must be satisfied with her decision to treat her desire to φ as reason-giving. If an agent meets this condition, it seems to me that she would have to approve of at least something about it.

Interestingly, for Bratman, the agent needs to be satisfied with the decision to treat the desire as reason-giving not just at the time of the decision but also at the time of action. This is evident in the following passage:

> In "The Importance of What We Care About," Frankfurt emphasizes that one can decide to care about something and yet "when the chips are down" fail to care about it. Perhaps, similarly, I might decide to treat my desire, say, to seek a reconciliation with an old acquaintance as reason-giving and yet, when the chips are down, find myself unable to treat it this way. I might find that, despite my decision, and despite the fact that I am satisfied with that decision, I do not care enough

> about reconciliation. In such a case it seems that I have not fully succeeded in identifying with my desire for reconciliation.[60]

This seems right, since the agent's approval at the time of action seems to intuitively be what matters.

Now, Bratman understands satisfaction with a policy not in terms of the presence of a particular attitude, but rather, in terms of the alignment and integration of the policy with the agent's other policies: "One is satisfied with such a decision when one's will is, in the relevant ways, not divided: the decision to treat as reason-giving does not conflict with other standing decisions and policies about which desires to treat as reason-giving." But notice that this analysis of satisfaction only makes sense as an analysis of satisfaction when we think of the sum of the agent's other policies as providing a guide to what the agent generally approves of doing. Again, here, agential approval of some form seems necessary on the view.

9.4 Approving is Necessary on the Caring View

Several aspects of the caring view might be thought to implicate approval. First, among "joy" and "elevation," "approval" is explicitly listed by Sripada as one of the positively valenced emotions that agents are disposed to experience when they act in ways that advance their cares.[61] If the emotional aspects of caring states are meant to take center stage on the view, it seems plausible that there must be some element of an approving emotion toward one's action in order for the agent to count as caring in the relevant sense.[62] There might be a sense of caring on which happiness

[60] Bratman (1999): 202.

[61] Sripada (2016): 8.

[62] Is approval an emotion? This may be more or less plausible depending on one's account of what emotions are. As I will explain in more detail in Chapter 2, I take approval, instead, to be a function of an agent's motivational profile.

towards acting in ways that advance the cared-for object without approval is sufficient for caring, as it seems possible in some sense to “care about something despite oneself.” For example, you might get a small twinge of pleasure from someone being called out on her bad grammar in an internet comment thread, once you’ve already renounced that practice as being classist. In one sense, it might seem appropriate to say that you still care about grammar, despite yourself. But it seems implausible that cares in this sense would be good candidates for deep self mental states. Even a wholly unwilling addict might get some pleasure from drinking, but a theory that gives the result that an unwilling addict is responsible since she cares about drinking alcohol seems counter to the aims of a Deep Self view.

In addition, the evaluative judgment aspect of the caring view further emphasizes the importance of the agent’s approval of her action. If an agent is attributionally-responsible for her action that furthers the end X, her action will be suitably related to the fact that she is disposed to form judgments that cast X in a favorable light. Recall again the earlier conclusion, regarding the valuing view, that the relevant sorts of evaluative judgments that can ground an agent’s attributional-responsibility must be related to her agency via the fact that she approves of those actions on the basis of her values. These same considerations apply to the evaluative aspect of the caring view as well. Sripada writes that the valuing view “places all elements of the relevant class of evaluative judgments within one’s deep self.” In contrast, “the care-based view allows that many evaluative judgments don’t bear any connection to the deep self, namely those that don’t bear the right dispositional tie to one’s cares.” In this way, the caring view is even more explicit about the fact that acting in accordance with evaluative judgments usually implicates internality because agents act on their judgments *because it matters to them* to act in accordance with what’s right. This sense of mattering, it seems to me, is fundamentally tied to approving.

Finally, the motivational and commitmental aspects of caring seem to implicate approval. For Sripada, if an agent cares about X she is intrinsically motivated to perform actions that promote the achievement of X. She also will want to continue caring about X. Together, this makes it the case that when φ-ing promotes the achievement of X, and X is something the agent cares about, the agent has an intrinsic desire with a positively valenced higher-order attitude towards being motivated by it. Although the consideration of the promotion of the achievement of X may not outweigh other factors in a given case, the agent still would seem to have to approve of being motivated to φ in at least a minimal or *pro tanto* sense for this to be the case.

It's important that the desire to promote the achievement of X is intrinsic, because this rules out cases where the agent might only be motivated to φ due to wanting to quell an external urge to promote the achievement of X, as I discussed in the section on the valuing view. If an agent acts on a desire to promote the achievement of X to quell such a desire it would seem strange to describe the agent as caring about X, and it would equally seem strange to describe the agent as approving of her action. Shoemaker makes a related point in discussing how his caring view differentiates non-attributable actions of unwilling addicts from attributable weak-willed actions:

> For the unwilling addict, it matters greatly that his desires for the drug are satisfied or eliminated—it may not matter how. For the merely weak willed, however, the primary object of one's care-dependent desire is not the mere satisfaction (or elimination) of some other non-care-derived desire.[63]

[63] Shoemaker (2003): 103.

Again here, the contrast seems best explained via the concept of approval. The weak-willed agent, unlike the unwilling addict, actually *approves* of something about acting as she does that goes above and beyond the alleviation of an urge.

And so, given each major Deep Self view, it is a necessary condition on attributional-responsibility that the agent approves to some degree of her action.

Chapter 2: A Sufficient Condition for Attributional-Responsibility[64]

1. Introduction

Chapter 1 surveyed some of the advantages of the Deep Self approach to attributional-responsibility. I showed that the appeal of each major kind of Deep Self view currently on offer lies in the fact that when an agent's action meshes with the proposed deep self mental states, her action is guaranteed to be caused by a process from which she is not alienated. I suggested that this lack of alienation amounts to the agent's approving of her action in some sense, and argued that proponents of each Deep Self view have reason to hold that approving of one's action to some degree is a necessary condition on attributional-responsibility.

In this chapter I argue that the notion of approval is not only necessary for attributional-responsibility, but also sufficient. First, I show that each major Deep Self view contains additional elements beyond securing the agent's approval of her action that are unnecessary for attributional-responsibility. None of the deep self mental states: valuing, caring, endorsing, or planning, as conceived of by proponents of traditional Deep Self theories, are necessary for attributional-responsibility. Some argue that attributional-responsibility might be secured by more than one mental state kind, and I carry this move to its logical conclusion. Instead of attempting to identify a kind of mental state or process that is invariably internal, I suggest that the way to proceed is to give an analysis of internality itself and hold that agents are attributionally-responsible for any instance of an action caused by an internal process. I analyze internality in

[64] This chapter greatly benefitted from comments on earlier drafts by Eugene Chislenko, Nathan Robert Howard, Renee Jorgensen Bolinger, as well as discussion by the participants of the fourth New Orleans Workshop on Agency and Responsibility, Temple University's Freedom and Responsibility Fall 2017 Seminar, and USC's Spring 2016 Dissertation Seminar.

terms of minimal approval, so an agent is attributionally-responsible iff her action is caused by a process the presence of which ensures that she minimally approves of acting on the motivational state that she in fact does.

To minimally approve of φ-ing, I suggest, is to have a hypothetical, partial desire to act on a motivation to φ for some further aim other than to get rid of the desire to φ. I motivate each aspect of this analysis in turn, and then turn to some advantages the view has over traditional Deep Self views. My view is comparatively agnostic regarding the processes involved in the production of action, and as such, it is better able to stand up to some of the most deeply entrenched problems for the Deep Self approach.

2. Unnecessary Components of Deep Self Views

As I argued in Chapter 1, each major Deep Self view is consistent with the idea that approving in some sense is a necessary condition on attributional-responsibility. However, proponents of each major Deep Self view take there to be additional elements to their account that are necessary for attributional-responsibility. This, I argue, is false. Each Deep Self view succeeds at locating a sufficient condition for attributional-responsibility just to the extent that it secures the fact that in each instance of intuitively attributable agency, the agent will approve to some degree of her action. Securing the fact that the agent approves of her action in some minimal sense is both necessary and sufficient for attributional-responsibility.

The first part of this argument will consist of showing that each major Deep Self view contains elements that are unnecessary for attributional-responsibility. In this section I show how the valuing, caring, and planning views contain unnecessary elements. I hold off on addressing the unnecessary aspects of the endorsing view until §4 and §5 of this chapter, as this part of this argument serves to set up my positive view.

2.1 Valuing is Not Necessary for Attributional-Responsibility

If we think of the valuing view as being supported primarily by the idea that the mark of agency comes from an agent's acting on her consideration of what she takes to be *best* to do, it is fairly easy to generate counterexamples. There are many actions for which an agent is attributionally-responsible but does not judge her course of action to be all-things-considered best. For example, in cases of normative silencing, an agent cannot form an all-things-considered judgment about what it would be best to do since there is no best option. Consider instead the view on which the agent must consider the action to be *among* her best options. Even on this view, there are counterexamples in which an agent may act in a way that she is attributionally-responsible for where she does not take her action even to be *one of* the all-things-considered best things to do but nevertheless embraces it on some level. Gary Watson, the original proponent of the valuing view, now endorses this objection and has come to see his earlier view as being too rationalistic. As he puts it,

> When it comes right down to it, I might fully 'embrace' a course of action I do not judge best; it may not be thought best, but is fun, or thrilling; one loves doing it, and it's too bad it's not also the best thing to do, but one goes for it without compunction.[65]

Cases in which people act out of love that they themselves take to be wholly irrational seem to be commonplace as well as particularly compelling examples. People who are in love wholeheartedly act in ways that are not among the options they consider all-things-considered best, nor do they sometimes even consider them good. On a fairly regular basis people return knowingly to undeserving exes, begrudgingly do unreasonable favors for loved ones, and support family members in ways that go beyond

[65] Watson (2004): 168.

the realm of reason. And they seem to do so with enough agency that we rarely think of these agents as being excused from responsibility for their actions as a result of their reason being undermined.

But, as David Shoemaker notes, despite their intuitive appeal, these cases are liable to be recast in terms of value by proponents of valuing views.[66] For example, Angela Smith argues that in these sorts of cases, agents merely speak in terms of 'having no reasons' to do as they do as a *façon de parler*.[67] When it comes down to it, agents really *do* judge what they are doing out of love to be best given their reasons, such as reasons of shared history, being entangled with someone in such a way, or the like. When an agent has the explicit thought that she ought not go back to her ex but does so anyway, she uses 'ought' only in the inverted commas sense. The thought she has, valuing view adherents allege, must really be something like *everyone would counsel me not to do this, but still I find it most valuable.*

Shoemaker appeals to cases of well-planned revenge as another example of actions an agent may be attributionally-responsible for despite undertaking them independently of what she values. During the plotting of revenge an agent may come to realize that her doing so is morally wrong and prudentially disastrous, and even that there is nothing valuable whatsoever in so-acting yet nevertheless carry out her plot. And yet, agents are paradigmatically attributionally-responsible for acting on such desires for revenge. While these seem like perfect cases for making the point that valuing is not necessary for attributional-responsibility, valuing theorists could argue that such agents are impossible. In the ordinary case, agents who act like this, contrary to appearances, *do* value retributivism either in the abstract or in the particular case, and they are either wholly self-deceived about their anti-retributivist values or else the values are merely

[66] Shoemaker (2015a): 129-134.

[67] Smith (2012).

aspirational: such agents only *desire to be such* that they value anti-retributivism. Otherwise, perhaps the agents are suffering from sort of condition that should undermine their being held attributionally-responsible.

It seems to me, frankly, implausible that every single case of acting out of love and every single case of acting out of revenge are realistically re-cast in such a way. One worry is that the intuitions that these cases are well described in terms of value are theory-laden and influenced by independent considerations that favor motivational judgment internalism, the view that there is a necessary connection between moral judgment and motivation. Motivational judgment internalism is only plausible when it includes the caveat that agents are motivated by their values except in cases of volitional disorder. The concern is that if finding motivational judgment internalism plausible influences intuitions about the cases in which an agent acts out of love or revenge, some sense of what differentiates volitional disorder from normal cases might already be smuggled in to the picture, when that is the very thing that is meant to be in question.

Part of the force of these revenge and love cases, and our resistance to recast all of them in terms of values, is that they are realistic, common-place, and seem to differ phenomenologically from the way they are described by proponents of the valuing view.[68] Perhaps the best case against

[68] Mere hypothetical cases are more fairly subject to be recast. To take a case that has been offered in support of the view that valuing is not necessary for attributional-responsibility that I do not think will be successful in convincing any valuing theorists, Shoemaker gives a case of an artist and philosopher who cares only about being the type of person who lives beyond the realm of justification such that when he judges something to be best for him he instead acts on a perverse desire to do the very opposite. When he acts in this way, according to Shoemaker, he clearly doesn't think his action is all-things-considered best, but he certainly seems to be attributionally-responsible for what he does. In order to make the case resistant to the recasting problem, Shoemaker specifies explicitly that this agent doesn't desire to live beyond the realm of justification because he values it, but I think this is unlikely to convince any valuing view proponents.

the valuing view is a real life example Shoemaker provides in which his own volitionally necessitated will came apart entirely from what he judged to be all-things-considered best so persuasive. He describes a time in which he encountered an injured mouse in his apartment, and knew that the best thing to do would be to take a hammer and kill the mouse instantly, saving the creature from a more protracted and painful death that would be caused by simply letting it out into the wild. But as he stood over the mouse with his hammer, he found he couldn't bring himself to do it, and chose to let the mouse free instead. This is a case on which he judged it clearly best to take one course of action but couldn't bring himself to act accordingly in such a way that nevertheless revealed something about what he is like as an agent. It seems very implausible to say that he was self-deceived about his view that putting the mouse out of its misery would be best; he is surely a very reasonable guy when considering what one ought to do. It makes most sense to suppose that he was instead motivated by some combination of squeamishness and contorted mercy, neither of which operated via encouraging him to think it was among his all-things-considered best options to let the mouse go free. And yet, despite the fact that he acted completely contrary to reason, he certainly seems attributionally-responsible for his action. And so it does not seem that con-

Valuing view proponents will describe such a man as engaging in paradoxical decision making. When he performs one of these contrary-to-judgment actions, the valuing theorist will argue, it must surely be that he takes himself to have justificatory reason to do so; he can even provide the reason: he does it *because* it is the opposite of what he judges best! In coming to this realization, the undertaking of this opposite action becomes what the agent judges it best to do, but now should he do the opposite of that? Since this kind of case is not common, there is no reason to doubt the claims of the valuing view adherent that this sort of case is paradoxical.

sidering a course of action to be among one's best options is necessary for attributional-responsibility.[69]

2.2 Neither Planning nor Caring is Necessary for Attributional-Responsibility

I reject caring and planning accounts of deep self mental states as being necessary for attributional-responsibility for reasons that apply to both, and so in this section I treat them in tandem.

Both caring and planning views posit that in order for agents to be attributionally-responsible they must act from a motivation that is in some way importantly tied to their past and/or future motivational architectures. As I discussed in Chapter 1, this aspect of those views is usually touted as an advantage. If agents are constituted over time by plans or cares, this offers a neat explanation as to why compulsive behavior that does not mesh with one's plans or cares is alienating: it literally stems from outside of the agent. While diachronic coherence in one's action arguably is a marker of agency *par excellence*, it is much less clear that acting in accordance with this kind of diachronic scaffolding is necessary for attributional-responsibility.

David Shoemaker, arguing against the necessity of Bratman's planning states, gives an example of a racist who prudently decides "to never let his

[69] Silverstein (2017) gives a compelling case that practical reasoning, understood as reasoning about what to do and normative reasoning, understood as reasoning about what one ought to do, are non-identical. An agent can settle the question of whether or not she should dine out for lunch while deciding to forestall reasoning about whether or not she will in fact do so until tomorrow. In this case normative reasoning has finished long before practical reasoning has even begun. While this by itself does not foreclose the possibility that attributable agents must act in accordance with reason, it does drive a further wedge into the supposed tight connection between valuing and ordinary attributable action.

hatred structure his will."[70] When he witnesses a KKK event though, one day, he says, "woo hoo!" In this case the man pretty clearly has not decided to treat his racist desires as giving him a reason for action, but nevertheless appears to be attributionally-responsible. Shoemaker explains that this man is attributionally-responsible because, despite the fact that he judged it best not to let his racism show, he *is* a racist; he has a longstanding hatred for Black people and these attitudes reveal what he is generally like as a person.

But consider a modification to the case that seems to tell against the caring view as well. Suppose the man who lets out a "woo hoo!" upon seeing a KKK event has never had a prior racist thought or attitude and never has one again. He is merely caught up in the racist fervor of the KKK demonstration and is temporarily overcome. Nevertheless, the man in this situation seems just as attributionally-responsible for his reaction. This indicates that it is the man's approval of his action in the moment, rather than across time, that matters for attributional-responsibility. Supposing we were confidently assured of the fact that this was and would always be his sole racist action, this might have some effect on the sorts of attitudes it would appropriate to direct towards him compared to the thoroughgoing active racist, but it does not affect the fact that his current act is attributable to him.

It's possible that mental states like caring or planning may help illuminate aspects of an agent's personality that are elusive when considering only time-slice properties, but I think we should be skeptical of any claims that such understanding is required for attributional-responsibility and think to do so would be to conflate two different senses of the Deep Self.[71] Agents who are swept up in moments of anger, passion, or fleeting interest can still act in ways that they stand behind; sometimes they even do so

[70] Shoemaker (2005a): 121.

[71] See Chapter 1, §3.1.

wholeheartedly. Thus, there is reason to think that even if the planning and caring views point to mental states that *guarantee* that an agent approves of her behavior in the right sense for attributional-responsibility, acting in accordance with these mental states cannot be *necessary* for attributional-responsibility.

Another, albeit less conclusive, reason to be suspicious that caring or planning states are necessary for attributional-responsibility is that they come with ontological baggage. Both Bratman's planning view and Sripada's caring view involve countenancing the existence of new kinds of mental states that are said to play a crucial role in agential architecture. Considerations of parsimony put the burden on these theorists to prove that adding new mental state kinds to our ontology is absolutely necessary. This concern is amplified by the fact that, at least outside of the domain of thinking about responsibility, it is intuitively plausible that plans and cares might be constructed out of beliefs and standing desires with particular sorts of content.[72] Parsimony considerations are always weighed against the explanatory power of views, and both Bratman and Sripada offer arguments elsewhere about the global roles of caring/planning states on agent architecture. These comprehensive pictures of agency could each be given dissertation-level treatments, and so I will certainly not attempt to discredit them here. I merely mean to mark the fact that, without ante-

[72] It may be worth considering the viability of a planning or caring view of attributional-responsibility coupled with the view that planning or caring states are in fact constructed out of simpler component parts. Shoemaker (2003) in some ways presents an example of such a view, and so is not subject to the same criticism I pose in the section for Sripada's and Bratman's views. In thinking about the viability of demarcating deep self mental states via their content rather than their kind, it may be worth considering the view put forth in (Velleman [1992]), on which agents effective motivation must be suitably related to a standing desire with a particular content: the desire to act in accordance with reasons. Depending on how one construes Frankfurt's second-order volitions, his endorsement account may also be construed in this way.

cedently agreeing with the arguments that these theoretical posits are necessary to explain other aspects of human agency, it is more difficult to get on board with the planning or caring view than it would be to accept a view that makes do with mental states that are more or less universally agreed to exist. There is reason to wonder whether we really need to posit a special *kind* of mental state to do the work of supplying a criterion for attributional-responsibility.

3. From Type-Disjunctive to Token-Disjunctive Views

Let's take stock. I have now argued that each traditional Deep Self view provides a deep self mental state that is sufficient but not necessary for attributional-responsibility. Each traditional Deep Self view succeeds insofar as it articulates a way that an agent may approve of her action such that she is not alienated from it. But the additional elements of each view that go above and beyond the fact that the agent approves of her action are superfluous; as long as the agent approves of her action and this is related to why she acts, it doesn't seem to matter much how she comes to do so.

One way of proceeding in light of this would be to hold a disjunctive view on which an attributionally-responsible agent's action may be caused by any of the several types of mental state that implicate that the she will approve of her resultant action. As mentioned in Chapter 1, §6, David Shoemaker's ecumenical view is a disjunctive view that is also motivated by the thought that there is no single type of deep self mental state that is necessary for attributional-responsibility. On the ecumenical view, an agent is attributionally-responsible for her action iff her effective motivation aligns either with her cares or her values. If endorsing and planning views also provide sufficient conditions for attributional-responsibility, one way to capture this would be to hold a disjunctive view on which an agent is attributionally-responsible for her action iff her effective motivation properly aligns with her cares or her values or her endorsement or her plans.

Recall, though, the distinction from Chapter 1 between the B-Tradition and the H-Tradition. On the H-Tradition, holding responsible is metaphysically prior to being responsible and to be analyzed in terms of the responses (usually conceived of as reactive attitudes) that are *fitting* in holding one another responsible. Since Shoemaker follows the H-Tradition, his ecumenical view is allegedly supported by the fact that agents' actions that mesh with either cares or evaluative commitments both seem to make fitting certain paradigm aretaic sentiments in response such as admiration and disdain. In Chapter 1, §5, I expressed my commitment to the B-Tradition, according to which being responsible takes priority over holding responsible. So, while the responsibility responses that appear to be fitting may provide a defeasible epistemic guide to the conditions of responsibility, I have argued that we also need to locate some further reason why those particular criteria must hold rather than some other criteria.

Given a methodological commitment to find some deeper reason related to the structure of attributable agency in virtue of which the proposed conditions for attributional-responsibility are in fact the correct conditions, the disjunctive view occupies a somewhat precarious position. Given the B-Tradition, any disjunctive view invites the further question: in virtue of what should we count the kinds of mental states the disjunctivist posits, and *only* the kinds of mental states she posits as constituting sources of the deep self? Suppose she answers this question by identifying some property of these kinds of mental states, *F*, which justifies their inclusion as states that can imbue first-order motivational states with the authority to speak for the agent. If the fact that states of that type have *F* is the reason they can speak for the agent, the disjunctivist deep self theorist should admit that a state of any kind, just so long as it has *F*, can imbue an agent's first-order motivational states with the authority to speak for her.

To maintain her position in light of this challenge, the disjunctivist would have an additional argumentative burden not shared by traditional Deep Self theorists. While it is open to traditional Deep Self theorists to claim that there is something uniquely special about endorsing, valuing, planning, or caring that makes that kind of state the only kind that can imbue first-order motivational states with the authority to speak for the agent, the disjunctive Deep Self theorist would have to admit that there is some property shared by each kind of Deep Self state that has that authority and then argue furthermore argue that no other state could in principle have that property.

To preserve the advantages of the disjunctive view without having to meet that challenge, one could instead hold a disjunctive view of the following form:

> **Type-Disjunctive Attributional Responsibility:** An agent is attributionally-responsible for φ-ing iff she acts on a motivational state that meshes in the relevant way with a further mental state type that has some property, *F*, by virtue of its being a token of that type.

But if *F* is a feature most fundamentally not of mental state types but of all the token mental states that fall under a type (for example, if *F* is a feature of every caring state), then this invites a further question: what is the relevance of mental state *types* to the theory? If the *type* of mental state plays no role in the theory except that of being a good predictor of whether the token will have *F*, then not including other mental state tokens that have *F* despite not being of a type that guarantees that all its members have *F* would be arbitrary.

A better alternative is to adopt what I'll call token-disjunctivism. According to token-disjunctivism, an agent is attributionally-responsible for any action caused by a motivational state that properly meshes with any further mental state or collection of states that have *F*. These mental states may be tokens of types whose tokens invariably have *F*, but the types of which they are tokens need not have only tokens that have *F*.

> **Token-Disjunctive Attributional-Responsibility:** An agent is attributionally-responsible for φ-ing iff she acts on a motivational state that aligns in the relevant way with a further mental state token (or group of mental state tokens together) that has some property, *F*.

What should we fill in for *F*? Since, as I have argued, each Deep Self view seems to provide a successful criterion for attributional-responsibility only to the degree that it ensures that the agent approves of her action, and that no additional elements are necessary, I suggest that for *F* we fill in "the fact that the agent approves to some degree of her action." I am proposing that the fact that an agent's action is brought about by a process that is non-coincidentally related to the fact that she approves of her action is both necessary and sufficient for attributional-responsibility.

In what follows I give an account of what it means for an agent to minimally approve of her action, and argue that a token-disjunctive account of attributional-responsibility based on minimal approval has a number of advantages over traditional Deep Self views.[73]

4. Partial Identification

I am going to build up an account of what it takes for an agent to approve of her action in the sense that I believe should ground an account of attributional-responsibility. Just as on Harry Frankfurt's endorsing view, I think the agential criterion for attributional-responsibility has to do with

[73] I argued in Chapter 1 that the "Deep Self" label is ill-defined, and that the positive view I advance may or may not count as a Deep Self view depending on the criteria that are prioritized in classification. However, I will adopt the convention of referring to my view as a "Deep Self view". That said, I don't put much stock in whether or not the Deep Self label rightly fits the view I put forth, and I only label it as such to bring out the fact that it clearly shares some of the advantages, both dialectical and substantive, of traditional Deep Self views.

the structure of one’s will and is built out of desires. Although the view I will put forward is quite different from Frankfurt’s, I arrive at it by noting the ways in which it contrasts with views like Frankfurt’s.

The first way in which I think an account of attributional-responsibility needs to diverge from Frankfurt’s view is that it ought to locate the criterion for partial identification rather than complete identification. In Chapter 1 I hinted at the view that identification with a course of action need only be partial in order for the agent to be attributionally responsible for it, since it seems that agents need only approve of their course of action in some way, or to some degree. This idea stands in contrast to most traditional Deep Self views. For example, recall that on Frankfurt’s view, in order to be attributionally-responsible for φ-ing, an agent must reflect on her first order desires and come to desire most to act on her desire to φ. Crucial to this view seems to be something like the idea that an agent must completely stand in favor of coming to act as she does. But views like this seem unable to respect the intuitive judgment that agents can be attributionally-responsible for actions that are the result of weakness of will.[74] In Chapter 3 I will argue that the view I put forth in this chapter is the most promising way for deep self theorists to handle weakness of will cases. For now I just want to show how weakness of will cases intuitively move us away from the view that complete identification is necessary for attributional-responsibility to the view that merely partial identification is sufficient.

[74] This objection has been raised, in various forms, by Vihvelin (1994), Haji (1998), Haji (2002), Fischer (2010), Fischer (2012), McKenna (2011), McKenna and van Schoelandt (2015), and Strabbing (2016). A complication that should be noted here is that not all Deep Self views are always presented as views of the conditions for responsibility, and some theorists propose these sorts of conditions merely as views of self-governance, intentional action, autonomy, or agency *par excellence.* This criticism should be taken to apply only to the (many) Deep Self theorists who do take acting in accordance with one’s Deep Self to be a criterion for responsibility.

Consider the following case:

> Murderous Max: Max strongly desires to go out on a killing spree this morning because he hates people and there is almost nothing he likes more in the world than shooting them—in fact he thinks of himself as having made an art of it, perfecting his technique more and more with each kill. There is one thing, however, that he cares about even more: he is extremely committed to his morning workout routine. As much as he wants to go out on a killing spree, he also realizes that if he does that, he'll have to forego his morning workout ritual. He knows that if he misses even one morning of working out, he'll probably fall off of his routine, and he'll thus sacrifice the progress he hopes to be making. So, after considering and giving some post-reflective weight to his options to act on both desires, he decides that acting on his desire to work out is what he most wants to do, it aligns best with his values, and is consistent with the plans he has set for himself. However, his desire to hone his murderous craft by going on that killing spree ends up just being so intense that he caves from lack of willpower and goes out and does the deed.

Max's killing spree certainly seems to reveal something about what he is like such that his action speaks for him for the purposes of attributional-responsibility; he is clearly blameworthy despite the fact that he endorses, in the senses relevant for most Deep Self accounts, going to the gym at the time of action. On Frankfurt's view, Max forms a second-order volition to go to the gym but some other first-order desire becomes his effective motivation. Max does seem to identify with his desire to go on a killing spree in a way that seems relevant. But if Frankfurt's account were the right account of identification, we would have to say that he does not and, as a result, is not attributionally-responsible for his killing spree. But this seems clearly to be the wrong result.

Traditional Deep Self views suffer from this problem because they are specially designed to show how acting on desires that agents themselves do not see as most favorable undermine agency in such a way as to make such actions non-attributable, as in the case of compulsion. But weakness of will is often described as the failure to act in accordance with what the agent finds to be the most favorable course of action, and yet intuitively we think weak-willed actions are attributable. These views tend to lack the resources to differentiate compulsion from weakness, so in exempting compulsive action they overextend to exempt weak-willed action.

The lesson we should draw from this is that while it makes sense to say that in compulsive cases agents are wholly alienated from their actions, it seems that there are cases, like the case of Murderous Max, in which an attributable agent doesn't stand behind his action as being *the* thing to do and yet still endorses it, in at least a partial way. While most Deep Self views give accounts of what would be required for an agent to *fully* stand behind her action, an agent does not need to wholeheartedly endorse her course of action in order to be responsible for it. If lack of identification with one's action is to be a relevant consideration in exempting an agent from attributional-responsibility, we'll need to understand this in terms of total lack of identification with one's action, rather than less-than-complete identification with one's action.[75]

In order to see the difference between what I'll call complete and merely partial endorsement, it will be easiest first to look at a case in which an agent feels herself being pulled in more than one direction by her first-

[75] This is not to deny the possibility that degree of identification with one's action may play some further role in determining features of blameworthiness. It seems at least *prima facie* intuitive that wholehearted embrace of morally wrong action could be cause for extra/more intense blame or blame of a special kind of character, and my proposal should not be taken as being incompatible with this. I am merely suggesting that we shift from a conception in which anything less than wholehearted identification falls short of attributability to one in which merely partial identification makes the cut, so to speak.

order desires and then explicitly deliberates about what to do at the second-order. Consider the following case:

> Three Desire Theresa: Theresa is currently at work and has three first-order desires, each with the potential to pull her in a different direction: she wants to complete the assignment she has been given by her boss, she wants to spend her time writing some thank you emails to some relatives who are waiting to hear from her, and she wants to slap her boss. She then considers each of these desires and considers which of them she most wants to act on. It's not that she is uncertain about what she ought to do: she knows she should complete her assignment; but she nevertheless is unsure prior to deliberating about which of her desires she most wants to act on. Upon reflection she gives some weight to the possibility of acting on her desire to get her work assignment done and some weight to acting on her desire to send her thank you emails. There's a part of her that wants to act on her inclination to do what she's supposed to do, and there's a part of her that would really prefer to act on her inclination to make sure her relatives hear from her today. Acting on her urge to slap her boss, she realizes, isn't even a contender for what she should do right now; even though she is struck by the raw urge to do it, she would never *really* want to do this. She's not even angry with her; it's just an occasional urge that pops into her mind. She decides, in the end, to work on her assignment.

Three possible outcomes in this scenario lead to three different levels of potential alienation. First, there is the outcome in which what she actually does is her work assignment. She feels no sense of alienation in this case, given that she acts on the desire of hers that she most wants to act on; she stands behind her action.

Next, there is the outcome in which, when she opens up her work assignment, she shifts over to her email and instead starts writing those thank you notes. It's possible this may feel somewhat alienating to her, as she is not, in the end, motivated by the desire that she decided she most wanted to act on. However, acting on a desire to write those thank you emails is in line with at least something that she wanted for herself. She may have decided in the end that what she wanted for herself *more* was to act on her desire to work on her assignment, but nevertheless she did want to some degree to act on her desire to write the thank-you emails. Furthermore, let's make the reasonable stipulation that it was no mere coincidence that she gave some weight to her desire to write thank you notes and the fact that she actually wrote thank you notes. The elements of her psychology that led her to give some post-reflective weight to her desire to act on her desire to write thank you notes were also involved in causing her to actually write the thank you notes. This act, I want to argue, is therefore not wholly alien to her since she at least partially endorses it.

Finally, there is the outcome in which she, despite never seriously considering it as a contender for the motivation she should act on, slaps her boss. If this happens, something seems to have gone seriously awry. Even though she had the first-order motivation to slap her boss, she experiences complete alienation since she acts on a motivation wholly outside of what was even in contention for what she wanted for herself to do upon reflection. In this case she is moved to action by a bit of her psychology that stands entirely outside of the complex psychological dispositions that affect choices about which motivations she would want herself to act on. Being moved to action when one would fail to even partially endorse its motive is, I contend, what makes actions that are compulsive in this way stand outside of one's agency, and thus fail to be able to speak for the agent. So, in order for an action to be able to speak for an agent it must be suitably related to the motivations the agent would at least give some second-order weight to in deciding on which motivation to act.

4.1 Partial Identification and Desire Individuation

While I have still not yet provided a complete analysis of minimal approval, since its foundation is the notion of partially endorsing a first order desire, and this partial endorsement is a special kind of desire, it is worth pausing to examine the notion of 'desire' at play here. I have argued that Theresa is attributionally-responsible for sending thank you emails because she gives some weight to acting on her desire to do so. In other words, she has a higher-order desire of some strength to act on her desire to send the emails.

Given certain conceptions of desire, though, one might think that whenever it is the case that Theresa wants to send the emails it is already the case that she has a desire to act on her desire to send the emails. For example, if 'desire' is given a simplistic dispositional analysis, Theresa's desire to send the emails just amounts to a disposition to send the emails. Since the only way to send the emails is by acting on her desire to send the emails, a desire to act on a desire to send the emails would seem to amount to nothing more than a disposition to send the emails. Thus, the fact that she gives some weight to her first-order desire can be nothing over and above her first-order desire itself.

However, there is reason to reject the simple dispositional analysis as inadequate. As I argue in "Depression's Threat to Self-Governance," in order to make sense of certain aspects of the phenomenology of agency, desires need to be individuated at least in part by their propositional contents.[76] The case of melancholic depression gives us reason to think that desires to φ and desires to act on desires to φ have distinct existences. This insight comes from thinking about cases in which an agent with melancholic depression wants to act on a desire to get out of bed but lacks the corresponding desire to get out of bed. The fact that getting out of bed is the satisfaction condition for the depressed agent's second-order en-

[76] Gorman (Unpublished Manuscript).

dorsement is no guarantee that she already possesses a first-order desire to get out of bed. Likewise, the fact that acting on a desire to send emails is the satisfaction condition of Theresa's first-order desire to send emails is no guarantee that she already possesses a second-order desire of some strength to act on her desire to send emails.[77]

That said, aside from this stipulation, I will not aim to mediate the dispute over the nature of desire here. In fact, my usage of 'desire' bears more similarity to the tradition of using the term 'desire' to refer to any motivational state whatsoever. When I use the term 'desire' in an example to refer to the each of the states that together make up a particular agent's partial endorsement of a course of action, the reader may substitute in her favorite motivational state for the term 'desire,' each time it is used, just so long as it is the kind of state that is individuated in part by its propositional content.[78]

5. Hypothetical Versus Explicit Endorsement

Besides requiring only partial rather than complete identification, I also think the kind of endorsement required for the type of minimal approval that matters to attributional-responsibility differs in another respect from

[77] Other views of desire may also be difficult to hold in conjunction with the view that partial endorsement matters for attributional-responsibility, even if they are not outright ruled out like the simplistic dispositional analysis. For example, consider the view that a desire to φ consists in nothing more than a judgment that one has some *pro tanto* reason to φ, coupled with an account of reasons on which *pro tanto* reasons are very easy to come by. If I take myself to have a *pro tanto* reason to act on a desire to eat my car, say because it has nutrients in it, such a view would describe me as having a weak desire to act on my desire to eat my car. If I then acted on an urge to eat my car, it is hard to see how the fact that I very weakly desired to act on my desire would make me attributionally-responsible for attempting to eat it. So views on which desires are this cheap to come by will not be particularly good candidates to pair with a partial-endorsement based view.

[78] Including but not limited to: intentions and evaluative judgments with motivational components.

the Frankfurt's traditional endorsement view. The traditional endorsement view posits that it is the actual *act* of endorsement that makes the agent's resultant action belong to her; the act of taking a stand results in ownership of one's motivation and its results.[79] However, not every case of attributable action involves the amount of intra-psychological reflection of Theresa or Max. In fact, we rarely go through the explicit process of taking a stand on which of our desires to act on in the course of deciding how to act. Yet, we are still able to say with some degree of confidence whether or not our actions align with the motivations we would have wanted for ourselves to act on if we *did* consider which motivations to act on. And this is what seems to make the difference between whether or not we identify at all with our actions. It is the fact that what the agent to some degree would endorse and what she in fact does align with one another, I propose, rather than any actual act of endorsement, that makes it the case that an action is suitably related to an agent's deep self such that it can speak for her. Since what's important here are the psychological dispositions and not the endorsement itself, the actual act of endorsement can be merely hypothetical. So the view, so far, is this. An agent's action can speak for her iff her act is suitably related to the fact that she would at least partially endorse (give some weight to) the desire that she acts on, if she were to consider which of her motivations she wanted to act on.

We frequently do not explicitly deliberate about or give additional weight to our competing first-order motivations *en route* to acting, and so views that require explicit endorsement predict that we will regularly fail to identify with the springs of our actions. However, our actions are almost always caused by the kinds of motivations to which we would give at least *some* weight if we were to deliberate, and in the rare cases in which a person's action fails to be produced by such a motivation, we generally

[79] This is crucial to David Velleman's understanding of Frankfurt's view. See Velleman (1992).

attribute the result to some sort of dysfunction. In moving from explicit endorsement to hypothetical endorsement, the view entails that relatively few actions fail to meet the requirements of attributability. I think this constitutes a point in favor of the hypothetical view for a couple of reasons.

First, it can help explain why some objectors to traditional Deep Self views are tempted to say that people who claim to be alienated from their actions must be making excuses. Recall that objectors to traditional Deep Self views sometimes object that they can too easily imagine situations in which they themselves might not go through the proposed attributability-granting processes and even, perhaps in hindsight, feel *some* sense of alienation from their actions, while they nevertheless maintain strong intuitions that they are attributionally-responsible for those actions. But a hypothetical endorsement view explains why complete alienation of the kind that undermines attributability is rare enough that it is entirely possible that these objectors have almost never met the conditions for it.[80]

Another benefit is that it can preserve the connection between felt alienation and actual alienation. While alienation in the sense that is relevant to attributability is not just a feeling, it is frequently accompanied by a feeling. One (perhaps non-essential) common feature of felt alienation is that it involves a sense that the process by which one comes to act is non-typical. The hypothetical partial endorsement view, since it holds that there are so few actions that fail to be attributable to agents, is compatible with the fact that our interpersonal relationships function in ways that assume that our actions generally issue from us in a way that speaks for us. If having many of our actions fail to speak for us was very common, the feelings that accompany any particular action failing to speak for us might

[80] Or, perhaps more accurately, most objectors have never experienced externality in the productions of actions that matter much. The scratching of itches at inopportune times might be a fairly common action that is often external.

feel less prototypically "alien" since it would be a more regular occurrence. And so in further restricting the number of actions the account deems alien, the hypothetical view better preserves the connection between felt alienation and alienation in the sense of failing to identify with one's action.

Despite these considerable benefits, some may protest that hypothetical endorsement is a poor substitute for actual endorsement. Why should we care about what motivations agents *would* endorse rather than what they actually do?

In an analogous discussion about hypothetical consent in law and applied ethics, David Enoch makes the point that the question of whether or not hypothetical consent is a poor substitute for actual consent must be answered by paying careful attention to whether or not hypothetical consent ever has the same kind of normative significance as actual consent. In order to answer that question, he says, we would need to know why actual consent matters and whether or not hypothetical consent can supply those same resources.[81] In the case of consent, Enoch argues that actual consent is sometimes important for reasons of sovereignty and sometimes for reasons of non-alienation. Hypothetical consent can answer to the latter but not the former concern.

Something very similar seems to me to be true of endorsement. There are certain ways of thinking about actual endorsement that cast it as granting motivations with special agential power. For example, David Velleman conceives of Frankfurt's second-order volitions as providing a compatibilist response to agent-causal libertarianism; requiring an endorsement process on his interpretation is an attempt to "put the agent back into the picture" of the causal story of action.[82] If this is the aim, then it's clear hypothetical endorsement won't do. But, in the next couple of

[81] Enoch (2017).

[82] Velleman (1992).

sections I will offer some cases in which non-attributable agents seem to make choices for themselves about what to do via processes that are very similar to the processes of attributable agents. This casts doubt on the idea that what goes awry in non-attributable actions is that the agent is missing from the picture in some way.

I have argued that what is really at issue is the fact that non-attributable agents are alienated from the motivational states that move them into action. And here, it is the coordinated symmetry between who the agent is and what she does and the fact that her action comes about due to a mechanism that plays a part in ensuring that coordinated symmetry that ensures non-alienation. Think about it this way: we're not frightened by the notion of failing to actively pick which motivation to act on and acting anyway—we do so all the time when we let ourselves run on autopilot. What's frightening is the prospect of being autopiloted in directions that have nothing to do with our own perspective of our interests—the motivations that we would give weight to were we to reflect.[83]

6. Approving as Endorsement with a Further Aim than Elimination

A hypothetical-partial-endorsement-based view provides an appealing explanation of why certain paradigm cases of non-attributable action are non-attributable, while, unlike many traditional Deep Self views, does not overextend its reach to classify actions performed out of weakness of will or ordinary actions undertaken *sans* deliberation as non-attributable. Actions that are caused by rogue motivations that stand wholly outside of

[83] What, though, would we make of being perpetually on autopilot in such a way? This is a bit of a frightening prospect, but it might be frightening for reasons that have little to do with motivational alienation. For example, we might just value having phenomenal consciousness of our agential processes for its own sake, or because it helps us clarify our self-perception or gain insight that leads to a sense of narrative coherence. Another confounding factor here is that we might need to not be on autopilot to *form* dispositions to partially endorse acting on certain desires.

the complex of psychological states that inform what agents want for themselves count as non-attributable. For example, straightforwardly compulsive actions, which the agent wholly disavows, will count as non-attributable. Frankfurt's unwilling addict, understood as someone who would give no weight to his desire to take drugs at the time of action, is another perfect example of a person whose action does not speak for him, according to this view.

Such a view also gives the verdict that willing addicts, even those who are merely partially willing at the time of action, are responsible for their taking drugs. That is, an agent who comes to takes a drug due to a chain of mental states that ensure that she, at the time of action, would give some post-reflective weight to her desire to take drugs, acts in such a way that is self-disclosing for the purposes of attributional-responsibility.[84] When we fill in the story in a certain kind of way, this seems to be the right result. For example, imagine a woman whose conflicting first-order motivations are the motivation to take heroin and the motivation to tend to her child, and then imagine her being such that if she were to deliberate about what to do, she would give some weight to each motivation. "On the one hand" she might think to herself, "I want to act on my desire to tend to my child, but on the other hand I do really love heroin a lot and maybe the fun of taking heroin is more important to me than my child's wellbeing…well, I guess my child's wellbeing edges out my desire for heroin, so I guess that's what I'll do?" In this case, if the woman ends up taking heroin due to her desire to have fun, despite the outcome of her actual or hypothetical deliberation, it seems that she is attributionally-responsible for her action.

[84] Other factors may differentiate the wholly and partially willing in terms of the nature of the response that will be appropriate. For example, if the partially willing addict's taking drugs in this case meets the remaining criteria for blameworthiness, this kind of agent may be less blameworthy than a wholly willing addict, but she is nevertheless an appropriate target of blame.

However, consider another kind of case that falls somewhere in between the paradigmatic cases of the willing and unwilling addict: the begrudgingly willing addict. Michael Bratman describes the case as involving an addict who, "since [he] is confident that his desire for drugs will soon so overpower him as to prevent him from acting intentionally, and since the struggle to remain drug-free is extremely painful… decides to cease resisting his desire, and to take the steps necessary for satisfying it." "To be sure," Bratman writes, "he would rather not perform **an** act of drug-taking. Nonetheless, given his options, he would rather perform ***this particular*** drug-taking act."[85] Such an addict does minimally endorse acting on his desire to take drugs, and so would count as attributionally-responsible for his action on the view currently under consideration, yet this doesn't seem like the right result.

To elaborate, the view apparently entails that if the agent were to wholeheartedly resist by not endorsing a desire to take the drug despite inevitably having his will being taken over by the urge, his resultant action would not be attributable. However, if he were to wisely recognize that failure would be the inevitable outcome of his resistance and get it over with more quickly by acting on the desire to take the drug to get rid of the urge faster, his action would be attributable. But it seems that the wrong thing makes the difference in these two cases; the action's connection to the agent shouldn't be determined by whether or not the agent decides to give in to an overwhelming impulse, but rather, by whether or not his desire is the kind of thing he'd wanted to be motivated by in the first place. The begrudging addict's action is still caused purely by non-agential neural activity, no matter how long the agent refrains from acting on it. He may feel just as alienated from his action as the unwilling addict, and would be right to feel this way since he feels compelled to endorse

[85] Bratman (1996).

acting on a desire he would not want to act on were it not for the feeling of horrifying inevitability.[86]

People whose psychological make-up has this structure are not limited to hypothetical addicts. Recent research into Tourette syndrome reveals that the initiation of ticcing behavior often comes about in a similar way.[87] Once thought to be akin to involuntary twitches like muscle spasms, Tourettic tics are now acknowledged to be intentional movements consciously undertaken in response to premonitory urges, which are experienced by agents as alien. More than 90% of people with Tourette syndrome report that their tics are "voluntary" in the sense that they believe they take an active (though subservient) role in the action's coming about.[88] The reported phenomenology echoes the neurophysiological findings. Ordinary action involves signals being sent from an agent's frontal lobe, the site of considered judgments about what to do, to the motor cortex, the initiator of action. But in the brains of people with Tourette syndrome, disordered neural connections between the basal ganglia and motor cortex cause errant signals to encourage the initiation of simple body movements (including the utterance of words) that build up as a sort of mental pressure over time, making the sufferer more and more uncomfortable until she chooses to act on the urge. People with Tourette syndrome can, with some

[86] Some theorists believe that all addiction functions similarly to the way it's described in these cases, that, in a sense, there are no truly unwilling addicts since addiction affects one's judgment about what one should do in this kind of way. See, for example, Buss (2012). Most Deep Self theorists, by way of contrast, take it as given that compulsion often involves an agent's deep self mental states being overpowered by a rogue urge. But this dispute needn't be settled here. It seems that our view should tell us that irrespective of the prevalence of cases of its type, if there were a case of addiction in which an unapproved desire simply overpowered one's approved desires, the agent's action would count as non-attributable.

[87] Schroeder (2005) argues for the importance of the integration of this research into our theorizing about responsibility.

[88] Leckman et. al (1999).

difficulty, use their judgment to postpone the discharging of these urges, but ultimately usually choose to give in to the urges as a way of alleviating the pressure.

Compare the person who utters slurs for the pure joy of harming minorities to the person with Tourette syndrome who utters slurs purely to relieve the unbearable pressure of a strong premonitory urge to tic. The former, unlike the latter, seems to be agentially involved in the right sort of way with the motivation of her action such that her action represents something about her, while the latter's action does not seem to be representative at all. According to the kinds of endorsement views under consideration so far, however, most agents' tics would be attributable, since it seems that people with Tourette syndrome do give weight to the desire to act on their urges, which comes from their desires to rid themselves of the uncomfortable urges.

Something seems to be going awry in the psychology of the begrudging addict and the person with Tourette syndrome such that the relationships they have to their effective motivating desires do not grant the desires the usual authority to speak for them. Here is what I think goes awry. There are oftentimes two functions of acting on a desire: acting on a desire to φ brings it about so that you are φ-ing, but it also in many cases gets rid of your desire to φ. For example, if you want to go to the library, and then you successfully act on this desire, you will be at the library and thus no longer have an occurent desire to go to the library. In most cases in which we endorse a desire to φ we do so because we want the satisfaction of the desire to φ, not merely because we want to no longer want to φ. In cases in which we endorse acting on a desire because we approve of doing so, we do so because we have some aim that acting on our desire satisfies other than its elimination through our action. When you 'endorse' giving in to such an urge due solely to its power, the force the urge exerts on you is no less purely mechanistic than when it overpowers your endorsement. To distinguish, I will call the kinds of endorsements that are not merely due to wanting to rid oneself of one's motivation with no further aim "approvals." So, it is approval rather than just any kind of endorsement that is relevant to attributability.

With all the pieces in place now, we can formulate an account of what it takes to "minimally approve". To "minimally approve" of acting in a certain way is to approve of it in just (at least) a hypothetical and partial sense.

> Minimal Approval: An agent minimally approves of φing-at-*t* iff at *t* she is such that if she were to reflect on her desire to φ at *t*, she would want to some degree to act on it with some further aim in doing so other than merely eliminating the desire.

One more kind of case that sits between the willing and unwilling addict cases is also worth discussing: cases in which a person appears to act in order to eliminate their desire and does so to further some aim she has in eliminating it. Consider the following case:

> Thoughtful Tourette Sufferer: Tom has a Tourettic urge to say the word "dyke," which he knows is a slur that targets lesbians. Tom has absolutely no animus against lesbians and, in fact, he aims to minimize the harms his frequent utterances cause for lesbians who hear them and do not know about his condition. He is currently in the room with one lesbian, and knows that another will additionally enter the room in 5 minutes. He feels an overwhelming urge to utter the slur, one that, with some great effort, he would be able to postpone, but only for about 5-6 minutes. He decides that he prefers to act on the urge sooner rather than later, so as to only harm one person rather than two, and does so.

Is Tom attributionally-responsible for uttering the slur? Our intuitions pull in both directions. Tom seems praiseworthy for avoiding harming an additional person by uttering the slur now rather than later, which seems to indicate that he is attributionally-responsible. But Tom does not seem blameworthy for harming the one person, since his utterance was due only to a need to eliminate a rogue urge. The Minimal Approval view can explain this by showing how he meets the conditions for attributional-responsibility under one description of his φ-ing, and fails to meet them under another.

> Tom seems praiseworthy for uttering the slur now rather than later because the actual sequence of mental states involved in the production of Tom's action guarantees that the agent at *t* is such that if he were to reflect on his desire to "utter the slur now rather than later" at *t*, he WOULD want to some degree to act on it with some further aim in doing so than merely eliminating it.

> Tom does not seem blameworthy for uttering the slur full-stop because the actual sequence of mental states involved in the production of Tom's action guarantees that the agent at *t* is such that if he were to reflect on his desire to "utter the slur rather than not utter the slur" at *t*, he would NOT want to some degree to act on it with some further aim in doing so than merely eliminating it.

While it is controversial whether we ought to individuate actions and desires this finely, and thus whether or not our concepts of praise and blame ought to be sensitive to these distinctions, I think the intuitive reactions to these kinds of cases tell in favor of doing so. The Minimal Approval View can help to make sense of our intuitions that initially seemed contradictory.

7. Minimal Approval and the Mechanism of Action

I have argued that minimally approving, in the sense I have articulated, should play a central role in attributional-responsibility, but I have not yet specified just what the relationship is between minimal approval and the causal chain that leads the agent to act. The fact that the agent would minimally approve of φ-ing and the fact that she in fact φs, at minimum, should not be wholly coincidental. Consider the following case:

> Coincidental Compulsion: Carrie has a compulsion to do a jumping jack at 3pm every day. She has absolutely no ability to stop herself from acting on her first-order urge to do it, but she almost always hates that she does it. It often interferes with whatever else she actually wants to be doing at 3pm, but her first-order urge takes over and she does the jumping jack anyway. Today, as 3pm rolls around Carrie is feeling a bit low energy, and thinks to herself that doing a jumping jack might actually be nice for helping her wake up a bit, and so she desires to some degree to act on her desire to do the jumping jack for a reason other than just to get rid of the urge. However, she will be in a work meeting at that time, and all things considered she definitely would not prefer to act on the urge, since it will be embarrassing and difficult to explain to her colleagues. At 3pm the urge overtakes her considered opinion about what she wants to do entirely via the exact same mechanism it does every day (one that is not sensitive to any information about what she approves of doing), and she does the jumping jack.

Carrie, it seems, is not attributionally-responsible for her jumping jack. Even though she did minimally approve of doing the jumping jack, this had absolutely nothing to do with what caused her to act. Minimal approval seems to matter only when it is non-coincidental that the agent acts in accordance with what she minimal approves of doing.

But, given the fact that minimal approval is hypothetical, we can't explain the connection by telling a simple causal story on which the approval causes the action. It *might* be the case that an attributionally-responsible agent explicitly reflects, approves of acting on one of her desires and this approval itself brings her to act. But this is not the only possibility on which her approval and the causal chain leading her to act might be non-coincidentally related. The process leading to her action, or some part of it, might cause her to be such that she would minimally approve. Or she might have a mental state that both her minimal approval and her action. The common principle shared by these relations is that for some particular chain of mental states leading to action, C, which leads the agent to φ, if the agent's φ-ing is caused by C, then the agent minimally approves of φ-ing.

But now consider the following case:

> Conditional Fallacy Frankie: Frankie is addicted to alcohol and would give anything to stop drinking, but despite his other plans for himself he is often overtaken by an extremely powerful urge to drink. One day, while feeling hopeless about his powerlessness against his urge he decides to think hard about whether or not he should just form and act on a desire in order to act on it before it overpowers him. Ultimately, he does. Given what I have been arguing, it seems he should not be attributionally-responsible. His thinking leads him to act and so is part of the mental states making up the causal chain that leads to his action. But, now imagine that this same thinking also leads him to put his head down on some objects on his desk, one of which he hits hard enough that it stimulates his brain to have certain mental states, mak-

> ing it the case that at the moment of action he minimally approves of acting on his desire to drink alcohol.[89]

In this case, Frankie's causal chain of mental states leading to his action does make it the case that he minimally approves of acting on his urge to drink. But, intuitively, this is not the kind of connection that would make him attributionally-responsible for his resultant action.[90] Instead I think the relevant connection is not that the causal chain of mental states leading to the agent's action actually makes it the case that the agent minimally approves, but rather that the causal chain of mental states disposes the agent to minimally approve in the vast majority of cases (all non-fluky cases).[91] [92] The thought here is that it is not any particular counterfactual that matters for attributional-responsibility, but rather the fact that the agent has the actual mental states that dispose her to approve, whatever they may be. We're not really interested in what the happens, that, generally speaking, in (most of) the worlds in which the agent considers this

[89] I call this case "Conditional Fallacy Frankie" because the problem here may be an application of some version of the Conditional Fallacy. For discussion of the Conditional Fallacy in general, see Bonevac, Dever, and Sosa (2006).

[90] Less science-fiction-y cases can also be generated, although it is less clear that they pose a genuine threat to the view as stated. For example, a person who never reflects on her motivations might undergo an organic personality change if she were to reflect on her first-order desires. It is not wholly clear that these post-personality-change hypothetical approvals are not what matters for attributional-responsibility. But if this seems problematic enough to provide reason to think that the view should say that certain aspects of agent's personalities ought to be held fixed, this gives additional reason to move to a dispositional view.

[91] This means that there will also be inverse Conditional Fallacy Frankie cases on which the agent does not actually minimally approve due to some fluke caused by the causal chain leading to action.

[92] The reader is invited to insert her own favorite theory of the metaphysics of dispositions here.

particular choice situation, she approves to some degree of her desire. In ruling out certain outlier worlds, moving to a dispositional account ensures that there is actually some feature of the agent's psychology that plays the relevant role here, rather than a mere fluke of the environment.

One final tweak must be made before an adequate account can be formulated. It seems important to attributional-responsibility that the mental states involved in the causal chain of action, as well as in the agent's minimal approval are really her own. By that I mean something quite modest, just that they are not inserted, say by an evil scientist or the like.[93] There is an extensive literature here, and I am aware that making this concession opens the door for potential incompatibilist arguments to get a foothold.[94] However, I reluctantly adopt a modest ownership condition since it seems to me a trade-off worth making.[95] Although the account loses theoretical simplicity, it gains the ability to make good on some of our strongest intuitions about attributability.

Putting all of these pieces together, we can now formulate the Minimal Approval view:

> The Minimal Approval View of Attributional-Responsibility: An agent is attributionally-responsible for φing-at-*t* iff the actual sequence of mental states involved in the production of her action is non-implanted and together with her other mental states makes it the case that at *t*, if she were to reflect on her desire to φ at *t*, she would be suf-

[93] See Matheson (2018) for a recent discussion of Manipulation Cases as they pose counterexamples to Deep Self views.

[94] The most famous of these argument, Derk Pereboom's Four Case Argument aims to show that there is no relevant difference on which to base an ownership condition between a case in which an agent's deep self mental states are implanted and when they are acquired in the normal fashion, given the truth of determinism (Pereboom 2001).

[95] For a recent defense of adopting a modest ownership condition to protect against manipulation cases, see McKenna (2016).

ficiently likely to want to act on her desire to φ at *t*, with some further aim in doing so other than merely eliminating this desire.[96]

8. Advantages of the View

8.1 Solving Frankfurt's Infinite Hierarchy Problems

Though the idea that higher-order desires have something to do with our agential identities is intuitive, the fact that higher-order accounts like Frankfurt's endorsement account have problems dealing with infinitely-ascending orders is often assumed to make higher-order accounts non-starters. Watson first frames his valuing view as a view that can avoid the problems of infinite hierarchies caused by hierarchical account. Frankfurt himself has been aware of these sorts of issues since his first article discussing the view, and has modified his view numerous times in order to try to better handle the problems. In this section I will show how two of the most commonly discussed problems of infinite hierarchies are caused in part by contingent features of hierarchical views that Frankfurt's account has but that the Minimal Approval view lacks.

In order to see how the first problem of infinite hierarchies arises, first consider a case in which there is a conflict at the second-order about which first-order desire to endorse. Consider first an agent deciding between going to the gym and doing some grading who then acts on the desire to go to the gym because she forms a second-order volition that is constituted by a desire to have her desire to go to the gym be the one to move her to action. But it seems very possible to have an agent who is not only first-order conflicted but also second-order conflicted; *this* agent not only wouldn't be sure whether to go to the gym or get some grading done,

[96] "φ-ing" should be understood here as standing for an action, and the account should be taken to cover only attributional-responsibility for *actions*. Attributional-responsibility for omissions and for consequences is, on my view, derivative on attributional-responsibility for actions. I explore some of these issues in Chapter 5.

but additionally, she wouldn't be sure whether to form the desire to act on her desire to go to the gym or to form the desire to act on her desire to get some grading done. Since the way Frankfurt seems to imagine that conflicts among desires of the same order are resolved is by moving to a yet higher order of reflection, we would then have to look at the agent's third-order desires to be moved by second-order desires to be moved by first-order desires. If these desires were also in conflict, we would have to look at the agent's fourth-order desires, etc. This process seems like it could too easily be such that it never terminates in wholehearted endorsement and so we might worry that this is just a bad candidate description for how the process that leads to autonomous action works in agents like us. Furthermore, it seems as though an agent who never decisively reaches a level at which she is wholly unconflicted can still act in ways that she should still be held attributionally-responsible for.

But we are now in a position to see how this kind of problem of infinite hierarchies is related to contingent features of a higher-order desire account. One feature of Frankfurt's view is that it seems to imply that wholeheartedness is necessary for an agent's process to issue in action. This is what makes conflict at the *n* level a *problem* for understanding how the agent could come to act, and forces us to examine level *n+1* to find the wholehearted endorsement. If we instead admit, as the Minimal Approval view does, that we may act despite having not settled conflicts among desires at the second-order, we are left without reason to worry about third-order desires as a way of mediating conflict in order to act. This is not possible on Frankfurt's view because according to Frankfurt the reason second-order desires matter is in large part because they help settle conflicts between first-order desires leading the agent to make a decision that issues in action.

The second, and more troubling, problem of infinite hierarchies is the one Watson speaks most directly to in "Free Agency" when he writes, "Can't one be a wanton, so to speak, with respect to one's second-order desires and volitions?"[97] Watson's concern is that just as an agent can be alienated from her first-order desire in the sense that she has not endorsed it, she can be alienated from her second-order desires as well. Since first-order desires on Frankfurt's view do not have the authority to speak for the agent unless they go through a process involving second-order desires, it seems as though there is no reason to suspect that second-order desires wouldn't also have to go through a higher-order process of their own to be granted such authority. Ever-higher levels would need to be appealed to in order to generate the appropriate authority, making it impossible for any desire to have the authority to speak for the agent. What we should conclude from this, Watson thinks, is that second-order volitions do not really have the authority to speak for an agent since they do not guarantee that the agent is not alienated from them.[98] Even if an agent forms a full-fledged unconflicted second-order volition, we have no reason to privilege the authority of that state over a third-order volition to not want to act in accordance with the desire that aligns with her second-order volition. Since second-order status does not automatically grant second-order volitions any special authority, it seems they have no more claim to represent the agent than any other candidate mental state.

But on Minimal Approval non-alienation is already established by bare reflective endorsement of any strength. It is a contingent feature of a hierarchical view that we should be concerned with determining the most decisive or most authoritative endorsement. On the Minimal Approval view, even if the agent has a wholehearted third-order desire that repudiates the second-order desire, as long as the second-order desire exists, then this is not enough to establish full alienation since the agent's action would still come about in a way that is related to something that the agent

[97] Watson, (2004): 28.

[98] Velleman (1992) argues that this problem is more general and is a problem for Watson's own view as well.

would want for herself upon reflection (unlike when she is motivated by a rogue first-order desire). While higher-order desires may reveal further truths about us, attributional-responsibility is not about revealing the deepest truths about ourselves, but rather, it is about meeting a very minimal set of ownership conditions of our actions ensuring that their source is not completely alien to our interests. To be attributionally-responsible for our actions we must only be invested in them in some reflective way that goes beyond a bare desire. Once this reflectivity is established, there is no need to worry about ever-ascending orders of reflection.

8.2 Capacity Without Process

We tend to think that human agents' actions can put them on the hook for what they have done in a way that non-agents', like non-human animals' behavior does not. We think that the kinds of moral responsibility responses that are appropriate for persons go above and beyond the ways it is reasonable to respond to a dog. Planning, endorsing, and valuing versions of Deep Self views generally aim to be able to make this distinction by positing some special kind of capacity that agents actively exercise when they undertake actions for which they are attributionally-responsible. For example, on the endorsing view, an agent exercises her capacity to choose which action to initiate by picking amongst her first-order desires.

But this feature of these views is also problematic, because the idea that we as agents go through some sort of special mental process each time we act in an attributable way is implausible.[99] The problem is that Frankfurtian descriptions of agents selecting which first-order desire to act upon, or Bratmanian descriptions of agents simultaneously acting and making choices about how to settle conflicts amongst desires in the future, just do not seem to be what we ordinarily do as agents in everyday life. Our conduct for which we are often rightly held morally responsible is

[99] See, for example, Arpaly (2002, 2006), Smith (2005), and Buss (2012).

sometimes spontaneous, initiated by subconscious motivation, out of character, or brought about in a fit of emotion. Furthermore, as the results of numerous social psychology studies appear to show us, we sometimes lack reflective access to some of the motivational influences on our actions, perhaps in ways that would implausibly preclude us from being attributionally-responsible for a large range of actions given the conditions of these more agentially demanding Deep Self views.

On the Minimal Approval view, though, action brought about by subconscious processes can still meet the requirements for attributability. The process that causes action needs to make it so that the action is in line with what the agent to some degree would want for herself, which requires the agent to be the sort of creature who has the capacity to form higher-order desires. But that capacity need not be exercised in the form of actual reflection, thus avoiding the charge that traditional Deep Self views face that in ruling out animal action, they rule out too much.

While the Minimal Approval view has this advantage, the minimalist nature of the view might seem to run a different risk, namely, being unable to plausibly explain why agents can be attributionally-responsible in a way that most non-human animals cannot. The Minimal Approval view is relatively silent about the nature of action-production and so even a creature with relatively instinctual or mechanical sorts of action production could, in theory, be eligible for attributional-responsibility. However, there is good reason to think that most non-human animals in fact do not have the capacity to form higher-order desires since they do not have the capacity for higher-order thought, and so would not be candidates for attributional-responsibility according to the Minimal Approval view.

While it is not wholly uncontroversial, the idea that humans are unique among animals in being able to have higher-order thoughts has a rich history of support in the literature on consciousness. Some argue that higher-order thoughts require possession of an I-concept in such a way that the thinker can understand themself *as* a self in a way that involves complicated linguistic capacities that lower animals do not possess.[100] Daniel Dennett argues that the ability to say which mental state one is in is more fundamental than having a higher-order thought, and provides the basis of having a higher-order thought. Since animals cannot say which mental state they are in, according to Dennett, they do not have the capacity for higher-order thought. Others question whether or not lower animals are capable of possessing mental-state concepts at all[101]; preliminary studies seem to indicate that they are not.[102]

So a significant advantage of the Minimal Approval view is that it can preserve the distinction between the way agents and non-agents reveal themselves through action due to their having special capacities, while simultaneously allowing that agents may not actually engage in such methodical deliberative processes when they act in ways that they can be held attributionally-responsible for.

8.3 Criterion Operates Independently from the Type of Mental State that Causes Action

A further advantage of the Minimal Approval view is that, unlike some Deep Self accounts, it is compatible with attributable actions being caused by mental states of any type, just so long as they meet the requirement of being appropriately related to hypothetical approval states. This agnosticism about what the actual process of action production looks like is an advantage for at least two reasons.

First, unlike several Deep Self views, accepting the view does not require accepting controversial positions in moral psychology. For example, the valuing view as advocated by Gary Watson assumes at least some form of motivational judgment internalism, and the caring views of Chandra Sripada and David Shoemaker rely on accepting that there exist

[100] See Quine (1995), Bermúdez (2003), and Bennett (1964, 1966, 1988).

[101] See Davidson (1984, 1985), and Bermúdez (2003).

[102] See, for example, Povinelli and Vonk (2004).

complex states or dispositions that we can identify as caring states, which are distinct from mere desires and play a central role in action production. In contrast, the Minimal Approval view is compatible with each of these pictures of action production, but its proponents can remain agnostic about which sorts of states have the ability to motivate. It is even consistent with extremely minimal theories of action production including simple forms of Humean psychology. This might make the view attractive to those who are averse to more traditional Deep Self views due to the more complicated systems of action-production that they posit.

But there is a further advantage to the fact that the Minimal Approval view does not posit that any particular kind of mental state must be involved in the causal chain in order for an action to be attributable: certain mental state types may sometimes produce attributable action and sometimes produce non-attributable action. In order to illustrate this point, I want to focus on a class of actions to which Deep Self theorists have perhaps paid insufficient attention: actions in which agents act directly "out of" emotions. Emotions are often thought to be partly constituted by motivational states or, at the very least, they are generally thought to have some unmediated influence on motivation. This accords with the common-sense ideas that we can "strike someone out of anger" or "hide out of embarrassment." Such actions often characteristically do not align with our plans, values, cares, or second-order volitions concerning what we think the most preferable thing to do is in a given situation. Imagine, for example, an anti-retributivist who nevertheless is swept up in a wave of vengeful anger, or an ethically non-monogamous person who has disavowed the appropriateness of jealousy being nevertheless moved to action by it. While these actions fail to align with Deep Self states, they nevertheless seem to be the sorts of things for which one can be attributionally-responsible.

On the other hand, if a theory were to hold that a person is attributionally-responsible for *any* action done out of emotion, it would not be viable.

Actions caused by psychological and neurological disorders that we intuitively tend to think exculpate an agent from attributional-responsibility can, it seems, cause action by impacting the agent's emotional state such that she "acts out of" a given emotion. If this is right, a theory of attributional-responsibility should have a way of distinguishing between cases of acting out of an emotion that involve one's agency in the right sort of way and ones that circumvent agency.

One lesser-known disorder illustrates the importance of drawing such a distinction. Misophonia is a neurobehavioral syndrome in which certain ordinary human-produced repetitive sounds, such as the sounds of others chewing, sniffling, or clearing their throats, trigger reactions of anger, disgust, and fear in otherwise psychologically healthy individuals.[103] While research on misophonia is in its infancy, it is hypothesized that the cause of such reactions is extra-connectivity between a set of emotional processing centers of the brain and the anterior insular cortex, the site of interoception (the ability to sense what is happening to one's own body) in the brain.[104] Due to this over-connectivity, ordinary sounds cause these sufferers to react emotionally as if these innocuous sounds are threats, setting off fight-or-flight reactions. When misophonia sufferers are in "fight" mode, their anger is not just an expression of being overwhelmed, but rather, tends to take the form of a directed expression of anger and disgust towards the source of the offending sound. To be clear, this is not just anger towards the person for making a sound that they know bothers the sufferer, as anger can be just as strong towards those making sounds who do not realize their sounds are upsetting to the sufferer. Crucially, at the very same moment in time that she acts out of anger, a person with misophonia is able to acknowledge that it makes no sense to be angry and that, for example, making sounds while chewing is entirely innocuous. Due to

[103] See Braut et. al (2018) for a cross-disciplinary review of the research on misophonia.

[104] See Kumar et. al (2017) , Edelstein et. al. (2017)

these irrational, embarrassing, and inescapable responses, people with misophonia often live increasingly reclusive lives as the disorder progresses in order to try to avoid both sounds and accidentally lashing out at those who they know have done nothing wrong.

Consider the following pair of cases:

> Manners Mary: Manners Mary was taught as a child to always chew with her mouth closed and greatly appreciated the value of the lesson. Following her parents, she grew up believing that a decline in manners in society was the root of much evil and that it is deplorable that some people chew with their mouths open. In her adulthood, she has come to see this as a bit overblown, but she has retained the sense that it's bad form to chew with an open mouth as well as an accompanying sense of disgust when she sees others behaving with such poor manners. At an important dinner party she notices her fellow guests chewing with open mouths, and thinks to herself that someone ought to tell them to stop, and that perhaps if no one else does, she should be the one to do so. However, she knows that these guests would only be offended and would not change their ways if she were to mention their behavior, and so decides that this would probably be a bad time to say something. However, she fails to hold her tongue, gets increasingly angry, and yells out "chew with your mouth closed!" despite knowing that everyone will only be offended and not change their ways.

> Misophonia Mary: Misophonia Mary does not care about manners in the slightest and makes no effort to chew with her mouth closed. However, she suffers from misophonia, which makes her inexplicably angry when she hears people making chewing sounds. Even when in the throes of an episode of misophonia she recognizes that there is nothing wrong or bad in any way about eating with one's mouth open, yet due to errant signals in her brain that trigger a fight-or-flight reac-

> tion, Mary feels compelled to flee or else lash out at those making the sounds. With nowhere to flee to at a dinner party, out of anger Mary yells out "chew with your mouth closed!" despite knowing that everyone will only be offended and not change their ways.

Intuitively, it seems we should hold Manners Mary attributionally-responsible but not Misophonia Mary, though they are both most directly motivated by their anger. This is some indication that our view should be consistent with the fact that acting out of anger is neither sufficient for attributional-responsibility nor disqualifying for it.

But because both agents' actions are motivated by anger and not suitably related to their plans, endorsements, judgments, or possibly even cares, traditional Deep Self views will have a difficult time explaining why Manners Mary's action is attributable and thus licenses a different response than Misophonia Mary's.[105]

The Minimal Approval, by contrast, is well-suited to explain the contrast. If we were, at the time of action, to ask Manners Mary to consider her motivation to yell out at the guests chewing with their mouths open she would give *some* weight to that option. After all, she thinks it is somewhat important that such ill-mannered behavior not go wholly ignored. But it does not seem appropriate to hold Misophonia Mary attributionally-responsible for her yelling, and the Minimal Approval view shows how her motivation stands outside of her agency. If we were to ask her whether or not she would like to be motivated to some degree by the desire to tell the dinner guests to stop chewing she would say, even in a moment of her anger, that she has no desire to be so moved. The only reason she

[105] It might be argued that Manners Mary's action is related to her judgments in a way that Misophonia Mary's action is not. However, it is clear that Manners Mary does not take her action to be among the best options. She might take herself to have a *pro tanto* reason to act as she does, but acting in accordance with one's *pro tanto* reason is not sufficient for attributional-responsibility. See Chapter 3, §4.2 for further discussion.

might give any weight to the desire to lash out would be to relieve the psychological pressure of not saying anything, caused by her involuntary fight-or-flight reaction, and thus she would fail to meet the conditions of the Minimal Approval view.

This pair of cases helps illustrate the fact that the Minimal Approval view can hold that agents who act out of emotions are often attributionally-responsible for their actions while leaving room for the possibility that emotional motivation may factor prominently in action caused by non-agential neurological activity for which we should not hold agents attributionally-responsible. The set of cases in which agents act out of emotions helps illustrate the broader point that in having a set of criteria for attributability that does not require any particular mental state *type* to feature in the action-causing sequence it has better flexibility for handling some of the nuances of attributability and neurological dysfunction. This gives the Minimal Approval view yet another significant advantage over traditional Deep Self views.

Chapter 3: *How Should Deep Self Theorists Account for Weakness of Will?*[106]

1. Introduction

In the first two chapters of this dissertation I articulated and defended a new view of attributional-responsibility, the Minimal Approval view. Along the way I illustrated some of its advantages over more traditional Deep Self views. One such advantage mentioned but not yet explored in depth is the view's ability to account for the attributional-responsibility of weak-willed agents. In this chapter I zoom in on this advantage of the view and argue that the Minimal Approval view is *uniquely* well-suited among Deep Self views to account for the difference between weak-willed actions, which are attributable to agents, and compulsive actions, which are not.

I begin by showing how deeply entrenched the weakness of will problem is for Deep Self theorists. Traditional Deep Self views lack the resources to adequately distinguish compulsion from weakness of will, which leads to their wrongly classifying certain attributable weak-willed actions as non-attributable. Most current solutions involve implausible bullet-biting, or else cede dialectical ground to control-based theorists. I suggest that Deep Self theorists instead must adopt an understanding of the self as having multiple strands that are competing yet individually have the power to speak for the person *qua* agent. The Minimal Approval view, I argue, is the only realization of this picture that appropriately separates non-attributable compulsive action from attributable weak-willed action. Thus the Minimal Approval view has a distinctive advantage over

[106] This chapter benefitted from discussion by the participants at a 2018 Eastern APA symposium session.

competing Deep Self views. I conclude by addressing a family of fairness concerns for the Minimal Approval view's treatment of these cases.

2. The Weakness of Will Problem

Consider the following case:

> Sam's Exam: Suppose Sam knows that she should be studying for her final exam, even though there's a party going on that night that she really wants to go to. She judges it would be best for her to study for her exam, endorses her desire to study for her exam, plans to study for her exam etc., and yet somehow she ends up going to the party anyway. Sam does not act on what she judges it best to do, or in line with what she plans to do, or on the desire she higher-order endorses. But common sense tells us that Sam's action is still self-expressive and so she nevertheless ought to be attributionally-responsible for her action.

This poses a problem for traditional Deep Self theorists because, according to their views, agents are not attributionally-responsible for their actions unless their actions are aligned with their planning, endorsing, valuing, or caring states. Traditional Deep Self views thus face an important objection: they counter-intuitively hold that we are not responsible for weak-willed acts, and so fail to provide a necessary condition for attributability. Perhaps, then, the bounds of the Deep Self could just be drawn differently so as to cast a wider net around the class of attributable actions?

The problem, however, is more deeply entrenched than this. Deep Self views are custom-made to show how acting on desires that the agent herself does not see as most favorable undermine agency in such a way as to make such actions non-attributable, as in the case of compulsion. But weakness of will is usually described as the failure to act in accordance with what one finds to be the most favorable course of action, and yet we do intuitively think weak-willed actions are self-expressive.[107] One of the motivations for Deep Self views is making sense of the claim that acting attributably just involves doing what one *really wants* to do, where "really wants" is qualified in such a way as to bracket off compulsive desires, which Deep Self theorists see as having the power to coerce agents 'from the inside,' so to speak. But they do this by bracketing off motivations that are counter to what agents judge best, endorse, or plan for, which also inescapably seems to bracket off weak-willed actions. So traditional Deep Self views not only incorrectly classify weak-willed actions as non-attributable, they also seem to do so almost by design.[108]

[107] I should note at the outset that I am operating under the assumption that compulsion is primarily a conative or volitional phenomenon rather than a merely cognitive one. It is possible that some things that we colloquially call 'compulsions' may instead be instances of acting in accordance with pathologically acquired beliefs about the world. For example, an agent might come to believe that her house would burn down if she didn't check that her oven was turned off 18 times. Suppose she does so instead of getting to a meeting on time. On a view like the Minimal Approval view, we could easily explain how she would be exempt from blameworthiness by showing how even though her oven-checking is attributable to her, given the background conditions, which include her epistemic state, she shouldn't be blameworthy for doing what she reasonably judged to be the best course of action. If this account of why compulsion exempts agents from blame could be made plausible, there are two ways it might be bolstered to at least address the distinction between compulsion and weakness of will. One possibility is that it could be argued that weakness of will is a truly volitional occurrence while compulsion is cognitive. Another route would be to argue that both compulsion and weakness are cognitive phenomena and the distinction lies in differences in the beliefs, or the manner of their acquisition. Even if some compulsive cases do function in this this way, I am deeply skeptical that all compulsive cases function this way. If that's right, there is a still a serious problem to solve here. I am grateful to Sarah Buss for pressing me to articulate this assumption.

[108] This problem has been articulated many times, notably in Vihvelin (1994), Haji (1998), Haji (2002), Fischer (2010), Fischer (2012), McKenna (2011), McKenna and van Schoelandt (2015), Strabbing (2016), and McKenna (*forthcoming*), and is often considered by many to be a knock-down objection to Deep Self views, despite Deep Self theorists having tradi-

Take, for example, Bratman's planning view. On Bratman's view an agent's act is only attributable if it meshes with her self-governing policies and plans about which desires of hers to act on. So if an agent makes a self-governing policy to always act on her desires to stay home and study before an exam, but then when the time comes is unable to get herself to act on her self-governing policy, her action will be non-attributable. But Bratman is also inclined to describe weakness of will in this way, as an agent's inability to get herself to act on her own self-governing policies. In short, weakness of will is usually defined as a failure of self-governance, and self-governance is usually what is required for attributable action. This makes the fact that weak-willed actions *do* seem blameworthy very difficult to accommodate on Deep Self views. This is a very serious problem for these theories since weak-willed cases appear to many as cases in which it is particularly appropriate to hold agents responsible; in fact, Gideon Rosen has even argued that agents are responsible for their actions *only* when they are weak-willed.[109]

We do, intuitively, want to say that compulsive cases are cases of non-attributable actions, and Deep Self views are able to give a richly explanatory account of why agents are not responsible for compulsion. When acting on compulsive desires, agents' standpoints are overpowered such that their resultant actions do not express anything about what it is they really want to do. Yet the attraction of this proposal is severely undercut by the

tionally relatively little to say about it. For example, Fischer writes of Frankfurt's view, "The problem of weakness of the will is, in my view, a decisive problem for Frankfurt's approach. Somewhat surprisingly, it has not received nearly as much attention as the so-called 'regress' problem. (Indeed, I am not aware of any discussion of the relationship between his account of acting freely and the problem of weakness of will by Frankfurt)" [Fischer (2010)].

[109] Rosen (2014) argues that agents could only be responsible for weak-willed actions because only weak-willed agents meet the high epistemic standards for knowing wrongdoing that he takes to be a prerequisite for blameworthiness.

fact that weak-willed cases have these same features yet intuitively these actions *are* attributable. It seems there must be some way of distinguishing compulsion from weakness of will as well as what makes the latter but not the former attributable, but traditional Deep Self views seemingly have no resources available to make such a distinction. Since it seems the distinction between compulsion and weakness of will must be found elsewhere, we may wonder if we should look elsewhere, too, for an explanation of what makes agents non-blameworthy for compulsive actions.

3. Current Solutions and Their Problems

It is crucial that Deep Self theorists find some solution to this problem. In this section I will discuss existing solutions to the weakness of will problem, but I want to be clear that I will only be addressing them as resources *for a Deep Self theorist* to tell compulsion apart from weakness of will and also to correctly classify weak actions as attributable and compulsive actions as non-attributable. While non-Deep Self accounts of responsibility such as control-based and reasons-responsive accounts will also have to differentiate compulsion and weakness of will, these cases do not necessarily pose any *special* problem for them, as the distinction might fall out as a result of the account's calibration of the degree of control or reasons-responsiveness required for responsibility. I will ultimately be arguing that this way of dividing up the cases is misguided, and hope to provide a compelling alternative picture, but giving a positive argument for my view over the solutions these non-Deep Self accounts posit is outside the scope of this chapter. Instead I will focus on the in-house debate over how to handle these cases that exists within the broadly Deep Self family of views. And so in discussing the options currently on the table, I will be analyzing them only as they are compatible with the aims of Deep Self theory.

3.1 Mismatched Accounts

One possible way out of the puzzle, though it has not been much discussed in the literature, would be for the Deep Self theorist to deny that it is the very same kind of mental state that weak-willed agents fail to act in accordance with that also grounds ascriptions of attributional-responsibility. For example, a Frankfurtian could say that weakness of will consists in an agent acting contrary to what she judges best, while attributable action does not require acting in accordance with what one *judges* best, but rather, with the desires one *endorses*, since Frankfurt leaves open the possibility that agents may endorse desires arationally. So it is possible for Sam to judge it best that she should stay home and study, while endorsing her desire to go to the party. This at least shows that it is not impossible for a Deep Self view to account for *some* weak-willed actions being attributable.

But unless it is plausible that in every case of weakness of will the agent endorses her course of action in Frankfurt's sense, the view still faces a weakness of will problem. All we need is one case to regenerate the problem that the view fails to identify a necessary condition for attributable action. A familiar case from Huckleberry Finn is helpfully illustrative.[110] Huck Finn befriends a slave named Jim and helps him escape from slavery. While on a raft being used for the escape, Huck is plagued by what he refers to as "conscience." He believes, as do other white people in his society, that helping a slave escape amounts to stealing, and that stealing is morally wrong. He judges that the moral wrongness of this action outweighs the demands of loyalty to one's friends, and never considers the idea of doubting what his society has told him is right. Though he resolves to turn Jim in because he judges it to be the right thing to do, he nevertheless finds himself psychologically unable to follow through on

[110] This case first appears in Adams (1985), and is discussed in regards to Deep Self views in Arpaly (2002).

this resolve, chastising himself for remaining a “bad boy.”[111] While this case has a non-standard feature of examples of weakness of will cases, since what Huck ends up doing is the right thing rather than the wrong thing, this is still clearly a case of weakness of will. It also seems that Huck is attributionally-responsible for helping Jim escape from slavery, and potentially even praiseworthy.

Having the mismatch of a judgment-based conception of weakness of will and an endorsement-based conception of attributability will not help with cases like Huck’s. On the judgment-based conception of weakness of will, Huck judges it best to turn Jim in but ends up helping him escape nevertheless. So we correctly count Huck as weak-willed. But Frankfurt’s traditional endorsement view won’t lead to the result that Huck is attributionally-responsible, because it does not seem plausible to suppose that Huck in Frankfurt’s sense endorses his first-order desire to help Jim escape. While it’s plausible that Huck may endorse a more general desire to help his friends, or desire *to some degree* to help Jim, he clearly does not form a second-order volition to act on his desire to help Jim escape. In fact, he desperately wants to be moved by his desire to do what he takes to be the right thing, so this looks like a case in which the desire he *most wants* to be moved by is his first-order desire to turn Jim in. So we have a case on which the mismatched view under consideration still incorrectly predicts that a weak-willed action will not be attributable, when intuitively it is.

Other mismatched accounts will be even less plausible, since it’s quite clear that there are weak willed actions that agents are attributionally-responsible for in which they neither judge the course of action they undertake to be best nor have they planned or committed to act in accordance with it.

[111] Arpaly (2002): 75.

3.2. Skepticism About Weakness of Will

While experiences of weakness of will seem to be very common, there is a long history of skepticism about their existence. As Gary Watson points out there are at least two different ways to be a skeptic about weakness of will.[112]

Traditional Socratic weakness of will skepticism consists in the claim that no one ever truly acts against her best judgment. While there is nothing *logically* incompatible about such denial and a Deep Self approach to attributability, denying the possibility of acting against one's better judgment is counter to one of the primary motivations of Deep Self theory. One of the main things that the Deep Self is used to explain is the fact that when an agent acts on a motivation external to her standpoint, like the unwilling addict does, she is not attributable for such action. So coupling such an approach with the idea that a person never acts in ways that conflict with what she judges best would be an unusual move. Socratic skeptics usually argue that all of our acts are what we at the time of action really judge to be the best courses of action. Otherwise, they allege, we wouldn't act in such ways. Coupled with Watson's view, the claim that we always do what we judge to be best leads to the conclusion that all action is attributable action. Frankfurt's view sits strangely with Socratic skepticism as well. One of the most interesting parts of Frankfurt's view is that it allows for some actions to count as attributable that are not judged best, but there just are no such actions if weakness of will skepticism is true.

A more natural kind of skepticism for the Deep Self theorist to adopt is skepticism that anyone ever *freely* acts in conflict with what she most wants to do.[113] For the Deep Self theorist this would amount to biting the

[112] Watson (1977).

[113] Gary Watson considers and argues for such a view in Watson (1977). However, his argument only speaks to views on which freedom is understood as being related to control, so it does not provide an argument for favoring skepticism over the view I ultimately de-

bullet and accepting the consequence of her theory that weak-willed actions for which agents seem to be attributionally-responsible are really compulsions for which they are not attributionally-responsible, or else are misdescribed. In the absence of a compelling argument for such skepticism, the Deep Self theorist opens herself up to a charge that such skepticism is *ad hoc*. But even more worryingly, such a view would be *highly* revisionary. While this in and of itself is not a knock-down argument, the view would have some very difficult consequences to accept. To illustrate, consider again the case of Murderous Max from Chapter 2 (reprinted below):

> Murderous Max: Max strongly desires to go out on a killing spree this morning because he hates people and there is almost nothing he likes more in the world than shooting them—in fact he thinks of himself as having made an art of it, perfecting his technique more and more with each kill. There is one thing, however, that he cares about even more: he is extremely committed to his morning workout routine. As much as he wants to go out on a killing spree, he also realizes that if he does that, he'll have to forego his morning workout ritual. He knows that if he misses even one morning of working out, he'll probably fall off of his routine, and he'll thus sacrifice the progress he hopes to be making. So, after considering and giving some weight to each option, he decides that acting on his desire to work out is what he most wants to do, it aligns best with his values, and is consistent with the plans he has set for himself. However, his desire to hone his murderous craft by going on that killing spree ends up just being so intense that he caves from lack of willpower and goes out and does the deed.

fend in this chapter. For critiques of the argument in Watson (1977), see Watson (1999), Strabbing (2016), and Watson (Forthcoming).

Is this case either misdescribed or else Max is not attributionally responsible for his killing spree, as the skeptic would have it? The characterization of his mental process seems normal enough (despite, of course, the strange content of his thought process). And it seems very difficult to believe that his action in this case is not self-disclosing for the purposes of attributional-responsibility. The fact is, he *does* desire to a great extent to act on his desire to kill people, and it doesn't seem right to excuse him for his action just because he wanted to act on another motivation *more*.

3.3. Attributability for Failure to Exercise a Deep Self Capacity

In a recent paper, Jada Twedt Strabbing suggests that, in response to the weakness of will problem, Deep Self views ought to allow that in addition to exercises of attributability-relevant capacities, *failures* of attributability-relevant capacities also are self-expressive for the purposes of attributional-responsibility. This is Strabbing's principle, which she says can be adopted by any Deep Self theorist:

> Having the Capacity (HC) Principle: An agent is attributionally responsible for an action A if and only if 1) A results from the exercise of his attributability-relevant capacity to do A or 2) A results from the failure to exercise his attributability-relevant capacity to avoid doing A.[114]

Now, just as objectively wrong actions are not by themselves sufficient for attributional-blameworthiness, failures to act may constitute wrongdoing but are not by themselves sufficient for attributional-blameworthiness. What Deep Self views add to wrongdoing to generate attributional-blameworthiness is some form of agential authorization or assent that shows us that the agent's action is not only wrong but also expresses something about what she is like through that action. So to take seriously

[114] Strabbing (2016): 14.

the fact that Strabbing offers her principle as a Deep Self principle, the most charitable interpretation of this principle as it applies to weak-willed action is to understand failures to exercise and act on attributability-relevant capacities as expressing implicit assent to either the omissions or to the alternate courses of action undertaken.

When the HC Principle gets applied to a particular Deep Self view, several ambiguities arise, and no interpretation of the view ends up feasible. I will illustrate these problems by applying the HC Principle to Frankfurt's endorsement view after I briefly preview the structure of the argument. The first ambiguity arises in interpreting what is meant by an "exercise of an attributability-relevant capacity." On one interpretation the "exercise" is the formation of the deep self mental state, and on another it is an agent's getting herself to act in accordance with that deep self mental state. Given the first interpretation, another ambiguity arises: is the attributionally-responsible agent the one who fails to form a deep self mental state to do something that would prevent her from acting impermissibly *if* she were to act in accordance with it? If so, the principle does not give the result that weak-willed agents are attributionally-responsible. Or is the attributionally-responsible agent the one who fails to form a deep self mental state that will actually be effective in getting her to avoid acting impermissibly? If so, I'll argue, the view mischaracterizes the character flaw revealed by the weak-willed agent. This leads an interpretation on which the attributionally-responsible agent is the agent who fails to get herself to act in accordance with her deep self mental state given that she had the ability to. But on this interpretation, I'll argue, the view cedes crucial dialectical ground and opens itself up to an especially difficult incompatibilist challenge.

Let's see what these interpretations of the HC Principle look like when applied to Frankfurt's view. Given the first interpretation, an agent "exercises her attributability-relevant capacity" by forming a second-order volition. If interpreted this way, we get the following view:

> Frankfurt-HC1: An agent is attributionally responsible for an action A if and only if 1) A results from the agent's forming a second-order volition to act on a first-order desire to A or 2) A results from the failure to form a second-order volition to act on a first-order desire to avoid doing A.

On this view, the only sorts of omissions for which a person can be attributionally-responsible are those that result from failing to form second-order volitions.

But unless we adopt a Mismatched Accounts strategy, it is natural for Frankfurt to think that weak-willed agents *do* form second-order volitions to avoid doing A but just don't act on them.[115] Sam, for example, forms a second-order volition to act on her desire to study, which, if it were successful, would lead to her avoiding doing A (in this case, going to the party), she just isn't motivated to action by it. On this interpretation there is no way to distinguish weak-willed actions from compulsive ones, and both seem non-attributable. Sam does form a second-order desire to act on a first-order desire to avoid doing A, namely, she forms a second-order desire to act on her desire to study, a way of avoiding going to the party.

[115] An alternate, but misguided interpretation of Frankfurt's "second-order volition" would have it that second-order volitions directly cause actions when unimpeded. On this view, we would still need to understand why weak willed urges don't count as impediments while compulsive urges do; this just moves the problem to a slightly different location. I call this approach misguided because, as I argue in "Depression's Threat to Self-Governance," melancholic depression's impact on the will seems best described by a disconnect between an agent's second-order volition and her action even in the absence of a countervailing first-order urge. See Gorman (Unpublished Manuscript) for further discussion both of this point and of the two competing conceptions of second-order volitions.

And so, on Frankfurt-HC$_1$, since she does not fail to form such a second-order volition, it seems Sam is not attributionally-responsible.

There is *something* that Sam fails to do, though: she fails to form a second-order volition to study that *actually* gets her to avoid going to the party. Even though the agent might believe that forming a second-order volition to study will get her to avoid going to the party, she is wrong. This leads us to the second possible interpretation. Could we instead interpret Frankfurt-HC$_1$ as delimiting weak-willed actions as the set of actions that involve an agential failure to form a second-order volition to do something that *actually* would make the agent avoid doing A, independent of her beliefs? Cases like the following one show why such a view would be unappealing:

> Secret Spinning Desire: Suppose Sam believes, incorrectly, that the only second-order volition she could form that would get her to avoid acting on her urge to go the party is the second-order volition to act on her desire to study. However, unbeknownst to her, doing so would be wholly ineffective. The only second-order volition she could form that would get her to avoid going to the party would be a second-order volition to act on a desire to spin around on her desk chair one time and then start studying. But Sam never considers acting on such a desire. Instead, she forms the second-order volition to act on a desire to study, which is ineffective, and she ends up going to the party instead.

While I think this interpretation of the view would correctly hold that Sam is attributionally-responsible for going to the party, it wholly mischaracterizes the character-flaw that weak-willed action expresses. It's not because she failed to realize that acting on a desire to spin around on her desk chair and then study that she is attributionally-responsible. This sort of epistemic flaw neither reflects the fact that an agent implicitly assents to the course of action (a crucial part of Deep Self explanations) nor does it seem to track our common-sense understanding of what it is about agents like Sam that is objectionable.

So, more charitably, I think we should interpret Strabbing's proposal as applied to Frankfurt's view in a third way:

> Frankfurt-HC$_2$: An agent is attributionally responsible for an action A if and only if 1) A results from the agent's getting herself to act on a desire to A that she second-order endorses or 2) A results from the failure to get herself to act on a first-order desire to avoid doing A that she second-order endorses.

The problem with this suggestion is that in some ways, just like weak-willed actions, compulsive actions might seem like failures to exercise and act upon general attributability-relevant capacities. Compulsive actions would thus count as attributable on this view when they are in fact non-attributable.

Instead, the view would need to be coupled with an argument that compulsive actions are not failures in the relevant sense to act in accordance with one's second-order volitions but rather, cases in which the agent lacks the ability to act in accordance with her second-order volition. This echoes a suggestion recently offered by Michael McKenna that the appropriate response to the problem of weakness of will for theorists like Frankfurt and Watson is to reintroduce talk of ability. For example, building on some textual support for this position in the article where Watson first proposed the valuing view, McKenna suggests that valuing theorists ought to adopt the following modified principle:

> We act freely just in case we are able to act in accord with what we value, as issuing from our evaluative system; we act unfreely just in case we are unable to act in accord with what we value.[116]

McKenna has in mind here the kind of freedom relevant to attributional-responsibility, and so we can derive the following principle:

> An agent is attributionally-responsible for her action just in case she *is able* to act in accord with what she values.

The idea here is that the valuing theorist could account for the fact that agents are attributionally-responsible for weak-willed actions despite the fact that such agents do not act in accordance with what they most value, because they have the (unexercised) *ability* to act in accordance with what they most value.

Something similar seems to be what Strabbing really has in mind when she writes, for example,

> But why is an action attributable to an agent just in virtue of resulting from his attributability-relevant capacity, even if that capacity is not exercised? On my view, the answer is this: when an action results from the agent's attributability-relevant capacity, he has control over the fact that he performs it. An agent clearly has control over the fact that he performs action A when A results from his exercising his attributability-relevant capacity to perform A. Yet an agent also has control over the fact that he performs A when A results from his failing to exercise his attributability-relevant capacity to avoid A, and this is precisely because he could have avoided A.[117]

[116] McKenna (Forthcoming): 5.

[117] Strabbing (2016): 22-23.

So weak-willed action on this view is self-expressive because, for example, Sam's going to the party when she possesses the relevant ability to stay home and study instead is tantamount to her implicit assent to the alternative course of action.

The fact that Strabbing invokes the notion of self-control within the context of attributability is notable, since focusing on an agent's volitional control is usually seen as a *competing* approach to a focus on attributability and the Deep Self. For example, Al Mele argues that the fact that we are not responsible for actions stemming from manias, compulsions, and addiction is better explained by the fact that we lack volitional control over such mechanisms, and so talk of the Deep Self seems unnecessary.[118] Given that it is presented as a friendly modification for Deep Self theorists, Strabbing's idea, instead, is that volitional control itself is not sufficient for responsibility, but rather, whether or not an agent exercises volitional control over her Deep Self-relevant capacities reveals something about what she is like. But it's just not clear what work the Deep Self is actually doing in this picture, or if, instead, this is really just the ability view in disguise. Ability, it seems, is really the heavy lifter on this interpretation of the account.

McKenna agrees. He argues that taking on board the notion of ability to distinguish weakness of will from compulsion amounts to giving up on the dialectical aims of Deep Self theory. As he puts it,

> Relying upon the notion of ability does not fit well with the strict aims of a [Deep] Self view and the attempt to treat compromises to free agency as *external* impediments to acting freely. If the weak willed non-addict's desire to take the drug is after all external to her, then it appears not to be the case that externality rather than internality is

[118] Mele (1992, 1995).

> what explains self-determination or lack of it. Something else is doing the work—an ability. But now, if it is merely an ability to resist a desire, what's it matter if it is internal or external to the agent? What matters is whether her ability affords her sufficient control over it to be free with respect to whether she acts on it.[119]

Dialectically, this view just cedes too much important ground to the ability theorist to really be attractive to a Deep Self theorist. In Chapter 1, §2.1, I explained how the Deep Self view's unique advantage over Classical Compatibilism is that it is able to bracket off incompatibilist concerns about the relevant kind of ability required for responsibility by shifting the conversation entirely away from talk about ability altogether. Any principle that reintroduces ability is subject to incompatibilist interpretation. On Strabbing's, weak-willed action is attributable to the agent only when she could have acted on her attributability-relevant capacity but didn't. But the incompatibilist will say that no agent can act otherwise from how she actually acts so long as she is determined. So, given the truth of determinism, no one is attributionally-responsible for any weak-willed actions. The burden is then on the compatibilist to prove that her compatibilist-friendly conception of ability is instead what differentiates failures to exercise a capacity from instances of entirely lacking the capacity. Although moving past this stalemate may not be an impossible task, it is a task that Deep Self views were designed to circumvent.

The great advantage of Deep Self views is the fact that they move beyond this stalemate by attempting to show how ability to do otherwise is irrelevant to questions of responsibility and by replacing this with another condition that is meant to be unquestionably compatibilist-friendly, namely, internality. This sort of move is leveraged by Frankfurt's arguments against the Principle of Alternative Possibilities. By analyzing internality

[119] McKenna (Forthcoming): 5.

in part in terms of ability, we lose part of the motivation for shifting to a focus on internality in the first place. And if a Deep Self theorist finds these arguments compelling, she should be wary of reintroducing talk of ability.[120]

Furthermore, some of the charges against compatibilist conceptions of ability are sharpened further by considering them in the context of a Deep Self view. In particular, the charge that compatibilist notions of ability fall short of explaining how an agent's actions/omissions are "up to her." For example, combining the view with Michael Smith's account of ability we get the view that failures to exercise capacities are cases in which an agent does not actually act on the desire she endorses when there is a raft of counterfactuals in which she *does* act in accordance with the desire she endorses.[121] This thin notion of ability is open to the following charge, which is well articulated by Pamela Hieronymi:

[120] Making the distinction normative rather than metaphysical does not help matters much. In both weak-willed and compulsive cases it is already true that the agent has, for example, endorsed her first-order desire to do the right thing. So to ask whether or not it is fair to expect the agent to exercise her attributionally-relevant capacity is tantamount to asking whether or not it is fair to expect the person to resist her desire to do the wrong thing. But attributionally-relevant capacities drop out of the explanation altogether here, as the answer to this question is just an answer about what amount of control over our desires we think is reasonable to expect of people to have, and the explanation makes no essential reference to anything about attributionally-relevant capacities. At the very least, the notion of the Deep Self would need to be quite different from how it is ordinarily understood to play any more significant role in this explanation. The boundaries of the Deep Self would need to be set by normative facts about when it is reasonable to take an act to be self-disclosing, rather than by facts that indicate whether or not the agent in fact (in at least some way) stands behind her action. See Chapter 5, §4.2 for further discussion of this point.

[121] See Smith (2003) for a development of this view.

> Suppose I have a heart attack. It may well be that, in a host of similar possible worlds, I do not have a heart attack. Further, the fact that I do not have a heart attack in those worlds may be explained by the underlying structure of my cardio-vascular system. Thus, I have the capacity, in Smith's sense, to have not suffered the heart attack. The truth of this claim does nothing to show that it was up to me whether I had a heart attack…[122]

In the context of a Deep Self view, this charge is particularly troublesome. If it's not up to you, nor does it have anything to do with what you really want, whether or not you act on the desire you endorse, how can whether or not you act on the desire you endorse determine whether or not your action is self-disclosing?

Even setting aside all of the dialectical problems with the reintroduction of ability, it's just not clear that the agent does assent to her action by failing to bring herself to act in accordance with her values/ endorsements/ commitments when she has the ability to. In order for this to be true, we would need an argument that this minimal kind of assent is the relevant kind of assent. One could make an argument that this notion of assent seems too minimal to do the justificatory work it's meant to do. Instead, my criticism comes from a slightly different angle: if we need to make the notion of assent much weaker than it's ordinarily understood in traditional Deep Self theories, we may as well just identify the kind of mental state that constitutes this form of assent, and make that mental state itself the criterion for attributability.

4. Adopting a Mosaic Conception of the Deep Self

The only truly feasible strategy for a Deep Self theorist to respond to the weakness of will problem is to alter the way she conceives of deep self

[122] Hieronymi (2007): 16.

mental states such that it makes sense to say that weak-willed agents *do* identify with the actions they undertake in a way that makes these actions self-disclosing. In order to understand the way in which an agent can assent to a course of action that is counter to what she takes to be her most favorable course of action, the traditional Deep Self theorist will have to alter her conception to allow that both the favored option and the weakly-willed option would, if undertaken, have the power to speak for the agent. So, for example, both Sam's desire to stay home and study, were she to act on it, and her desire to go to the party would both count as being part of her Deep Self. The idea behind this approach is to in some way weaken the pre-requisites for agential self-expression so that an agent *does* express something about who she is directly via her weak-willed action.

Chandra Sripada calls the view of the Deep Self that corresponds to this alternate picture on which more than one mental state and their resultant actions can speak for the agent a "mosaic" conception. Mosaic conceptions of the Deep Self permit conflicts and tensions within the Deep Self since, as Sripada writes, "conflict can and often does extend all the way to our very practical foundations."[123] He contrasts this with "homogenous" conceptions of the Deep Self, according to which the Deep Self can contain no conflicts, and all apparent conflicts are merely illusory. On such views, for an action to issue from a person's Deep Self is for the agent to authorize an action, and an agent can only authorize an action by uniquely and decisively picking it out in some way. I have shown that most traditional homogenous Deep Self theories are unable to accommodate weak-willed actions as attributable actions because weak-willed actions are caused by motivations that conflict with the motivations the agent has authorized herself to act on. This suggests that the way to solve the weakness of will problem for a Deep Self theorist is to adopt a mosaic

[123] Sripada (2016): 24.

picture of the Deep Self and correspondingly, a weakened criterion for attributability.

4.1. The Minimal Approval View as a Mosaic Deep Self View that Can Solve the Weakness of Will Problem

Adopting the Minimal Approval account solves the weakness of will problem because it explains the way in which even the weak-willed agent may assent to her course of action: *via* minimally approving of her course of action. It is able to solve the problem precisely because it is a mosaic view: more than one course of action can be assented to since conflict is permitted at the agential foundations that are relevant for responsibility-ascriptions.

Recall that the Minimal Approval account just requires the following for attributional-responsibility:

> The Minimal Approval View of Attributional-Responsibility: An agent is attributionally-responsible for φing-at-*t* iff the actual sequence of mental states involved in the production of her action is non-implanted and together with her other mental states makes it the case that at *t*, if she were to reflect on her desire to φ at *t*, she would be sufficiently likely to want to act on her desire to φ at *t*, with some further aim in doing so other than merely eliminating this desire.[124]

The account accommodates the fact that weak-willed agents have less-than-complete identification with the courses of action they undertake, and are nevertheless attributionally-responsible for them. For example, it makes sense to think that Murderous Max's action comes about in part

[124] "φ-ing" should be understood here as standing for an action, and the account should be taken to cover only attributional-responsibility for *actions*. Attributional-responsibility for omissions and for consequences is, on my view, derivative on attributional-responsibility for actions. I explore some of these issues in Chapter 5.

due to the fact that he wants to some degree to act on his desire to go on a killing spree, and that this desire is not just to relieve some urge but is rather related to his love of the art of murder. Therefore, his action, though weak-willed, is clearly attributable.

It might be helpful at this point to see how the Minimal Approval view comes apart from traditional ability-based views in terms of what it posits about the difference between weakness and compulsion. Now it may be, as a matter of empirical fact, that it is often harder to get oneself not to act on a desire that has no correspondence to what you endorse, making it such that you have less control over acting on such desires, and so there may be fairly substantial overlap in the extensions of the two theories. The difference is that, on my view, the difficulty one has in resisting a piece of behavior is not essential to what makes something non-attributable, nor to what makes something a compulsion. An agent may be such that she could easily have avoided the compulsive behavior if she tried, but simply due to absent-mindedness forgot to. She would still not be responsible for her compulsive action because it does not express anything about who she is. On the other hand, Sam, whether she has the capacity to resist her weak-willed desire to go to the party or not, still does something she is attributionally-responsible for because her doing so is related to the fact that she would want to act on this desire due to some aim, and this tells us something about what Sam is like. I think is precisely what a Deep Self theorist should want to say about the cases, not only because it is extensionally adequate, but also because it gives the right *kind* of explanation for why weak-willed but not compulsive agents are attributionally-responsible.

4.2 Can Other Deep Self Views Adopt a Mosaic Conception to Solve the Weakness of Will Problem?

The reason the Minimal Approval view is able to solve the weakness of will problem while Frankfurt's endorsement view is not is that it does not require a homogenous conception of the Deep Self and so can offer a weaker criterion for attributability. This raises the question: can any other Deep Self view follow suit and adopt a criterion that is compatible with a mosaic notion of the Deep Self in order to solve the weakness of will problem?

Not every Deep Self makes sense in mosaic form. Bratman's planning view, for example, seems inextricably bound up in a homogenous conception of the Deep Self since the agent's deciding decisively on a certain unique course of action to prioritize is meant to be precisely what authorizes attributional-responsibility for that action. Bratman emphasizes the role of cementing agential coherence in planning agency, and adopting a mosaic conception of planning would certainly forgo these appeals of the view. However, even if a version of the planning view did allow for an agent to make conflicting plans, it is not feasible to think that in all attributable weak-willed cases the agent must have a plan to act in the way she does, nor need her action be related to any of her larger plans. In fact it's quite a natural thought that part of the nature of weak willed actions is the very fact that they deviate entirely from our plans for ourselves.

In Chapter 2, I argued that each traditional Deep Self view proposes a criterion that contains additional elements that are unnecessary for attributability. Mosaic versions of valuing and caring views, I'll now argue, are subject to a parallel criticism in their handling of weakness of will cases. In both cases weakening the criterion to accommodate a mosaic Deep Self conception does not weaken it enough to capture all cases of attributable weak-willed action.

Strabbing considers and then, for this very reason, dismisses a mosaic version of the valuing view on which rather than requiring that agents act on what they *most* value in order to be attributionally-responsible for their actions, agents just need to act in accordance with something they (*pro tanto)* value at all. So according to this view, in Sam's case, we could say that while she values studying for her exam the most, she does value going to the party as well to some degree. But there are cases, as Strabbing points out, in which a weak-willed agent does not even *pro tanto* value the course of action she takes, yet she still seems attributionally-responsible for her action.[125]

Watson's own examples of a mother overcome with a sudden urge to drown her bawling baby in the bathtub, and a squash player with a desire to smash his opponent in the face with his racquet can be used to illustrate here. [126] There are cases in which a person desires something without valuing it at all, yet we appropriately take a person's acting on such a desire to say something about what she is like. Agents who act on *weak-willed* desires to do such things provide counterexamples to the claim that weakening the criterion of the valuing view in this way can solve the weakness of will problem.

[125] Another problem with this view might be that it seems in other respects too expansive—it gives the result that agents who are intuitively compulsive who nevertheless see *something* of value in taking a drug, etc. are in fact attributionally-responsible. This point is not crucial to my argument in this chapter since there is, I take it, already sufficient reason to reject this view. But it is worth mentioning since a similar though somewhat less serious worry exists in regards to the Minimal Approval account. Consider a case in which an alcoholic desires to some small degree to act on her desire to drink, not just because she wants to get rid of the desire to drink, where her second-order desire is not explained by any sort of special epistemic circumstances, and where acting on a desire to drink is morally problematic. My view may seem unreasonably austere in classifying such agents as attributionally-responsible and blameworthy. While I won't fully develop it here, if this result is unpalatable, one possible route is to say that attributability is gradable such that the alcoholic's action is *less* attributable than the action of an agent who wanted most to satisfy her desire to drink (where such desire to satisfy is not explained by wanting to rid herself of a desire to drink. (In theory, a mosaic value theorist could make some sort of analogous move to circumvent this issue.) A natural idea might be that the gradability comes from the strength of the desire to drink relative to her other desires to some degree satisfy first-order desires. I hope to explore this issue further in future work.

[126] Watson (1987).

This is an application of the charge against the valuing view I made in Chapter 2. It is just a fact that we are attributionally-responsible for some things that we don't value at all, and there are many weak-willed cases that illustrate this lesson particularly well. As Watson himself now acknowledges, the valuing account is just too rationalistic; this is a problem for accounting for the fact that even some non-weak-willed actions are attributable in the absence of valuing.[127] In such cases an agent *wholeheartedly* decides to do something that she does not judge to be good, and intuitively is attributionally-responsible for her action.

The caring view is a better candidate for solving the weakness of will problem, and in fact, it is usually advanced in an explicitly mosaic form, as a view that is therefore well-positioned to handle cases of weakness of will.[128] As with the valuing view, though, a somewhat parallel criticism to the one I advanced against caring views in Chapter 2 extends to the view's treatment of weak-willed cases.

The issue is what Strabbing calls the problem of weak-willed whims.[129] The idea is that oftentimes the desire an agent acts on in a weakness of will case is a whimsical desire that intuitively is not related to what the agent cares about. Nevertheless, intuitively, the agent is responsible. Certainly some cases that we are tempted to describe as weak-willed whims may truly be instances of very minor compulsions, and the line here may not always be clear. Whereas attributional-responsibility for acting on whims may seem inconsequential in many cases, in weak-willed cases acting on these whims is in many cases what makes the agent fail to do what she believes she ought to do. Supposing she is right about what she ought to do, if her responsibility for failing to do what she ought to do is derivative on her responsibility for acting as she in fact does, then it matters very

[127] Watson (1987).

[128] See, for example, Shoemaker (2003), and Sripada (2016).

[129] Strabbing (2016): 16.

much whether or not she is attributionally-responsible for her whimsically-motivated action.[130]

Suppose Sam has decided to stay home and study for her exam and is overtaken by a sudden whim to pace around her dorm room instead. It seems strained to say that Sam cares about pacing around the room or that her desire to do so would be suitably related to any sorts of distinctive caring states she has. But it is likely that Sam's desire to pace around the room is not just a random fluke either. In fact, it may be that whimsical desires always stem from subconscious intrinsic desires. Arpaly and Schroeder argue that "in each case [involving someone acting on a whim] it is easy enough to imagine credible intrinsic desires that each person might have such that the person's whim is instrumental toward, or a realizer of, the content of the intrinsic desires." It seems plausible that there will always be similar stories to tell about whimsical desires even if the agents themselves don't always have access to the explanations.[131]

If whimsical desires are always instrumental or realizer desires of subconscious intrinsic desires, it is easy enough to see how a weak-willed agent who acts on a whimsical desire would satisfy the requirements for Minimal Approval for these actions. The fact that she would act on her whimsical desire is related to the fact that if she were to reflect she would have a desire (not necessarily consciously held) to act on a motivational state that would further her intrinsic desire (her further aim).

But dialectically speaking, care theorists cannot appeal to the presence of a mere intrinsic desire to show that the agent is attributionally-responsible, because they are at pains to show that caring states are not reducible to mere intrinsic desires, but rather, involve a complex set of

[130] I think it makes the most sense for mosaic Deep Self theorists to consider responsibility the actions that are omitted in weak-willed cases as derivative from responsibility for acting on the motivation the agent in fact acts on. I develop this idea more fully idea in Chapter 5.

[131] Arpaly and Schroeder (2013): 10.

dispositions. Chandra Sripada does take caring states to be partially constituted by intrinsic desires, and so he may tell a similar story to show how the presence of these intrinsic desires is *evidence* that they agents whimsical desires are suitably related to her caring states. However, it is just not clear that the sorts of intrinsic desires for which whimsical desires are instrumental towards or realizer desires of need always be related to what she cares about in Sripada's sense.

It is plausible that Sam's pacing around the room is caused by a subconscious fear of failure that is suitably related to the fact that she cares about being a good student. But it is equally plausible that her pacing is due to a spontaneous and fleeting intrinsic desire to not think so hard; she is moved to act on a desire that realizes an intrinsic desire she has no long-term or emotional investment in whatsoever. This echoes the objections to the caring view I advanced in Chapter 2.

It is interesting to note that David Shoemaker, who also advances a care theory, seems to openly embrace the fact that according to his view agents are not attributionally-responsible for acting on whims.[132] The costs of this move are mitigated by the fact that, according to his picture, agents who act on whimsical desires may be candidates for answerability-responsibility and/or accountability-responsibility. These forms of responsibility, for Shoemaker, are serious forms of moral responsibility that do not require agents' actions to be attributable (at least not in the kind of sense required by attributional-responsibility). For care theorists who take attributability to be necessary for the most central kinds of moral responsibility, however, denying that agents who act on weak-willed whims are attributionally-responsible comes at a much higher cost.

This shows that there is a distinctive advantage to adopting the Minimal Approval View. While the main innovation in responding to the weakness of will problem is to lower the criterion for attributability such

[132] See Shoemaker (2015b): 113.

that conflicted strands of the self can both speak for the agent, the Minimal Approval view is the most promising implementation of this strategy.

5. Responses to Fairness Objections

Despite the advantages of the Minimal Approval view's handling of the weakness of will problem, it does have some features that may seem unintuitive due to worries about the fairness of demarcating compulsion and weakness in the way that I have suggested we should. In this section my aim is to respond to these concerns.

First, some may object that on my view we let people off too easily for their compulsions. I am committed to the claim that it can be the case that a person is not attributionally-responsible for her action even if she could have done otherwise had she resisted. On my view, compulsions in theory need not even be particularly difficult to resist to make a person exempt from attributional-responsibility for the resultant actions. To the extent that this strikes some people as implausible, I take it this is motivated by a concern about fairness.

I have several lines of response to such worries. First, there is much contested ground over which questions about free will and moral responsibility should be answered in the domain of metaphysics and which should be answered in the domain of first-order ethical theory, but I take there to be a methodological problem with raising concerns about fairness at least with *this* part of the picture. This is not to say that questions of fairness need not enter consideration at all, just that if they do, they do so at a different level of generality. We might reasonably ask the question, *given considerations of fairness, what is the appropriate basis on which to judge people to be attributionally-responsible for their actions?* Once we have decided that the answer to *that* question is that it is appropriate to hold people attributionally-responsible for their actions just in case they reveal something about what they are like as agents, and commit to a Deep Self view, our further question is now: *what does it take for someone to reveal something* *about what she is like as an agent through her action?* Our criterion for answering that question should be based on how well the proposed state or process corresponds to the proper notion of self-disclosure that is relevant to praise and blame. Reintroducing concerns about fairness in answering *this* question seems to admit a certain kind of defeat for the metaphysical inquiry of the B-Tradition. We sought to discover the conditions that tell us when an act *actually is* self-disclosing, not just when it would be best for us to think of someone's act as self-disclosing given the consequences of doing so as they bare on considerations of fairness.

Setting aside these methodological concerns, it is not even clear that considerations of fairness would make us favor an ability-based theory of when people are exempt from attributional-responsibility for compulsive action over the view I have put forward.

There are at least two different concerns about fairness that might be raised in regards to my proposal. First, one might think the victims of moral transgressions that are caused by compulsive agents who could have easily resisted their compulsive actions have a right to blame the people who wronged them. Moral wrongdoing could have been prevented easily, and so, given considerations of fairness, blame seems warranted.[133] Second, given that on my view agents are attributionally-responsible for weak-willed actions that are very difficult to resist, it seems unfair to exempt similarly situated compulsive agents who would not have the same difficulty resisting were they to try. One thing to note is that the view is a view about attributional-responsibility, not just blameworthiness, so presumably the inverse fairness concerns could equally be raised for praiseworthiness. These concerns might be thought in a way to balance

[133] I actually agree that such victims might deserve an apology, but do not think they have the right to blame compulsive agents. I develop a brief sketch of a sense of responsibility that might make the former but not the latter response appropriate in Chapter 5, §5.

each other out. But given that susceptibility to blame may be more burdensome than susceptibility to praise is a benefit, noting this may do little to quell worries.

But isn't it also unfair that compulsive agents are expected to shoulder the burden of resisting external urges? If compulsive urges are in fact external, why should they be the responsibility of the agent to have to manage them? One could try to somehow balance this consideration of fairness against those raised against the proposal, but I think to make fairness the sole criterion for the theory would be to make a methodological mistake. It would amount to reducing blame and responsibility practices to mere burdens rather than acknowledging them as practices that are inextricably bound up in a context that makes such practices apt.

If remaining concerns linger, the following is at least dialectically open to the Minimal Approval theorist. She may admit that even if concerns about fairness do not influence ascriptions of attributability, they may inform the background conditions of the agent's action such that they influence blaming practices, thus affecting the way in which it is appropriate to interpret and respond to the meaning of the agent's action. The Minimal Approval theorist may even accept that while it plays no role in assignment of attributional-responsibility, difficulty resisting countervailing motives may have a role to play in distinguishing between whether or not, for example, an agent reveals through her action an overtly malicious trait or a mere lack of moral fortitude. In this way the Minimal Approval theorist can accommodate the datum that degrees of difficulty resisting does seem to play some role in our assessment of agent blameworthiness without ceding the crucial point that control plays no role in setting the bounds of attributability.

Chapter 4: *The Finely Individuated Trait View of Blame's Content*

1. Introduction

My central goal in this dissertation so far has been to advance an account of the conditions of attributability, the Minimal Approval view. I have argued in favor of an account that locates attributability in a set of agential conditions, and accordingly, my focus has primarily been on the attitudes of the responsible party. In this chapter I illuminate some of the core features of blame itself, and so my focus is instead on the attitudes of the blamer, rather than the person blamed.

Recall that the view I articulate in the first three chapters is meant to be, in theory, compatible with a wide range of stories about the route from attributability to blameworthiness and blame, and about the relationship between attributional-responsibility and other kinds/faces of moral responsibility. In this chapter I will develop one possible story about the relationship between attributability and blame that I think ought to be adopted, and in Chapter 5 I will say more about what I take the relationship between attributional-responsibility and accountability-responsibility to be. But it should be noted that Chapters 1-3 are, in a way, self-standing. One may accept the claims of the first three chapters without also accepting the claims I will make in this chapter or the next.

That said, an accompanying account of blame can make a particular account of its preconditions more or less plausible. In particular, the Minimal Approval view leaves us with a view on which the "self" that is implicated in attributional-responsibility may be both arational and fragmentary. This is a less robust conception of the self or will than might be required for certain accounts of blame. Does the possibly arational and fragmentary nature of the self that accompanies the Minimal Approval picture threaten our ability to tell a compelling story about blame? My main goal in this chapter will be to argue that it does not.

Developing a complete theory of the nature of blame would require providing answers to two different questions. First, there is the question of what kinds of mental states blame consists in. Candidate answers to this question in the current literature include judgments, emotions, desires, and dispositions to communicate or protest. Second, there is the question of what the content of these blaming attitudes are. For example, suppose blame amounts to targeted resentment: does B blaming A mean just that B resents *the fact that A φ-ed*? The judgment that A's φ-ing reasonably implicates? A's quality of will as exhibited by her φ-ing? A's character as exhibited by her φ-ing?

These two questions may not be wholly independent, but since I will not be able to offer a complete account of the nature of blame in this short chapter, I focus primarily on the latter question, which speaks more directly to the challenges that arise for articulating a theory of blame to go along with the Minimal Approval view. My aim is to sketch a theory of blame's content that is attractive in its own right, but also one which is particularly well suited to be accepted in conjunction with the Minimal Approval view.

2. From Attributability to Blameworthiness

Once it is determined that an agent is attributionally-responsible for φ-ing, what else must be added to this fact in order for it to be appropriate to blame the agent on the basis of her φ-ing? If the kind of blame in question is moral blame, then one thing that must be added is that the agent does something that is in some sense morally objectionable. I take the relevant sense of an agent doing something morally objectionable to be that she does something that she morally ought not to have done relative to her non-moral beliefs at the time of action.[134] Although it is controversial, I

[134] I say her "non-moral beliefs" because I believe that unlike other sorts of ignorance, moral ignorance does not exempt, but I will not here be able to give a defense of the fact

take it that this way of conceiving of the sense in which a blameworthy agent does something morally wrong avoids the need to posit an additional epistemic requirement on blameworthiness[135] or an additional requirement to rule out cases in which an agent acts under severe duress. It is not that these factors are irrelevant to an agent's blameworthiness, but rather that they are factors that play a role in normative ethical theorizing about wrongdoing itself. There is no need to offer additional criteria here beyond the fact that the agent attributably does something morally objectionable for blameworthiness because I take it that it is just not morally wrong in the sense relevant to blameworthiness to tell a lie due to your child being threatened, or to give someone a glass of poison when you believe that it is gin.[136]

A proponent of the Minimal Approval view need not be beholden to this picture of the relationship between moral wrongness and blameworthiness, however; there are interesting and largely underexplored issues regarding which of these issues ought to fall under the domain of theorizing about moral wrongness and which should fall under the domain of

that moral ignorance is special in this way. For recent discussion of this issue in the literature see the papers collected in Robichaud and Wielend (2017).

[135] A belief-relative notion of moral wrongness subsumes several kinds of proposed epistemic conditions on blameworthiness including but not limited to Susan Wolf's "sanity condition." See Wolf (1987).

[136] Many think we need to leave room in the picture for the possibility of morally wrong action committed by agents who are excused from blameworthiness due to duress in order to explain intuitions about cases of, say, murder or assault under duress. Although I won't argue for it here, I suspect these intuitions can be explained instead by a combination of uncertainty about the degree of duress that makes typically wrong actions become permissible, and sufficient attention to the importance of the gradability of moral wrongness.

theorizing about blameworthiness.[137] While these are interesting issues in their own right, I will largely bracket them here.

Aside from the fact that the agent does something morally objectionable, my proposal is that there are no additional requirements for being an appropriate target of blame over and above the fact that the agent's action is attributable to her. She need not act with ill-will, she need not possess any (further) kind of free will, nor need she possess any special sort of communicative or normative-competence.

3. Robust Traits and Implicit Judgments: Between Scylla and Charybdis

But blame is not merely the recognition that an attributable moral wrongdoing has occurred. It is important to respect the fact that, as George Sher points out, it is central to our conception of blame that blame is fundamentally "a reaction *to a person on the basis of the wrongness of what he has done*" in which we take "wrong acts to…reflect badly on the agents who perform them"(7). Even a moral responsibility skeptic might allow that there are morally wrong acts and that we are justified in reacting to the fact that

[137] To give just one example, Taylor (2003) argues that hierarchical accounts of autonomous agency ought to be rejected given that they generally fail to account for cases of severe duress since they wrongly predict that agents are still autonomous when they act under duress. The Minimal Approval view does in a certain sense propose conditions for "autonomous agency" but the sense of autonomy relevant to attributional-responsibility is taken to be rather minimal. (There is a good case to be made that the fact that acting under duress is not compatible with autonomy in its various stronger senses invoked outside of discussions of responsibility, including medical and political contexts.) While I take it that it is important that the complete specification of blameworthiness make room somewhere to explain the fact that agents are not blameworthy in cases of severe duress, it is not obvious to me that this must be done via showing how duress undercuts attributable agency, given just how minimal the sense of attributable agency required for the Minimal Approval view is. But views that posit that less minimal agential conditions and thus stronger senses of autonomy are required for attributability may seem more objectionable if they relegate exemptions for duress to the normative domain instead of explaining them via their accounts of autonomous agency.

such actions have occurred. But this seems to fall short of genuine blame, which goes above and beyond a mere judgment that someone has committed a wrongdoing or the admonishment of such an act. Dispassionately telling a murderer that what she did was wrong and that she ought not do so again in the future falls short of blaming her on a fundamental level. When we blame her we, in some important respect, react to *her* on the basis of her action. The central question an account of blame needs to answer, therefore, is what blaming adds over and above a judgment that someone has acted wrongly that somehow relates her wrongdoing to a reaction to the wrongdoer herself.

However, two straightforward ways of implicating the agent herself in the content of blaming attitudes: via robust traits or implicit judgments, are not available to proponents of the Minimal Approval view.

According to various traditional Deep Self views, an agent's act is attributable iff it is caused by a mental state that has an especially tight connection to an agent's practical standpoint or character. For example, Michael Bratman posits that agents' actions must align with their planning states, and their planning states, when taken all together, jointly constitute an agent's diachronic practical identity. On certain readings of Frankfurt's theory, such as on David Velleman's interpretation, second-order volitions play the role of being functionally identical to the agent herself such that blaming attitudes directed at the initiation of an attributable action *just are* blaming attitudes directed at the agent herself. Blame's sting, on such views, comes from the fact that one's attributable acts express one's deepest commitments and so criticism of an agent's attributable action impugns the core of her being. Sometimes this idea is coupled with the idea that an agent's diachronic commitments or values make up her character traits, and so when we blame an agent due to her action, we are really blaming her for having certain morally problematic traits, which are expressed through her action. Call this the Robust Trait view.

But given a simple correlation between attributability and attributional-blame, the Minimal Approval theorist cannot avail herself of the Robust Trait view. If an agent's act is attributable, according to the Minimal Approval theorist, all we know is that *some* part of her self stands behind it, not that who she most deeply is stands behind it. Given the way the Minimal Approval view handles weakness of will, it could be possible for a person to be generally kind, and even to most strongly endorse doing the kind thing in every scenario, but still be blameworthy for acting unkindly in a one-off weak-willed scenario. For her to blameworthy, doing that unkind thing must be something she approved of to some minimal degree, but unkindness needn't be a part of any larger or more defining feature of her will.

On another view, one that is meant to be compatible with more minimal conceptions of attributability, the content of blame is the not the character of the agent, but rather, the *meaning* of the agent's action, which is in part a function of the agent's position with regard to the person doing the blaming. As T. M. Scanlon puts it, the meaning of an action for a person is "the significance that person has reason to assign to it, given the reasons for which it was performed and the person's relation to the agent."[138] On Angela Smith's version of the view, which she calls the Rational Relations view, the content of blame is the judgment of the agent taken to be implicit (by the blamer) in her so acting.[139] It is because Smith takes it that attributable actions can reasonably be taken to reveal the judgment of the agent that agents are answerable, or can be called upon to provide justification for their actions. But according to the Minimal Approval view, agents do not need to take themselves to have normative reasons to perform the attributable actions they perform; they may approve of them for

[138] Scanlon (2008): 54.

[139] Smith (2005): 17.

no reason at all. So it would be unreasonable for blamers to assume that agents take their attributable actions to be justifiable.

4. The Ledger View

Given that the Minimal Approval view neither requires attributable actions to be expressions of robust character traits nor claims that attributable actions make it the case that others can reasonably assume that they reflects in some way on the judgments of the agent, there is reason to adopt a view on which the contents of blame are taken to be much more minimal. One candidate view is the Ledger View of blame. Ledger views hold that what blame adds to the judgment that someone has done something wrong is that the wrong act itself adds a 'negative mark' to the wrongdoer's 'moral record.' The content of blame then, is the wrongdoing itself *as it bears on the blamed person's overall record.* A theory of attributability, then, would give the conditions for when an agent's action does/ does not reflect on her moral record.

The insight of the Ledger view that is worth preserving is that it is the fact that the agent (attributably) committed the wrongdoing itself that reflects poorly on the person blamed rather than the fact that the agent's traits more generally align with the propensity to commit similar wrongdoings. What should it matter to the victim of a horrific crime if the perpetrator was acting in a way that, while she endorsed it in the moment, does not reflect her more general character? The Minimal Approval view can explain why the fact that an agent acts "out of character" sometimes seems to explain why she is exempt from blame due to the fact that acts that she does not minimally approve of are oftentimes out of character. But it is not some further requirement on blame that actions need be related to past or future traits or dispositions. The Ledger view's focus on the attributable act's mark against the person's character itself better reflects this.

Nevertheless, the Ledger view, in the ways it is usually defended, suffers from a couple of serious problems. First, in focusing on a person's overall 'score' it overemphasizes the degree to which the blamed person's wrongdoing is blameworthy due to diminishing her moral standing in general. When a person with a generally exemplary history of moral behavior commits a wrongdoing, she is still blameworthy, though her scorecard may still be much better than average. The Ledger view leaves us with unanswered questions about why such people should be blameworthy for failing to achieve overall moral perfection.

Second, our practices of holding one another responsible as they actually exist often seem quite removed from the practice of moral grading and accounting posited by the Ledger view. As a result, this sort of view can run the risk of distorting our interpersonally engaged social practices to make it seem, as Gary Watson puts it, "as though in blaming we were mainly moral clerks, recording moral faults... from a detached and austerely 'objective' standpoint."[140] Furthermore, even if this did provide an apt description of our *actual* processes, I doubt that a system of demerits and point-scoring could really be the institution that many of us seek to so fiercely defend.

A suitable account of blame's content to accompany the Minimal Approval view ought to borrow from the Ledger view the idea that an agent may be blameworthy on the basis of evidence from an individual wrongdoing without it necessarily revealing a larger character flaw, while aiming to avoid these pitfalls of the view.

5. The Finely Individuated Trait View

5.1 A Paradigm Shift for Thinking About Aretaic Traits

In order to avoid the unappealing consequence of thinking in terms of a person's overall 'score' being the focus of appropriate blame directed at a

[140] Watson (2004): 226–227.

person, it will be helpful to reconsider the relevance of aretaic traits. On the orthodox view of aretaic traits, traits are robust character dispositions that influence agents across a variety of circumstances. However, recent evidence from social psychology threatens to show that robust traits as philosophers have often conceived of them simply do not exist. The Situationist Challenge, as it's been called, argues that our morally relevant traits are highly contextual and influenced by morally trivial situational factors. Since a single agent usually has evaluatively inconsistent dispositions triggered by these various contextual factors, it is very rarely apt to ascribe traits like "viciousness" or "kindness" to agents.[141]

To take a commonly cited study, participants were 84% more likely to help a woman pick up her papers if they found a dime in a phone-booth just prior to the papers scattering. One natural conclusion, given a wealth of similar data across other studies, is that it's the dime finding and not the good will that leads participants to help the woman.[142] A possible response to this concern is to hold that we ought not utilize the concept of traits at all in our moral practices. But does it mean that we ought not blame a study participant for failing to help the woman (supposing it is morally wrong not to help in the scenario)? It seems that we still can coherently blame the participant because she does reveal something about her character; she reveals that she is the kind of person who wouldn't help a woman whose papers are scattering, at least in the case in which she has not first found this dime. Nothing in the Situationist critique tells against the fact that it is appropriate to have a blaming response to the fact that someone attributively acts unkindly in a particular situation. Even if an attributable morally wrong action is aptly described as being influenced by

[141] See Doris (2002, 2015).

[142] Although see Earp and Trafimow (2015) for the alleged replication crisis for social psychology, which may cast doubt on some of the data used to bolster criticism of the existence of robust character traits.

a morally trivial situational factor, and is thus not correctly described as stemming from a more general trait like cruelty or dishonesty, that doesn't mean we shouldn't respond to the more finely individuated trait that is evidenced just from the agent's attributably performing that action alone.

In order to say that we blame agents for their traits in these sorts of scenarios, we would need to consider a paradigm shift in thinking about the metaphysics of aretaic traits. Ordinarily, more robust traits are taken to be fundamental and to consist in dispositions to perform token actions of a certain type. For example, viciousness is taken to consist in a collection of dispositions to, say, steal candy from a baby, ruthlessly punch someone, con someone out of out of money, etc. Instead, I think we should think of the paradigm examples of traits as being more finely individuated, such as: *being the kind of person who would on at least one particular occasion steal candy from a baby*. Arguably, we can still appropriately use terms like "vicious" to point to patterns of more finely individuated traits that are similar or co-occur due to a common cause, and we can do so without countenancing the existence of viciousness as having an independent existence or as having the power to shape behavior more globally across an agent's psychology.

The orthodox way of thinking about aretaic traits obscures the fact that we do learn something about an agent's moral character when she attributably commits a moral wrongdoing, even on a view like the Minimal Approval view with very minimal conditions for attributability. Namely, we learn that she is the kind of person who would attributably φ, where φ-ing is morally wrong, in the kind of circumstances in which she in fact φs. The proper content of blaming attitudes, I want to suggest, are these very finely individuated aretaic traits, individuated roughly as finely as actions themselves. While this may seem too fine to individuate traits compared to the way they are often conceived of in the philosophical literature, I believe we do in fact blame people on the basis of quite finely individuated traits all the time. For example, consider the following paradigmatic blaming statements: “I can’t believe you are the kind of person who would do that!” or “I didn’t think, when I first started dating her, that she was the kind of woman who would *ever* say something like that to me!” While these statements are focused on the character of the person blamed, their focus is quite narrow. The man who blames his wife for cheating on him may have an interest in whether or not her act is part of a more general propensity to be unfaithful, but it would be absurd to insist that he should relinquish his blame entirely if he were to be presented with conclusive evidence that it was a one-time thing.

In individuating traits this finely it might seem that the view runs the risk of collapsing into the view that blame’s content is just the morally wrong action itself. But if we allow that blaming involves attitudes beyond mere judgments about or that something has occurred, the role of finely individuated traits becomes less obscure.

For example, on George Sher’s account, blame centrally involves a desire not only for the bad action to not have occurred, but also for the agent to have *been* different, although the evidence that the agent has a trait that we blame her for is fully supplied by her attributable act itself. In blaming it seems we want not just for the horrible thing to not have been said, but for the agent to not have had something in her psychology that she would countenance at all that would lead to her saying such a thing. As Sher puts it, blame, in its most characteristic form, often seems bound up in the frustration of a desire that leads “not to the generalized frustration that we feel when we get stuck in traffic or botch a plumbing repair, but rather to bad feelings that are directed specifically at the wrongdoer or bad person himself.” To see why this is not surprising, Sher continues,

> we need only remind ourselves of the peculiarly close connection between that person and what is wanted. When we have an unsatisfiable desire to scape a traffic jam or fix a broken drain, we may indeed want other people to act in certain ways…but we want this only because it

> would produce a father result that does not essentially involve them. By contrast, when we have an unsatisfiable desire that someone…not have a bad character, our desire is directed at that person not merely in the superficial sense that we want something that he could bring about, nor yet in the somewhat deeper sense that we want something that we cannot fully describe without mentioning him, but rather in the deepest sense that we want him to have exercised his own decision-making capacities in a certain way.[143]

This kind of view makes sense of the fact that something about the agent herself is the content of blaming attitudes, rather than the act itself.

Consider also a view on which some sort of negative behavioral response to an agent is made appropriate due to the fact that the agent is blameworthy. For example, blame may consist in part in the blamer's altered patterns of attention towards the blamed person and/or the revocation of charitable interpretations of the blamed person's behavior more generally. The licensing of these sorts of responses seems to be a feature of the blamed agent revealing something about what she is like through her action rather than of wrong actions themselves.

5.2 Why Blame Can Sometimes be Emotional and Impair Relationships

Recall that one criticism of the Ledger view is that it seems to advance a picture of blame as a detached objective assignment of a demerit, and this threatens to distort the emotional and interpersonal dimensions of blame. While simply attending to the phenomenon of blame in real life gives us plenty of evidence that blaming someone can be infused with emotion and can play some role in modifying or even ending relationships, there is a further question as to whether either of these features might be essential to blame. But even if neither of these aspects plays an essential role in a theo-

[143] Sher (2005): 105.

ry of blame, there ought to be a coherent story to tell about why these features are often present, even if they are secondary to blame itself. The Finely Individuated Trait view is well suited to explain these connections.

Some philosophers who aim to give an account of blame solely in terms of fitting reactive emotions draw a sharp distinction between agent-focused anger or resentment on the one hand and generalized frustration on the other. According to these views blameworthy agents are defined by the fact that they are the appropriate targets of fitting resentment. But notice that the Finely Individuated Trait view is well-positioned to both help distinguish between frustration and resentment as well as explain why resentment may often be warranted when an agent is blameworthy. Anger in its most general form tends to be a reaction to the threat of something taken to be valuable. It is plausible that people take it to be valuable that people who are important to them (and perhaps people in general) not behave in certain ways. When a person attributably acts in one of these disvalued ways, she is in a way both the person who threatened the thing taken to be valuable as well as, in a way, the lost value itself. By locating what is blameworthy as an aspect of the blamed agent's self in some sense, we can better see why agential anger, or resentment, is often present in an episode of blaming rather than mere frustration.

In many cases these sorts of valued social norms that prohibit people from acting in certain ways are part of the implicit contracts that make up the bounds of our interpersonal relationships. I think we should accept the Strawsonian idea that we attach great importance to what the actions of others reveal about their attitudes towards us. We all have normative expectations of one another that sustain our social practices, it would seem. We rely on the fact that most of our friends are not the kind of people who will intentionally harm us, our wives will not laugh in our faces, and our students will not throw paper airplanes at us while we're lecturing.

In some but not all of these cases, our relationships are conditional on not seeing a person in such a way and, in revealing what they are like, the person being blamed impairs her relationship with those affected. The frustration of the affected persons' expectations may be grounds for withdrawal from the relationship on their part. We think we know what kind of people we are dealing with when we form certain kinds of bonds with them, and when we come to learn, through their actions, that they are not the sorts of people we took them to be, we may come to question our connections with them. We needn't agree with T. M. Scanlon's idea that blaming consists in part in taking one's relationship to be (at least partially) impaired by the blamed person's action to see why via the process of blaming people often come to see that there are impairments to their relationships.[144] Notice that our reactions in these sorts of cases are better encapsulated by statements like "I just can't be friends with someone who would say such a thing to me" than "you've reached 10 demerits so your moral scorecard is too low for me to respect you as a fellow moral agent." The particularities of the situation matter and relate to sometimes highly specific relationship-customized norms and expectations about participants' characters.

5. The Time-Slice Property Objection

On the Finely Individuated trait view of blame, the contents of blame involve a property of an agent that is time-specific. This leads to a couple features of the view, which, on the face of it, might seem puzzling.

First off, even if a blamer knows that the blamed person's action reveals that she has the property of *being the kind of person who would φ given circumstances C*, the blamer might be in a position to know that the blamed person will never again be in circumstances C. Despite the fact that after the act the blamed person would still have the property of being such that she would act in an immoral way in circumstances C, we might wonder

[144] See Scanlon (2008, 2013).

why we should care about that when that trait has no bearing on the person's current or future actions.

Secondly, the blamed person might be immediately regretful such that she changes after the act so that she no longer *does* have the trait of being such that she would attributably φ in circumstances C, say because she comes to realize that you will blame her for it. Since she no longer approves of her action, we may wonder why it would be permissible to blame her on the basis of a time-slice property she no longer possesses. Another way to think about the worry, specific to the Minimal Approval View, is this. It seems that at t_2 an agent can be alienated from the desire she approved of at t_1, but it is only in virtue of not being alienated from her desire that she is blameworthy for it. Why then, should we not take the kind of alienation she has at t_2 as disqualifying her from blame at t_2?

But in both cases, I want to give an initially somewhat flatfooted response. The action is related to an agent, even one who has changed, because she is still the one who performed the action. The content of the blame is not time-indexed; it just contains the details about the circumstances within its content and specifies that this time-particular trait is true of the agent. The agent now is numerically identical, if not qualitatively identical to the agent who performed the action, and so it is related to her by being a part of her agential history. The affected parties still learn something about the blamed person and it's still perfectly reasonable in many cases for someone to take learning that someone is the kind of person who could *ever* be capable of doing what they've done to be grounds for taking a blaming stance towards that person. Worrying that the blamed agent is alienated from her action is misplaced, because we know that if her action is attributable then at the time of action she was not alienated from it, and so she met the ownership conditions for it. It might make sense to speak about some kind of alienation that the agent now has from her past action, but this is not the kind of externality-grounding alienation that is relevant to responsibility. The function of inquiries about alienation is just to find out if the agent meets the minimum conditions for ownership over her action such that it is expressive at the time of action.

However, this might seem to make the view subject to an opposite worry. It seems that the fact that an agent has changed for the better after the fact does and ought to in some cases modify our stance towards her. If an agent who does something she is blameworthy for becomes such that she no longer would do such a thing, apologizes, and does restitution, it seems that forgiveness becomes appropriate. On the view I have been arguing for, though, given that she is still numerically identical to the person with the blameworthy trait, it might seem that the view would predict that instead, blame continues to be warranted.

But the view I have put forth should not be confused with the view that once a morally problematic trait is revealed, blame is the one set response that is all-things-considered best. I have only argued that the contents of blaming attitudes are finely-individuated morally problematic traits as revealed by agent's actions. Depending on what kind of attitudes these are, other factors may influence their appropriateness in a given circumstance.

For example, take the view that warranted blame involves the appropriateness of a certain kind of directed attention. One possible scenario is that the aptness of blame gives you *pro tanto* reason to focus on the wrongdoer through a certain lens, but, given her repentance, you have more reason focus your attention elsewhere. Another possibility is that it may be equally permissible to blame or to withdraw one's blame, since the appropriateness conditions for focusing one's attention plausibly involve lots of cases in which its permissible to go in any one of several different ways.

Similarly, an emotion-based view of blame can, arguably, make sense of cases on which, for example, there is reason to be mad at a person on the basis of their having exemplified a certain trait, but there is also countervailing reason to feel some other emotion that is mutually-exclusive with remaining angry. In general, it should be emphasized that describing the content of blaming attitudes does not yet settle an ethics of blame.

I admit, however, that it is hard to see how countervailing considerations of an agent's post-action transformation could act to mitigate blame on the basis of finely individuated traits given a view on which blame consists merely in beliefs or judgments of some sort. On a crude version of the view on which blaming just is the judgment *that* someone has a morally objectionable trait, and traits attach to persons across time, then blameworthy people seem inexorably doomed to lives on which it once and forever all-things-considered appropriate for others to blame them. It is possible that more sophisticated accounts of the judgment view, when paired with the Finely Individuated Trait view, might be able to concoct ways to avoid this problem. However, it may be more promising to pair the Finely Individuated trait view with a different account of the kinds of constitutive attitudes of blaming.

6. The *Minority Report* objection: Pre-Blame

In the Stephen Spielberg film *Minority Report*, fortune-tellers known as pre-cogs are able to apprehend would-be criminals by foreseeing that they will commit a crime. The idea that someone ought to be punished on the basis of a crime without *actually* committing the crime (yet) is, on the face of it, unappealing. Certain compatibilist views have been accused of having the unintuitive consequence of allowing, at least in theory, for the possibility of morally sanctioned pre-punishment.

While the Finely Individuated Trait view is not a theory of the ethics of punishment, it does face an analogous problem: what I'll call the problem of "pre-blame." Suppose that neuroscientists were able to conclusively show that someone is the kind of person who would do some morally wrong thing, say Ψ-ing, in a set of circumstances, but it is a set of circumstances that she is not yet in, or perhaps will never be in. According to the Finely Individuated Trait view, since such a brain scan or other neurological test seems to reveal the relevant information about the person, it seems that she would be attributionally-blameworthy for being the kind of person who would Ψ in a particular set of circumstances, despite the fact that she has not actually Ψ-ed. This seems counterintuitive, presumably because we tend to think that it is only through a person's actual action that we can rightly blame her. While my strategy to respond involves biting the bullet, I want to show that focusing on several possible factors that drive its underlying intuition can diffuse the force of the counterintuitive consequence.6.1 Actual Action as Usual Evidence

In real life as we know it, the only completely decisive evidence we are ever given that someone has such a finely individuated trait comes via their action. Given the current state of science, it is not possible to definitively prove that someone is the kind of person who would φ in some maximally specified set of circumstances. That means that the *only* time we have conclusive evidence that someone is the kind of person who would φ in circumstances C is when we know that that person attributively attempts to φ in circumstances C. Given that these things always overlap in the actual world, it's not surprising that we would come to the conclusion that it is only appropriate to blame someone when she actually φs. We have deeply embedded norms, perhaps supported by principles of morality, about not assuming the worst about someone and not acting as though we know how they will act before we have conclusive evidence of it. Despite the fact that we *would* have conclusive evidence if we were able to scan people's brains, according to the objection, it's possible that our intuitions are nevertheless influenced by the fact that we are so used the only conclusive evidence of the relevant kind of trait being an actual action.

We don't know what it would be like to live in a world in which we could tell exactly what someone would do before they have done it, and there is just a general strangeness in imagining the scenario which may end up pervading all of our intuitions about it. Imagine we reject the view that it is appropriate to blame someone for their finely individuated trait in favor of a view on which her action must actually occur. Whatever kind of attitude we take blame to consist in, it is strange to think that in a world in which we knew exactly what was going to happen we would be able to entirely withhold that attitude until the action took place. Once we know that someone will commit a horrific murder we will resent them, make certain judgments about them, desire that they are otherwise, potentially want them to suffer or be punished, have reason to modify our relationship to them, etc. These cases in which we know what someone will do before they actually do it just seem to lead to seemingly unpalatable consequences about blame no matter whether we hold that people in such a situation can be blamed before their actions or only afterwards. This is reason not take a counterintuitive consequence of what a view posits about pre-blame to be a decisive objection.

6.2 Victim Relation and the Standing to Blame

Take another case in which the Finely Individuated Trait view seems to predict that it is appropriate to blame someone that might seem unintuitive.

> Trapped Tara: Tara wakes up one morning and declares that she going to murder someone, and you have every reason to believe that, if given the opportunity, she would follow through on her declaration. Little does Tara know, though, that while she was asleep last night her bedroom was air-lifted to a deserted island with her in it, so no murdering will be possible. Escaping from the island, let's suppose, is also impossible.

In this case you are given good evidence that Tara has the finely individuated traits of being such that she would murder someone in many circumstances in which she has people available to murder. However, you also know that she will never be in those circumstances. The Finely Individuated Trait view says that blaming Tara would be apt, but this may seem somewhat unintuitive. After all, Tara didn't actually do anything wrong, and she hasn't harmed anyone.

But I think our reaction is not so much that it's unfair to blame Tara, but rather that it's not so clear why we should bother with blame in this case. In particular, part of the intuition that blaming Tara is inappropriate seems to derive from the fact that it seems that that some of the standard ways of outwardly expressing blame would be inappropriate in a case in which no wrongdoing was committed. But it is standard to draw a *prima facie* distinction between the permissibility of blame and the permissibility of outwardly expressing blame or confronting the blamed person. The latter might take into account additional moral reasons one might have to express or refrain from expressing blame, and it also might take into account the standing of the blamer.[145]

It is controversial just what conditions give a person the standing to express blame, but three commonly cited factors thought to undermine one's standing to blame are complicity, hypocrisy, and meddling.[146] One suggestion as to what might unify these disparate seeming conditions is that standing is at least in part a function of the blamer's relationship to the victim of the wrongdoing.[147] When the blamer is complicit, she is partially at fault for the victim's situation; when the blamer is hypocritical, she cannot honestly align herself with the victim; and when the she is meddling she lacks the requisite connection to the victim.

If there's no possibility of Tara actually committing the wrongdoing, there is no real victim. Even if the content of blame strictly speaking

[145] Although, one might also need certain standing to appropriately blame in the unexpressed sense. This may seem more or less plausible depending on the account given of blaming attitudes and their contents. It is less controversial that standing matters for outward expressions of blame than for blame simpliciter.

[146] See Coates and Tognazzini (2013).

[147] See, for example, Bell (2013).

doesn’t require a victim, merely the existence of a trait, the permissibility of expressing it might be mitigated by the absence of a victim. It might be inappropriate to outwardly direct blame at Tara when there is no one she has actually harmed. To a lesser degree, the appropriateness of expressing blame might be mitigated in a similar way in pre-blame cases. While there is a future victim, there is no current victim, and that might alter the appropriateness of expressing blame. Since it can be difficult to untangle intuitions about the appropriateness of blame from the appropriateness of expressing blame, this is one further reason that we should give pause to putting a lot of stock in the seeming counterintuitiveness of appropriate pre-blame.

6.3 Stubborn Incompatibilist Intuitions

The intuition that the permissibility of pre-blame is unacceptable might also stem in part from sticky incompatibilist intuitions, which, methodologically speaking, we shouldn’t permit at this stage to provide a knock-down objection to the view. Libertarians will hold that it isn’t *really* possible to have 100% certainty about what someone will do, even in a world that has access to the most advanced possible scientific discoveries. In an indeterministic world, therefore, anytime someone predicts the future, the prediction can only ever be a confident guess. If a fortuneteller predicts that a person will act in an immoral way, it is, however unlikely, within that person’s power to overcome the situationally-influenced factors and change courses due to a pure exercise of the will. It’s possible that what rubs some of us the wrong way about these scenarios in which a person is pre-blamed is that *deep down* we think blaming someone on the basis of a prediction is unfair since she might not act in the way she is predicted to act. If so, then we have clearly failed to isolate the appropriate intuition.

Of course, an incompatibilist might use the fact that we seem unable to move away from these kinds of intuitions that tell against the permissibility of pre-blame as evidence that we have unshakable incompatibilist intuitions. This could be bolstered into an argument that the compatibilist project has gone awry at some point. However, it is part of the nature of the dialectic that we do have stubborn intuitions that pull us in favor of libertarianism, but that we also have stubborn intuitions favoring aspects of compatibilism and skepticism as well. All three of the following statements have intuitive appeal, and yet they are jointly incompatible: "We are morally responsible" "Being morally responsible requires having multiple paths available at the moment of choice." "Determinism is true, and rules out responsible action." The point I want to make here is just that insofar as our intuitions about cases may stem from intuitive support for one of these general ideas, we have reason to at least be cautious about using them to rule out views.

6.4 Denying the Existence of Resultant Moral Luck

I have, so far, acknowledged that the fact that it is somewhat counterintuitive that pre-blame would hypothetically be appropriate on the Finely Individuated Trait view, though I have argued that we ought to have a measured response to its counter-intuitiveness. But it should also be noted that this same consequence of the view actually helps explain our intuitions about another issue: resultant moral luck.

It seems that, in the actual world, it is appropriate to treat a successful murderer and a murderer whose plan is thwarted by something wholly outside of his control differently. This is puzzling, though, since the only real difference in the two cases is something that has nothing to do with any features of the two individuals.[148] But an adherent of the Finely Individuated Trait view could make use of what is known as the epistemic argument for denying the existence of resultant moral luck. In other words,

[148] Discussion of these issues in the contemporary literature tends to center on the treatment of the issue provided in Williams and Nagel (1976), although discussion also appears earlier in Feinberg (1962). For one recent treatment of the issue as it pertains to blameworthiness and praiseworthiness, see Hartman (2017).

the Finely Individuated Trait view is consistent with and even helps bolster a powerful argument that, despite appearances, it really is not appropriate to treat successful and (un)luckily unsuccessful murderers differently when all else is held equal.

As I have been arguing, the epistemic argument against resultant moral luck holds that the reason we often treat the two cases differently is because in the real world we rarely know the strength of someone's commitment to undertaking a certain course of action unless we have evidence from the fact that they actually went through with it.[149] As Dana Nelkin explains,

> Thus, rather than indicating our commitment to cases of resultant moral luck, our differential treatment of successful and unsuccessful murderers indicates our different epistemic situations with respect to each. If we were in the unrealistic situation of knowing that both agents had exactly the same intentions, the same strength of commitment to their plans, and so on, then we would no longer be inclined to treat them differently.[150]

The Finely Individuated Trait view helps explain why we ought to treat the two the same if we were in the same epistemic circumstances: what matters is the fact that the person has the quality of being such that she *would* φ in circumstances C.

Countenancing moral luck is generally taken to be a *problem* of some sort, and so it is meant to be counterintuitive that moral luck should make a difference. Since something like the Finely Individuated Trait view is needed to bolster a crucial argument that moral luck doesn't make a difference, this should count in its favor. The same aspect of the view that

[149] See Richards (1986), Rescher (1993), Rosebury (1995), and Thomson (1993).

[150] Nelkin (2013).

under one lens looked counterintuitive, also has strong intuitive appeal when focusing on issues of resultant moral luck. Thus the problem of preblame does not give us anywhere near conclusive reason to reject an account of blame's content that seems as promising as the Finely Individuated Trait view.

Chapter 5: *The Case of Forgetting*

1. Introduction

In the previous four chapters I developed the Minimal Approval account of attributional-responsibility. I suggested that an agent can be attributionally-responsible for acting when the production of her action meets fairly minimal ownership conditions. In Chapter 3 I showed how this account explains why agents are attributionally-responsible for weak-willed actions: unlike in cases of compulsion, agents minimally approve of their weak-willed actions. Given this explanation, it seems right to say that weak-willed agents are also derivatively attributionally-responsible for *failing* to do what they *would* have done were they to have been strong-willed.

In this chapter I consider the prospects for extending a similar line of thought to account for cases in which agents seem responsible for forgetting to do something they ought to have done. I argue that this move is not as appealing as it might seem, and, in addition, I argue that similar attempts to expand the notion of attributional-responsibility by other theorists to account for these kinds of cases are misguided.

Instead, I argue that agents are responsible in a non-appraising, role-responsibility sense for these sorts of forgettings, but they are not attributionally-responsible. This sets an important limit on the project of the dissertation as a whole; while I argue that blaming practices that center around finely individuated traits as expressed by actions of which agents minimally approve play a crucial and significant role in our moral responsibility practices, they are not exhaustive of our responsibility practices.

2. The Minimal Approval View and The Case of Forgetting

Consider the following case from George Sher’s book, *Who Knew? Responsibility Without Awareness*:

> Alessandra, a soccer mom, has gone to pick up her children at their elementary school. As usual, Alessandra is accompanied by the family's border collie, Bathsheba, who rides in the back of the van. Although it is very hot, the pick-up has never taken long, so Alessandra leaves Sheba in the van while she goes to gather her children. This time, however, Alessandra is greeted by a tangled tale of misbehavior, ill-considered punishment, and administrative bungling which requires several hours of indignant sorting out. During that time, Sheba languishes, forgotten, in the locked car. When Alessandra and her children finally make it to the parking lot, they find Sheba unconscious from heat prostration.[151]

Assuming that Alessandra is a generally caring and thoughtful person who loves Sheba, it is not easy to get clear on our intuitive responses to such a case. On the one hand, we might feel bad for Alessandra for making such an upsetting mistake, one that does not on the face of it seem to reflect any sort of characteristically morally wrong personality traits. On the other hand, she seems to have acted negligently—we think she should still be held responsible in some sense for what happened. To bring out this intuition, imagine that Alessandra were to offer no apology to her children and family for what she had done.[152] This, I think, would seem

[151] Sher (2009): 24.

[152] Although, perhaps apology is warranted even in cases that involve no responsibility. See, for example, Talbert (Forthcoming): 17, and Scanlon (2008): 150. Perhaps a similar intuition could be brought about, though, by considering the case in which Alessandra feels no remorse or special duty to comfort others affected as a result of her causal responsibility.

highly inappropriate. While her family might rightly be sensitive to the fact that Alessandra herself might be suffering from the tragedy of the incident, we might not think them out of line to expect her to make amends of some sort. The verdict here is unclear, but there is at least a *prima facie* case for Alessandra's being responsible in some way.

What does the Minimal Approval view say about Alessandra's neglect of Sheba in this case as it pertains to her attributional-responsibility? It might be thought that the Minimal Approval view actually has resources that other views in the Deep Self family lack to show why Alessandra is, after all, attributionally-responsible for leaving Sheba in the car. In Chapter 3, I argued that agents could be attributionally-responsible for acting out of weakness of will. When Sam acts out of weakness of will and ends up going to a party rather than studying for her exam, she is attributionally-responsible for going to the party. It makes sense in these cases to also say that she is derivatively attributionally-responsible for not studying. If being attributionally-responsible for what you actually do can render you attributionally-responsible for what you fail to do as well, then we might think we should conclude from the fact that Alessandra is (let's stipulate) attributionally-responsible for staying to talk to the school administrators, that she is also attributionally-responsible for leaving Sheba in the car.

In order to evaluate whether or not this is right, though, we'll need to pay more attention to the principle that lets us move from Sam's attributional-responsibility for her action to her attributional-responsibility for her omission. One principle that might get us from Sam's responsibility for going to the party to Sam's responsibility for failing to study is the following:

> Modal Bridge Principle: If an agent, A, is attributionally-responsible for φ-ing-at-*t*, then A is also attributionally-responsible for not ψ-ing, where ψ-ing is anything else A could have done at *t*.

Adopting this principle would also give the result that Alessandra is responsible for not tending to Sheba, since, arguably, she *could* have tended to Sheba rather than stayed to talk to the administrator. But notice that adopting this principle reopens the issue that Deep Self views were partially created to avoid. Given the fixity of the past and the truth of determinism, Incompatibilists will argue, there is nothing that Alessandra might have done at t other than what she actually did—stay and talk to the administrator. Given the hard work of maneuvering around these questions and arguing for their irrelevance in the case of attributionally-responsible action, it would be unfortunate if they were simply to reappear in the case of attributionally-responsible omission.

There is also a larger problem with this view. Consider the following case:

> Oblivious Ollie: Ollie is walking to work and meets the conditions of Minimal Approval for walking to work. Unbeknownst to him, there is a small child drowning in the river behind him, but he never turns around and notices the child drowning. Is Ollie attributionally-responsible for failing to turn around and help the child? It would seem that he should not be, but given a plausible (compatibilist) construal of the Modal Bridge principle, he is attributionally-responsible for walking to work, and could have instead turned around and saved the drowning child.

So perhaps, instead, we ought to adopt a principle like the following:

> Evidence Bridge Principle: If an agent, A, is attributionally-responsible for φ-ing-at-*t*, then A is also attributionally-responsible for not ψ-ing-at-*t* where ψ-ing is any act that the agent has sufficient evidence is a potential course of conduct for her.

This principle both correctly exempts Ollie from attributional-responsibility, and, at least on the face of it, might even avoid the incompatibilist worry. But consider the explanation it gives as to why Alessandra is responsible for leaving Sheba in the car. She is responsible for talking to the school administrators when it is the case that she also had evidence (to which she failed to attend) that she could go tend to Sheba to avoid catastrophe at that moment instead. But what role does the existence of the evidence to which Alessandra failed to attend play in explaining *why* she is responsible? Alessandra may fail to attend to the evidence due to a random misfiring in her brain, and so the fact that this evidence existed doesn't tell us anything about what Alessandra is like agentially. It may be true that it tells us that a surface-level normative fact is true about Alessandra: there was evidence that an alternative course existed, evidence to which she ought to have been responsive. But grounding the explanation of an ascription of attributional-responsibility in a normative fact like this would require accepting a sort of deep foundational asymmetry between attributability for acts and omissions. Since attributability for actions on the Minimal Approval view is determined by pure metaphysical conditions of agency, it would be odd if first-order normative theorizing needed to take place in order to determine the conditions for attributability for an omission,.

Instead, I think the Minimal Approval view should be coupled with a principle that ties attributional-responsibility for omissions more closely to the fact that the mechanism of the agent's action relates to the fact that the agent would approve to some degree of what she does do instead of the omitted action. This will explain why Sam is attributionally-responsible for failing to study, but will have the result that Alessandra is not attributionally-responsible for failing to tend to Sheba. This is the principle I think the Minimal Approval theorist ought to adopt:

Contrastive Approval Bridge Principle: If an agent, A, is attributional-

> ly-responsible for φ-ing-at-*t*, then A is also attributionally-responsible for not ψ-ing-at-*t*, when ψ-ing is an alternative considered in the worlds in which A meets the conditions of minimal approval.[153]

Sam is attributionally-responsible for not studying, not because she could or should have studied, but because of a fact about her agency. Sam minimally approves of going to the party *even though* she knows that she could be studying, and this is related to what leads her to go to the party. Notice that, given this bridge principle, Sam's responsibility for her omission actually has something to do with her approval. She's responsible for going to the party because she minimally approves of it, and she's responsible for not studying because she minimally approved of doing something else *instead*. Her responsibility for her omissions comes from the fact that we know something about what she is like when she is faced with the

[153] What grounds these bridge principles? One possibility is that they are derived from principles that would explain why agents are responsible not just for their actions, but also for the consequences of their actions. On this strategy, the fact that an agent does not ψ instead when she φs is just a consequence of φ-ing like any other. Different principles of this form correspond with the modal, normative, and contrastive approval bridge principles. For example:

Modal: *If an agent, A, is attributionally-responsible for φ-ing-at-t, then A is also attributionally-responsible for all of the consequences of φ-ing that she could have foreseen, (possibly including the fact that A will not ψ instead).*

Evidence: *If an agent, A, is attributionally-responsible for φ-ing-at-t, then A is also attributionally-responsible for all the consequences of φ-ing that she had enough evidence to foresee, (possible including the fact that A will not ψ instead)*

Normative: *If an agent, A, is attributionally-responsible for φ-ing-at-t, then A is also attributionally-responsible for all of the consequences of φ-ing that she should have foreseen (possibly including the fact that A will not ψ instead).*

Contrastive Approval: *If an agent, A, is attributionally-responsible for φ-ing-at-t, then A is also attributionally-responsible for all of the consequences of φ-ing she foresees in the worlds in which she reflects on which of her motivations to act on at t and approves of φ-ing, (possibly including the fact that A will not ψ instead).*

choice situation in which the option of studying features—there's a part of her that wants to go to the party *anyway*.

The same can't be said of Alessandra. Her minimal approval of talking with the administrator doesn't tell us anything about what she cares about with respect to Sheba, since considerations about Sheba aren't among those that would factor into her agential psychology whatsoever when she so approves.

Note that Alessandra's failure to think about Sheba's being in the car is plausibly one of the things we ought to hold fixed when assessing whether or not she has the disposition required for being attributionally-responsible for talking to the administrator according to the Minimal Approval view. To see why this should be so, notice that people are plausibly attributionally-responsible for φ-ing even if there are other options that they do not remember that would silence their desires to φ were they to come up with those options. If I had remembered that I had spaghetti in the pantry I would never have given any weight to a desire to go to the grocery store, but this doesn't mean that I'm not attributionally-responsible for going to the grocery store.

Alessandra never considers Sheba, and so there is no bridge that shows that just because she is attributionally-responsibility for talking to the administrator, she should be attributionally-responsible for forgetting Sheba. This leaves us needing to look elsewhere to explain the intuition that Alessandra is in some sense responsible for failing to tend to Sheba.

3. The Tracing Strategy

One possible way to show that Alessandra might, after all, be attributionally-blameworthy for failing to tend to Sheba would be to use the tracing strategy.

The tracing strategy is used to explain why agents are responsible for certain acts that are not directly attributable to agents. According to proponents of the tracing strategy, an agent may also be attributionally-blameworthy for acts that should have been and could have been prevented by an earlier act, when the agent is attributionally-blameworthy for the prior omission. For example, an out of control drunk driver who hits another car may not meet the attributability requirements for hitting the car, but nevertheless be blameworthy. Given the tracing strategy, we can explain her attributional-blameworthiness for her behavior via the fact that she is attributionally-blameworthy for her earlier acts of getting intoxicated and choosing to drive rather than catch an Uber to and from the bar. Note that the agent must be *blameworthy* for the prior action in order for it to explain why the agent is blameworthy for the current action. It cannot merely be that the person made a choice in the past that lead to the bad outcome. If a mother decides to pick her child up from school rather than have him take the bus home, and on the way home the child gets injured in a car accident due to no fault of the mother's driving, she does not become blameworthy just because she made the prior choice to pick him up by car.

While some have voiced concerns with the viability of the tracing strategy in general, for the sake of argument I'll grant that the tracing strategy, generally speaking, works, and provides a good explanation for a broad range of cases of blameworthiness.[154] The question then is whether

[154] Tracing strategies are more often invoked by reasons-responsiveness and control-based theorists of responsibility. While the tracing strategy is in principle open to defenders of the Deep Self view, defenders of Deep Self views sometimes couple their theories of attributability with defenses of the H-Tradition of responsibility and aim to defend a Strawsonian quality-of-will thesis on which blameworthy agents' actions must express ill-will. It's not clear that a tracing strategy is compatible with this sort of view. Although I've developed the Minimal Approval account as B-Tradition account of responsibility, interestingly, the hypothetical nature of the reflectiveness required for Minimal Approval might make the tracing strategy somewhat less necessary in certain standard cases, and so adopting it might help quality of will theorists avoid the tracing strategy. Another recent set of objections to the tracing strategy concerns certain articulations of the strategy's

or not a Minimal Approval theorist could use the strategy to explain the intuition that Alessandra is attributionally-blameworthy.

I think such an explanation is bound to fail. There are things that Alessandra could have done that would have made it so that she didn't forget Sheba in the car. She could have tied a string around her finger or set a phone alarm reminding her that the dog was in the car, or she could have avoided bringing the dog altogether. She failed to do those things, but in order to make it the case that she is attributionally-blameworthy for leaving Sheba, we would also need to know that she is attributionally-blameworthy for those prior omissions. Given the reasonable assumptions that she would be no more than a minute or two picking up her kids, that she doesn't have any general tendency to forget important things when her plans get modified, and doesn't have any particular tendency to forget about Sheba, it's just not clear that she would be attributionally-blameworthy for not setting a phone alarm.

Using just our intuitive notion of attributional-blameworthiness, it is just not clear that Sheba does anything that reflects a bad trait prior to her leaving Sheba in the car. And using the Minimal Approval view combined with the Contrastive Approval Bridge Principle doesn't seem to give us the result that there is anything that she is attributionally-blameworthy for either. In order for her to be derivatively attributionally-blameworthy for leaving Sheba in the car due to a tracing explanation, it would have to be the case that in every specification of the story that generates the intuition that Alessandra is responsible for leaving Sheba in the car, she is also such that she is blameworthy for minimally approving of what she was doing instead of setting a phone alarm, or blameworthy for minimally approv-

compatibility with reasonable epistemic conditions on responsibility. For this line of argumentation, see Vargas (2005), and Shabo (2015).

ing of what she was doing instead of thinking about ways to keep herself vigilant about Sheba's wellbeing.

Furthermore, I don't think tracing can give us the right *kind* of explanation for our intuition that Alessandra is responsible in some sense. Take a case in which Alessandra has complete knowledge and full awareness of the fact that there is a 1/10,000 possibility that she will leave Sheba in the car at some point. If Alessandra were to fail to set a phone alarm, given the other demands of her busy life as a mother, I think few would blame her for this. Nevertheless if she were to in fact leave Sheba in the car I think this would still elicit our reaction that she owes some form of recompense to her family in light of leaving the dog in the car. Even when she has done nothing wrong in the past, we still think she ought to be held responsible in the present.

4. An Argument Against the Minimal Approval View as the Correct View of Attributability?

Although I have presented the Minimal Approval view as a view that is robust across a wider variety of different causal stories than traditional Deep Self views, cases like Alessandra's open up the possibility of the objection that it is not robust across a wide *enough* variety of causal stories. One such challenge comes from proponents of a loosely associated set of views sometimes called New Attributionist views.[155] In this section I raise the possibility that the case of Alessandra could be leveraged into an objection that agents are in fact attributionally-responsible for an even broader range of acts and omissions than what the Minimal Approval view can account for. I consider George Sher's and Angela Smith's views as offering competing views of attributional-responsibility that give the

[155] New Attributionists include George Sher, Angela Smith, Matt Talbert, and T. M. Scanlon. For a discussion of this term and its application, see Talbert (2016). Matt Talbert, however, explicitly takes the view that Alessandra is not responsible [see Talbert (Forthcoming)].

result that Alessandra is attributionally-responsible for leaving Sheba in the car, but argue that neither account succeeds at this task.

4.1 Sher's Agent-Constituting Mechanism View

On Sher's account, agents need not approve of their motivations in any sense in order to be attributionally-responsible for them. The criterion for attributional-responsibility on his view is instead just that the action or omission in some way stems from the constitutive features of the agent *qua* the particular rational agent she is. He takes these features to include not just the agent's cares, values, and plans but also her general dispositions and tendencies, as well as the relevant neurophysiological mechanisms that cause such things.[156] In this way, he is able to hold that agents may be attributionally-responsible for their actions and omissions even if they are not endorsed by the agent in any way, so long as they are caused by one of the kinds of mechanisms that are involved primarily in agency.

In Alessandra's case, is the fact that she leaves Sheba in the car caused by her constitutive agential features? According to Sher, it is. Things like "her concern for her children, for example, or her tendency to focus intensely on whatever issue is at hand" are just the sorts of things that make her the agent that she is. And even though Alessandra would never approve or judge it best to leave Sheba in the Car, according to Sher, "…we must locate the significance of Alessandra's failure to remember Sheba not in what it reveals about her judgments about reasons, but rather in its being caused by the same psychophysical structure that sustains her ability to *make* such judgments."[157] And so, he thinks, we should conclude that she is in fact attributionally-responsible for her omission.

I'd like to echo a general line of criticism taken up by Angela Smith and others to Sher's larger account, though.[158] The problem is that, while

[156] Sher (2005): 122.

[157] Sher (2009): 131.

[158] See, for example, Smith (2008), and Mason (Provisionally Forthcoming).

he articulates a promising-sounding necessary condition for attributional-responsibility, Sher fails to provide a sufficient condition, since a mere causal connection between these kinds of structures and action is intuitively not sufficient make an agent attributionally-responsible. Take the following example: Imagine a case in which an agent has the quality that she thinks really really hard about her decisions just about all the time. But this same feature of her agency sometimes causes her brain, at other times, to short out due to being in overdrive, making her collapse in place. Suppose that she does not know that these incidences are related. One day she collapses and falls into someone on the way down, injuring this person. Pretty clearly, this agent is not attributionally-responsible for her falling, and yet her behavior is, it seems, caused by one of the features that makes the agent who she is *qua* rational agent.

Perhaps Sher could tell a story about how this behavior is not caused in the right sort of way by her tendencies, but it is unclear what that story could amount to without changing the extension of the view pretty significantly. The broader lesson here is that the criterion for attributional-responsibility can't just be that the behavior is a side effect of the agent's rational-agent tendencies, as Sher would have it.

4.2 Smith's Rational Relations View

In a series of articles, Angela Smith defends the Rational Relations view of responsibility according to which, broadly speaking, if an agent's behavior displays a lack of rational concern, it is appropriate to hold that agent responsible for the behavior.[159] Her view is generally taken to be an account of attributional-responsibility that can accommodate the fact that agents like Alessandra are attributionally-responsible for things like forgetting. However, I will show that depending on the interpretation of Smith's view, it either does not give the result that Alessandra is responsible, or is not really a theory of *attributional*-responsibility.

[159] Smith (2005, 2008, 2012, 2015).

I'll start with the first horn of the dilemma. It is understandable that Smith's view would generally be taken to be able to give an account of Alessandra's attributional-responsibility, since Smith's own paradigm case of responsibility that she uses to motivate her view is very similar to the Alessandra case, and Smith seems to explain responsibility in this case as being grounded in a revealed lack of concern via an omission. This certainly sounds like attributability language. In Smith's case, she forgets a friend's birthday and the friend takes her to be morally responsible for her forgetting, as her dispositions and patterns of attention reasonably convey to her friend a lack of care.

Smith describes the birthday-forgetting case in the following way:

> I did not *intend* to hurt my friend's feelings or even *foresee* that my conduct would have this effect. I just forgot. It didn't occur to me. I failed to notice. And yet, despite the apparent involuntariness of this failure, there was no doubt in either of our minds that I was, indeed, responsible for it. Although my friend was quick to pardon my thoughtlessness and to dismiss it as trivial and unimportant, the act of pardoning itself is simply a way of renouncing certain critical responses, which it is acknowledged, would, in principle, be justified.[160]

But does Smith's forgetting her friend's birthday reveal an objectionable degree of concern about her friend? Certainly it might. Someone self-absorbed or who undervalued her friendship might make this mistake frequently precisely for the reason that she failed to cultivate the appropriate level of concern. But it is not obviously true that any such description is true of Smith in this case. It might be that perhaps a better friend would focus her attention so intently on the birthdates of all her friends with the special intent to ensure this never happens. But if Smith has no

[160] Smith (2005), 236.

special reason to worry that she'll forget, behaving this way would seem to be supererogatory. It's possible, it seems, to have an appropriate level of concern for one's friends without doing this.

Commenting on the preceding passage from Smith, Matt Talbert says the following:

> Though Smith doesn't say that she is blameworthy here, it's a natural way to read the case and it's easy to imagine a person in the position of Smith's friend responding to her with the negative attitudes involved in moral blame. I take Smith's view to be that *if* blame is appropriate in this case, then what makes it appropriate is just that in forgetting her friend's birthday, Smith reveals something objectionable about her orientation toward her friend.[161]

Presumably Talbert's reading of Smith would also apply to Alessandra's case. If Alessandra's forgetting did display objectionable lack of concern for Sheba, then this would be sufficient for appropriate blame. The question then becomes, in in Alessandra case: is blame, after all, appropriate in the particular case due to objectionable lack of concern for Sheba?

I agree with Talbert that if it's true that we can tell the story such that Alessandra more generally cares a great deal about Sheba then we should conclude, on an account like Smith's, that leaving Sheba in the car does not reveal anything objectionable about her orientation towards the dog, and thus Alessandra is not attributionally-responsible for her omission. And so we are left with cases on which it seems intuitive that Alessandra is responsible in some sense, despite the fact that leaving Sheba does not reveal anything objectionable about her level of concern for Sheba. Again, the only way of avoiding this outcome for Smith would be to argue that the very fact of Alessandra's omission makes it the case that it is impossi-

[161] Talbert (Forthcoming).

ble for her to have sufficiently cared about Sheba. But this seems implausible.

As Talbert points out, the idea that anytime someone leaves a dog in the car this shows an objectionable attitude of lack of care must be mistaken, because were an extraordinary circumstance to occur, even a high level of concern for Sheba would not be enough to prevent forgetting her in the car. But perhaps we could modify the view to something like, in the absence of extraordinary intervention (and presumably Alessandra's discussion with administrators should not count as extraordinary) if Alessandra had properly cared for Sheba she wouldn't have left her in the car. But, as Talbert is I think right to point out,

> Even under normal circumstances, people sometimes forget or fail to notice things about which they care very much. More generally, I take it that even under normal circumstances, what we notice, what we remember, what fails to occur to us, and so on, doesn't *necessarily* indicate what we value or how much we value it.[162]

It might be that forgetting a birthday or forgetting a dog in the car *generally* indicates that a person has an objectionable lack of concern, but this is not sufficient evidence that a lack of concern is *actually attributable* to any particular agent. This is why it's crucial in an account of attributability to have some condition that makes reference to the agent's psychology, not just the agential psychology that her action/omission makes us warranted in believing might be the case. As a theory that can explain why Alessandra's responsibility for leaving Sheba must be grounded in some trait that is attributable to Alessandra, Smith's theory falls short.

This leads to the other horn of the dilemma, though. It seems to me that Smith might be able to agree that Alessandra is responsible even in

[162] Talbert (Forthcoming): 16.

the cases in which her forgetting doesn't indicate an objectionable lack of concern for Sheba. In fact, in various places it seems as though Smith conceives of the connection between our tendencies and our evaluative judgments to be that our tendencies merely give those around us *prima facie* evidence of our evaluative judgments such that we are required to answer for our tendencies—to either own up to the fact that they do reveal what we are like in a given instance or, crucially, to explain how in the particular case they came about in some way that shows that they did *not* stem from our lack of concern. Smith writes that:

> If one judges some thing or person to be important or significant in some way, this should (rationally) have an influence on one's tendency to notice factors which pertain to the existence, welfare, or flourishing of that thing or person. If this is so, then the fact that a person fails to take note of such factors in certain circumstances is at least some indication that she does not accept this evaluative judgment.[163]

I am not sure whether Smith means here that it's reasonable for the *blamer* to take it as some indication or it's reasonable for us as philosophers knowing the full specification of the story including the mental-causal story of the potentially blamed person's behavior to take it as indication that she does not accept the evaluative judgment. If we read Smith on the former interpretation, though, then this gives her a way to preserve the idea that Alessandra is responsible even if Talbert's claims are all true. In Alessandra's case, her leaving Sheba in the car gives others a *prima facie* reason to think she does not care sufficiently about Sheba, to which she ought to respond. And the required response may involve apology and/or other recompense for what happened that helps to make clear that she does not have any lack of general care about Sheba or her family.

[163] Smith (2005): 244.

But just because some fact is evidence that Alessandra doesn't care enough about Sheba, this doesn't suffice to show that Alessandra actually doesn't care enough about Sheba. And so it seems that, whatever the merits of this view, it is no argument against the Minimal Approval view as a view of *attributional*-responsibility since it is possible no negative trait is, in fact, attributable to Alessandra, (though it might not be unreasonable for her family to initially attribute one to her.) It might be possible to use the word "attributability" to refer not the set of actions that reveal something about what the agent is like such that the agent is properly subject to praise and blame on the basis of those traits but instead to the set of behaviors for which it is *reasonable* to respond to *as if* they revealed something about what the agent was like due to the fact that they have a special kind of significance for moral-social relations.[164] But doing so puts the view out of conversation with the Minimal Approval view, which aims to give the fully specified conditions on which a behavior does, after all, reflect a blameworthy trait of the agent. So, on this second horn, the reading of Smith's view may be more charitable, but it can't be leveraged into an objection that the Minimal Approval view misidentifies the bounds of attributional-responsibility due to its handling of the Alessandra case.

5. Accountability Without Attributability

Let's take stock. I have argued that while it seems that Alessandra is in some sense responsible for leaving Sheba in the car, the most promising

[164] For evidence that there is more than one sense in which the word "attributability" tends to be used in this literature, see the discussion in the comments thread of this post at the PEA Soup blog: http://peasoup.typepad.com/peasoup/2009/05/scanlon-on-moral-responsibility-blame-part-1.html Here, T.M. Scanlon writes that the special relevance of attitudes that are in principle judgment-sensitive comes from the fact that they are "particularly significant for our relations with each other (not because they are "attributable" or "belong to" the agent in a sense in which other attitudes to do not)." "Or, to put the same point differently," he writes, "we tend to think of these attitudes as 'attributable' in a way that others are not only because their content has this greater significance."

bridge principle to explain attributional-responsibility for omissions on the Minimal Approval view gives the outcome that she is not attributionally-blameworthy. The tracing strategy also gives the result that Alessandra is not attributionally-blameworthy even in cases in which she still seems responsible. Alternative views of attributional-responsibility that are able to return the result that Alessandra is attributionally-blameworthy for her omission seem either to face serious problems of their own or else not truly be views of attributional-responsibility in the sense I have been concerned with.

One option would be to bite the bullet and conclude that Alessandra is, after all, exempt from all moral responsibility for leaving Sheba in the car. Sometimes theories should prompt us to give up our common sense intuitions about every day cases, and our intuitions in this case seem far from clear. Some have defended the view that mistakes like Alessandra's are not the kinds of things for which a person can be morally responsible, and this is not always seen as a difficult bullet to bite.[165] This line of thought can be further supported by the fact that Alessandra ought to reassure her family that her omission was not due to lack of concern for Sheba. Furthermore, people may often be in poor epistemic circumstances in regards to knowing whether they are in fact attributionally-responsible for their mistakes, and so if Alessandra were to confidently proclaim her innocence, given the stakes we might rightly find her to lack appropriate humility. Biting the bullet here is not obviously wrong, and is certainly open to the Minimal Approval theorist.

But I don't think the intuition that Alessandra is in some sense responsible should be written off so easily. One reason is that there's an important difference between an uncertain intuition and an ambivalent intuition. And our intuitions about Alessandra seem to me to be of the latter

[165] See, for example, Zimmerman (1986), and Talbert (Forthcoming).

type. It's not just that we're not sure if Alessandra's case counts as a case of something one can be morally responsible for, it's that she does seem responsible in one sense while not seeming responsible in another sense. Following David Shoemaker's approach to ambivalent intuitions about responsibility, I think we should take this ambivalence in our intuitions as evidence of the fact that Alessandra's omission fails to satisfy the conditions for one type of responsibility but does satisfy the conditions for another type.[166] In particular, I think that our intuitions reflect the fact that this is not the kind of thing that does reveal something about what Alessandra is like as a person, and so character-implicating responses of the kind that are appropriate in cases of attributional-responsibility are inappropriate. And yet, it still seems that Alessandra *owes* something to her family in light of her mistake. Even in the case in which her family is completely assured of her lack of attributional-responsibility, it seems that they would be justified in expecting and feeling entitled to an apology. And it seems she owes it to them. This sense of owing recompense of some sort plausibly implicates some form of responsibility response that goes above and beyond the mere the appropriateness of agent-regret since it is not just that she Alessandra ought to feel badly about what happened; it seems that her family *deserves* something from her: at minimum, an apology.

My proposal is that Alessandra is not attributionally-responsible but is responsible in a non-appraising accountability sense. The relationship between attributional-responsibility and accountability-responsibility is frequently thought to be that attributability is sufficient for mere appraisals while an agent must meet the further requirements for accountability in order to be subject to proper sanctioning responses of any kind. On this picture, attributability is a necessary condition for accountability and in order for an agent to be accountability-responsible an agent's behavior

[166] Shoemaker (2009, 2011, 2015).

must not only be attributable but she also must possess some further capacities such as normative-competency or even libertarian free-will. On this sort of picture, if Alessandra is not attributionally-responsible for leaving Sheba in the car, she is certainly not accountability-responsible for doing so. I want to instead adopt an alternate picture on which attributional-responsibility and accountability-responsibility are wholly distinct.[167]

I will not be able to give anywhere near a full account of accountability-responsibility, but I do want to very briefly sketch once possible picture. Accountability-responsibility, on this picture, is derived from role responsibility.[168] Many of the roles we play in relationship to others, such as

[167] As Zheng (forthcoming) points out, this way of distinguishing concepts, while not as popular in moral philosophy has a number of analogues in in political philosophy. Two distinct concepts of responsibility, an appraising and non-appraising sense, have sometimes been distinguished by using different terminology. For example, Schmidtz and Goodin (1998) distinguishes "blame responsibility" from "task responsibility" and Lavin (2008) distinguishes "liberal" from "post-liberal" responsibility.

[168] A somewhat similar picture is sketched in Zheng (2016, Forthcoming). In Zheng (2016) the distinction between attributional-responsibility and accountability-responsibility is made by grounding it in a "conceptual genealogy" meant to show how attributability and accountability concepts arise from distinct sources of philosophical concern. Zheng shares my general conception of attributional-responsibility as arising from a metaphysical question about how to make room on a naturalist picture of the world for our sense that we are in some important way the authors of our own conduct such that our behavior can open us up for appraisal-based praising and blaming responses. The concept of accountability, in contrast, concerns issues of fairness and arises from strictly practical normative concerns in moral and political philosophy. For Zheng, accountability-responsibility derives from the fair division of labor. Individuals have duties to shoulder certain social burdens that they are involved in causing because, as Zheng puts it, "when a person's action brings about some negative consequences for others, this generates a social problem that simply cannot go unaddressed. These costs must be picked up somehow and by someone, even if there is no bad intention or fault on the part of the person involved, because there are victims who deserve redress. This means that under a fair system of distributing burdens, it will often be appropriate for the person who performed

our roles as caretakers, guardians, teachers, and bosses, generate certain kinds of role-related responsibilities. As Sam Scheffler puts it, "...it is a familiar fact that such ties are often seen as a source of special responsibilities. Indeed, we would be hard pressed to find any type of human relationship to which people have attached value or significance but which has never been seen as generating such responsibilities."[169] Role-related responsibilities might be more or less moral depending on the particular role in question.

It might be thought that all we mean when talking about role-responsibili*ties* is that we have some duties or obligations that obtain in virtue of our relationships. However, I suspect the connection between role-responsibilities and moral responsibility is tighter than that. It seems to me that we are *directly* accountable to others for failing at our role-related responsibilities, and when our responsibilities are moral, we are generally straightforwardly morally accountable for doing so. [170] The question of whether or not I am responsible in this sense completely circumvents the question of whether or not my action/omission expresses something morally bad about me. I might take responsibility for the welfare of another person's children while they are on my watch, but if one of these children gets hurt on my watch this does not necessarily reflect something bad about what I am like as a person, such that it would be appropriate to

the action to bear a large share of the costs: she can be asked to compensate for damages, make reparations, or to change her practices to prevent future failures. But notice that none of this requires an assessment of character, intentions, attitudes, or values—she can justifiably be required to bear (at least some of) the costs of her behavior whether she performed them out of malice, negligence, or sheer (non-culpable) ignorance or accident." Zheng (forthcoming) builds on this foundation and argues that accountability in this sense is grounded in role-related responsibility.

[169] Scheffler (1997): 190.

[170] Similar ideas are developed in several different contexts in: Scheffler (1997), Miller (2004), Cane (2016), Zheng (2017, forthcoming), and Lawford-Smith and Collins (2017).

blame me in any kind of appraising sense. I still owe my friend an apology and special concern for the child's recovery given that I am accountable to her.

This is not to say that there is no set of excusing conditions for accountability-responsibility, but these conditions do not come from failing to meet certain metaphysical agential criteria. Rather, they might pertain to the *fairness* of being held to have certain role-related responsibilities at a given time and circumstance. For example, the fairness of the conditions under which a person comes to take on a role-related responsibility may be one relevant factor. As David Miller puts it when discussing a similar concept of responsibility,[171] fair ascriptions of role-related accountability may also

> ... rest on implicit norms concerning the capacities that human beings can be expected to possess. Thus we do not hold people responsible for the consequences of their actions in cases where those consequences could only have been avoided by a superhuman display of strength. Likewise we do not hold people outcome responsible in cases where they are coerced into acting as they do, and in making these judgments we rely on our intuitive sense of how much pressure a normal person could be expected to resist. At the same time, diminished capacity does not relieve a person of outcome responsibility. People cannot escape it merely because through ignorance they failed to anticipate the results of their actions. An unusually clumsy person can be held responsible for the damage he causes as he blunders about the world[172]

[171] Miller speaks of "outcome responsibility" which roughly maps on to what I am calling "accountability" and contrasts this with "moral responsibility" which he describes as playing a similar role to what I have called "attributional-responsibility."

[172] Miller (2004): 245.

I suspect that in many cases when a person "takes responsibility," it is in this role-related accountability sense. In taking responsibility an agent clarifies or reaffirms what she takes to be her reasonably assigned role-related duties for which she is directly accountable even in the absence of displaying any immediate morally objectionable traits in failing to meet those duties.

Returning to the case of Alessandra, her role as Sheba's caretaker makes her liable for something bad happening to Sheba on her watch, and this holds completely independently of whether or not her failure to do so reflects a bad (even finely individuated) aretaic trait of hers. Alessandra, while not attributionally-responsible for leaving Sheba in the car does owe redress to her family members for her mistake. Her family would be wrong to think that she has a bad trait, or to alter their perceptions of her or their relationships to her in light of her mistake. Her family, however, would be right to expect that Alessandra should apologize and make any and all attempts to make it up to them. She should do these things not just because she is the mother of a family that has just experienced a horrific event but also because of the specific accountability to others affected that she has, given that she had a role-based obligation to care for the dog at that point in the day.

It should be noted that, were Alessandra to attributably fail to live up to the demands of her role responsibility in her *response* to the situation, she *would* then additionally be properly subject to appraisal-based attributional-blame. In this way, we can capture the full force of the intuition that if Alessandra were to coldly fail to offer an apology, she would not only be failing to deliver something that is owed to her family but she would also be revealing a morally bad trait through her behavior.

Though this sketch of the relationship (or lack thereof) between attributional-responsibility and accountability-responsibility is admittedly incomplete, I believe it represents a promising starting point for clearing away conceptual confusion regarding ascriptions of moral responsibility, and this is not limited to cases of forgetting. I'll briefly mention three other potentially fruitful applications of distinguishing the concepts in this way.

Responsibility for implicit bias may be helpfully thought of as involving accountability without attributability in cases in which an agent does not minimally approve of her action that causes the a harm associated with acting in accordance with implicit bias. These agents nevertheless harm others through their actions, and perhaps in doing so they fail in their roles as citizens. As a result they owe something to the recipients of these harms.

Similarly, white people might have role-related responsibilities as citizens to help undo the effects of structural racism, even the effects that they personally had no role in bringing about.[173] Even though, as an individual, a particular white person might not have undertaken any attributable actions that would reveal objectionably racist traits, it might be that group-membership generates role-related responsibilities to actively mitigate the effects of structural racism. Omitting these actions, then, can make white people accountably-responsible for harms to black people. Of course, once awareness of this fact is raised for the white person's consideration, she can then become attributionally-blameworthy for actively ignoring it.[174]

[173] See Zheng (Forthcoming).

[174] Avia Pasternak makes a similar point when she says that "group members are not necessarily blame-responsible when their group acts badly (that would be determined in light of their own behaviour). But a group's [collective moral responsibility] affects group members in other ways: First, it generates feelings that are associated with moral wrong-doing, such as 'we shame' and regret for what the group has done, and possibly also a sense of personal shame for being associated with the group. Moreover, the [collective moral responsibility] of a group grounds certain task-responsibilities for its members, such as the duty to change their group's practices and norms, or the duty to share the burden of repairing the damage their group caused. When group members end up having these forward-looking duties, then a failure to comply with them would amount to a personal moral failure" (Pasternak [2011]).

Finally, consider also the various agency-impairing conditions that I have argued exempt agents from attributional-responsibility: from the person with Tourette syndrome who utters slurs to the person with misophonia who verbally lashes out every time someone chews gum. While these agents are subject to forces that undermine their agency, this does not give them license to harm people without explanation or apology, because this would be a display of a lack of regard for the role-related responsibilities they take on in relation to those their behavior affects.

Again, this sets an important limit on claims of being exempt from responsibility due to failures to meet the conditions of the Minimal Approval account. While it is crucial for agents to respond to the behavior for which they do meet these conditions and so are attributionally-blameworthy, this does not exhaust the extent of what is required of their humanity in terms of responding to the harms they have a duty to ameliorate, whether they come about by choice or chance.

Bibliography

Adams, Robert Merrihew (1985). Involuntary sins. *Philosophical Review* 94 (1):3-31.

Arpaly, Nomy (2002). *Unprincipled Virtue: An Inquiry Into Moral Agency*. Oxford University Press.

Arpaly, Nomy & Schroeder, Timothy (1999). Praise, Blame and the Whole Self. *Philosophical Studies* 93 (2):161-188.

Arpaly, Nomy & Schroeder, Timothy (2013). *In Praise of Desire*. OUP.

Bell, Macalester. "The Standing to Blame: A Critique." In Justin Coates, D. & A. Tognazzini, Neal (2013). *Blame: Its Nature and Norms*. Oxford University Press USA.

Björnsson, Gunnar & Pereboom, Derk (forthcoming). Traditional and Experimental Approaches to Free Will and Moral Responsibility. In Justin Sytsma & Wesley Buckwalter (eds.), *Companion to Experimental Philosophy*. Blackwell.

Bonevac, Daniel, Dever, Josh & David Sosa, and (2006). The conditional fallacy. *Philosophical Review* 115 (3):273-316.

Bratman, Michael E. (1996). Identification, Decision, and Treating as a Reason. *Philosophical Topics* 24 (2):1-18.

Bratman, Michael E. (1999). *Faces of Intention: Selected Essays on Intention and Agency*. Cambridge University Press.

Bratman, Michael E. (2003). A desire of one's own. *Journal of Philosophy* 100 (5):221-42.

Braut, Jennifer J. (2018). "Investigating Misophonia: A Review of the Empirical Literature, Clinical Implications, and a Research Agenda." *Frontiers in Neuroscience.* 12

Buss, Sarah (2012). Autonomous Action: Self-Determination in the Passive Mode. *Ethics* 122 (4):647-691.

Buss, Sarah. (1994). "Autonomy Reconsidered." *Midwest Studies in Philosophy* 19(1), 95- 121.

Cane, Peter (2016). Role Responsibility. *The Journal of Ethics* 20 (1-3): 279-298.

Clarke, Randolph (2003). *Libertarian Accounts of Free Will*. Oxford University Press USA.

Clarke, Randolph (2009). Dispositions, Abilities to Act, and Free Will: The New Dispositionalism. *Mind* 118 (470):323-351.

Clarke, Randolph (forthcoming). Free Will and Abilities to Act. In *Streit um die Freiheit: Philosophische und theologische Beiträge*. Paderborn: Schoeningh/Brill.

Clarke, Randolph, McKenna, Michael & Smith, Angela M. (2015). *The Nature of Moral Responsibility: New Essays*. Oxford University Press.

Coates, D. Justin & Tognazzini, Neal A. (2013). The Contours of Blame. In D. Justin Coates & Neal A. Tognazzini (eds.), *Blame: Its Nature and Norms*. Oxford University Press. pp. 3-26.

David Lewis "Dispositional Theories of Value II" in Smith, Michael, David Lewis, and Mark Johnston. "Dispositional Theories of Value." *Proceedings of the Aristotelian Society, Supplementary Volumes* 63 (1989): 89-174.

Doris, John M. (2002). *Lack of Character: Personality and Moral Behavior*. Cambridge University Press.

Doris, John M. (2015). *Talking to Our Selves: Reflection, Ignorance, and Agency*. Oxford University Press.

Earp, Brian D. & Trafimow, David (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology* 6.

Edelstein, Miren et al. "Misophonia: Physiological Investigations and Case Descriptions." *Frontiers in Human Neuroscience* 7 (2013), 296.

Enoch, David (2017). Hypothetical Consent and the Value of Autonomy. *Ethics* 128 (1):6-36.

Fara, M. (2008). Masked Abilities and Compatibilism. *Mind* 117 (468):843-865.

Feinberg, Joel (1962). Problematic responsibility in law and morals. *Philosophical Review* 71 (3):340-351.

Fischer, John Martin (2010). The Frankfurt cases: The moral of the stories. *Philosophical Review* 119 (3):315-336.

Fischer, John Martin & Tognazzini, Neal A. (2011). The Physiognomy of Responsibility. *Philosophy and Phenomenological Research* 82 (2):381-417.

Fischer, John Martin. (2012). "Responsibility and Autonomy: The Problem of Mission Creep." *Philosophical Issues* 22(1), 165-184.

Fischer, John Martin. (2012). "Semicompatibilism and Its Rivals. *Journal of Ethics* 16 (2), 117-143.

Frankfurt, Harry (1987). Identification and Wholeheartedness. In Ferdinand David Schoeman (ed.), Responsiblity, Character, and the Emotions: New Essays in Moral Psychology. Cambridge University Press

Frankfurt, Harry (1992). The Faintest Passion. *Proceedings and Addresses of the American Philosophical Association* 66 (3):5-16.

Frankfurt, Harry G. (1969). Alternate Possibilities and Moral Responsibility. *Journal of Philosophy* 66 (23):829.

Frankfurt, Harry G. (1971). Freedom of the will and the concept of a person. *Journal of Philosophy* 68 (1):5-20.

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

Frankfurt, Harry G. (2006). *Taking Ourselves Seriously & Getting It Right*. Stanford University Press.

Gorman, A.G. (Unpublished Manuscript.) "Depression's Threat to Self-Governance." University of Southern California, Department of Philosophy.

Haji, Ishtiyaque. (2002). "Compatibilist Views of Freedom and Responsibility." In Robert H. Kane (ed.), The Oxford Handbook of Free Will. (Oxford: Oxford University Press.)

Haji, Ishtiyaque (1998). *Moral Appraisability: Puzzles, Proposals, and Perplexities*. Oxford University Press.

Haji, Ishtiyaque (2002). Compatibilist views of freedom and responsibility. In Robert H. Kane (ed.), The Oxford Handbook of Free Will. Oxford University Press

Haji, Ishtiyaque. (1998). *Moral Appraisability: Puzzles, Proposals, and Perplexities*. (Oxford: Oxford University Press.)

Hartman, Robert J. (2017). *In Defense of Moral Luck: Why Luck Often Affects Praiseworthiness and Blameworthiness*. Routledge.

Hieronymi, Pamela (2007). Rational capacity as a condition on blame. *Philosophical Books* 48(2):109–123.

Jaworska, Agnieszka. (2015). "Identificationist Views," In Meghan Griffith, Neil Levy, and Kevin Timpe (eds.), *Routledge Companion to Free Will*, (Abingdon: Routledge).

King, Matt (2014). Two faces of desert. *Philosophical Studies* 169 (3):401-424.

Kumar, Sukhbinder et al. "The Brain Basis for Misophonia." *Current Biology*27.4 (2017): 527–533. *PMC*. Web. 14 Sept. 2017.

Lawford-Smith, Holly & Collins, Stephanie (2017). Responsibility for states' actions: Normative issues at the intersection of collective agency and state responsibility. *Philosophy Compass* 12 (11):e12456.

Lavin, C. (2008) The politics of responsibility. University of Illinois Press, Champaign

Leckman, James F. and Donald J. Cohen. (1999). *Tourette's Syndrome—Tics, Obsessions, Compulsions: Developmental Psychopathology and Clinical Care* (Hoboken: Wiley-Blackwell), 27.

Levy, Neil (2011). Expressing who we are: Moral responsibility and awareness of our reasons for action. *Analytic Philosophy* 52 (4):243-261.

Lippert-Rasmussen, Kasper (2003). Identification and responsibility. *Ethical Theory and Moral Practice* 6 (4):349-376.

Mason, Elinor, (Provisionally Forthcoming) "Taking Responsibility." *In Oxford Studies in Agency and Responsibility Volume 6*, ed. David Shoemaker, Oxford University Press.

Matheson, Benjamin (2018). The Threat from Manipulation Arguments. *American Philosophical Quarterly* 55 (1): 37-50.

McKenna, Michael (2008). Compatibilism. *Stanford Encyclopedia of Philosophy*.

McKenna, Michael (2008). Frankfurt's argument against alternative possibilities: Looking beyond the examples. *Noûs* 42 (4): 770-793.

McKenna, Michael (2011). "Contemporary Compatibilism: Mesh Theories and Reasons-Responsive Theories." In R. Kane, ed., 2011, *Oxford Handbook of Free Will*, 2nd ed. (New York: Oxford University Press): 175-98.

McKenna, Michael (2012). *Conversation & Responsibility*. Oup Usa.

McKenna, Michael and Chad Van Schoelandt (2015). "Crossing a Mesh Theory with a Reasons-Responsive Theory." In A. Buckareff, C. Moya, and S. Rosell, *Agency and Responsibility* (Palgrave Macmillan).

McKenna, Michael and Chad Van Schoelandt. (2015). "Crossing a Mesh Theory with a Reasons-Responsive Theory." In A. Buckareff, C. Moya, and S. Rosell, *Agency and Responsibility* (Basingstoke: Palgrave Macmillan).

McKenna, Michael. (2011). "Contemporary Compatibilism: Mesh Theories and Reasons- Responsive Theories." In R. Kane, ed., 2011, *Oxford Handbook of Free Will*, 2nd ed. (New York: Oxford University Press), 175-98.

McKenna, Michael (2016). A Modest Historical Theory of Moral Responsibility. *The Journal of Ethics* 20 (1-3):83-105.

McKenna, Michael. (Forthcoming). "Watsonian Compatibilism." In *Oxford Studies in Agency and Responsibility*, Justin Coates and Neal Tognazzini, eds. Vol. 5.

Mele, Alfred R. (1992). *Springs of Action: Understanding Intentional Behavior*. Oxford University Press.

Mele, Alfred R. (1995). *Autonomous Agents: From Self-Control to Autonomy*. Oxford University Press.

Mele, Alfred R. (1999). Kane, luck, and the significance of free will. *Philosophical Explorations* 2 (2):96-104.

Mele, Alfred R. (2006). *Free Will and Luck*. Oxford University Press.

Miller, David (2004). Holding nations responsible. *Ethics* 114 (2):240-268.

Mitchell-Yellin, Benjamin (2014). In Defense of the Platonic Model: A Reply to Buss. *Ethics* 124 (2):342-357.

Mitchell-Yellin, Benjamin (2015). The Platonic model: statement, clarification and defense. *Philosophical Explorations* 18 (3):378-392.

Nelkin, Dana K. (2013). Moral Luck. *Stanford Encyclopedia of Philosophy*.

Pasternak, Avia (2011). The collective responsibility of democratic publics. *Canadian Journal of Philosophy* 41 (1):99-123.

Penelhum, Terence W. (1971). The importance of self-identity. *Journal of Philosophy* 68 (October):667-78.

Pereboom, Derk (2005). *Living Without Free Will*. Cambridge University Press.

Rescher, Nicholas, 1993, “Moral Luck”, in *Moral Luck*, D. Statman (ed.), Albany: State University of New York Press.

Richards, Norvin, 1986, “Luck and Desert”, *Mind*, 65: 198–209; page reference is to the reprint in Statman 1993b.

Robichaud, Philip & Wieland, Jan Willem (eds.) (2017). *Responsibility - The Epistemic Condition*. Oxford University Press.

Robin, Zheng (2016). Attributability, Accountability, and Implicit Bias. In Jennifer Saul & Michael Brownstein (eds.), *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*. New York: Oxford University Press. pp. 62-89.

Robin, Zheng (2016). Attributability, Accountability, and Implicit Bias. In Jennifer Saul & Michael Brownstein (eds.), *Implicit Bias and Philosophy, Volume 2: Moral Responsibility, Structural Injustice, and Ethics*. New York: Oxford University Press. pp. 62-89.

Rosebury, Brian, 1995, “Moral Responsibility and Moral Luck,” *Philosophical Review*, 104: 499–524.

Rosen “The Alethic Conception of Moral Responsibility” nature of moral responsibility.

Rosen, Gideon(2004).Skepticism about moral responsibility. *Philosophica l Perspectives* 18 (1):295–313.

Sartorio, Carolina (2016). A Partial Defense of the Actual-Sequence Model of Freedom. *The Journal of Ethics* 20 (1-3):107-120.

Scanlon, T. M. (1998). *What We Owe to Each Other*. Belknap Press of Harvard University Press.

Scanlon, T.M. (2008). *Moral Dimensions: Permissibility, Meaning, Blame*, Cambridge, MA: Harvard University Press.

Scanlon, T.M. (2013) “Interpreting Blame”, in *Blame: Its Nature and Norms.* Ed. Coates & Tognazzini: 84–99.

Schmidtz D., and Goodin R.E. (1998) Social Welfare and Individual Responsibility: for and against. Cambridge University Press, Cambridge

Schroeder, Timothy & Arpaly, Nomy (1999). Alienation and externality. *Canadian Journal of Philosophy* 29 (3):371-387.

Schroeder, Timothy. (2005). Moral Responsibility and Tourette Syndrome. *Philosophy and Phenomenological Research* 71(1), 106–123.

Shabo, Seth (2015). More Trouble with Tracing. *Erkenntnis* 80 (5):987-1011.

Sher, George (2005). *In Praise of Blame*. Oup Usa.

Sher, George (2009). *Who Knew?: Responsiblity Without Awareness*. Oxford University Press USA.

Shoemaker 2012 http://peasoup.typepad.com/peasoup/2012/06/survey-says.html

Shoemaker 2015a “Ecumenical Attributability”

Shoemaker, David (2009). Responsibility and disability. *Metaphilosophy* 40 (3-4):438-461.

Shoemaker, David (2013). Qualities of will. *Social Philosophy and Policy* 30 (1-2):95-120.

Shoemaker, David (2015). *Responsibility From the Margins*. Oxford University Press.

Shoemaker, David W. (2003). Caring, identification, and agency. *Ethics* 114 (1):88-118.

Shoemaker, David. (2003). "Caring, Identification, and Agency." *Ethics* 114 (1):88-118.

Shoemaker, David. (2011). "Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility." *Ethics* 121(3), 602-632.

Shoemaker, David. (2015). "Ecumenical Attributability" in *The Nature of Moral Responsibility* ed. Randolph Clarke, Michael McKenna, and Angela Smith. (Oxford: Oxford University Press.)

Silverstein, Matthew (2017). Ethics and Practical Reasoning. *Ethics* 127 (2):353 - 382.

Smith, Angela M. (2005). Responsibility for attitudes: Activity and passivity in mental life. *Ethics* 115 (2):236-271.

Smith, Angela M. (2008). Character, blameworthiness, and blame: comments on George Sher's In Praise of Blame. *Philosophical Studies* 137 (1):31-39.

Smith, Angela M. (2012). Attributability, Answerability, and Accountability: In Defense of a Unified Account. *Ethics* 122(3):575-589.

Smith, Angela M. (2008). Control, responsibility, and moral assessment. *Philosophical Studies* 138 (3):367 - 392.

Smith, Angela M. (2015). Responsibility as Answerability. *Inquiry : An Interdisciplinary Journal of Philosophy* 58 (2):99-126.

Smith, Michael (2003). Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion. In Sarah Stroud & Christine Tappolet (eds.), *Weakness of Will and Practical Irrationality*. Oxford: Clarendon Press. pp. 17-38.

Sripada, Chandra. "At the Center of Agency, the Deep Self"
https://umich.app.box.com/s/vlknn6s4ggnc7tuefft2

Sripada, Chandra (2016). Self-expression: a deep self theory of moral responsibility. *Philosophical Studies* 173 (5):1203-1232.

Sripada, Chandra (2017). Frankfurt's Unwilling and Willing Addicts. *Mind* 126 (503):781-815.

Strabbing, Jada Twedt (2016). Attributability, weakness of will, and the importance of just having the capacity. *Philosophical Studies* 173 (2):289-307.

Strabbing, Jada. *Moral Responsibility: Attributability, Accountability, and Capacities.* Dissertation. Princeton University. (2011).

Strawson, Peter F. (1962). Freedom and resentment. In Gary Watson (ed.), *Proceedings of the British Academy, Volume 48: 1962*. Oup Oxford. pp. 1-25.

Talbert, Matthew (2012). Accountability, Aliens, and Psychopaths: A Reply to Shoemaker. *Ethics* 122 (3):562-574.

Talbert, Matthew (2016). *Moral Responsibility: An Introduction*. (Cambridge: Polity).

Talbert, Matthew (Forthcoming). "Omission and Attribution Error" in D. K. Nelkin and S. C. Rickless (eds), *The Ethics and Law of Omissions* (OUP).

Taylor, James Stacey (2003). Autonomy, duress, and coercion. *Social Philosophy and Policy* 20 (2):127-155.

Thomson, Judith Jarvis, 1993, "Morality and Bad Luck", in *Moral Luck*, D. Statman (ed.), Albany: State University of New York Press.

Timpe, Kevin; Griffith, Meghan & Levy, Neil (eds.) (2017). *The Routledge Companion to Free Will*. New York: Routledge.

Tognazzini, Neal A. (2011). Understanding Source Incompatibilism. *Modern Schoolman* 88 (1/2):73-88.

Vargas, Manuel (2005). The Trouble with Tracing. *Midwest Studies in Philosophy* 29 (1):269-290.

Vargas, Manuel (2011). Revisionist Accounts of Free Will: Origins, Varieties, and Challenges. In Robert Kane (ed.), *Oxford Handbook on Free Will, 2nd Edition*. Oxford University Press.

Velleman, David J. (1992). "What Happens When Someone Acts?" *Mind* 101(403), 461- 481.

Velleman, J. David (1992). What Happens When Someone Acts? *Mind* 101 (403):461-481.

Vihvelin, Kadri (2004). Free Will Demystified: A Dispositional Account. *Philosophical Topics* 32 (1/2):427-450.

Vihvelin, Kadri (2011). How to think about the free will/determinism problem. In Michael O'Rourke, Joseph Keim Campbell & Matthew H. Slater (eds.), *Carving Nature at its Joints: Natural Kinds in Metaphysics and Science*. MIT Press. pp. 314--340.

Vihvelin, Kadri (2013). *Causes, Laws, and Free Will: Why Determinism Doesn't Matter*. Oup Usa.

Vihvelin, Kadri. (1994). "Are Drug Addicts Unfree?" In S. Luper-Foy and C. Brown (eds.) *Drugs, Morality and the Law* (New York: Garland), 51-78.

Watson, Gary (1975). Free agency. *Journal of Philosophy* 72 (April):205-20.

Watson, Gary (1996). Two Faces of Responsibility. *Philosophical Topics* 24 (2):227-248.

Watson, Gary (1999). Disordered Appetites: Addiction, Compulsion and Dependence. In Jon Elster (ed.), Addiction: Entries and Exits. Russell Sage Publications

Watson, Gary (2004). *Agency and Answerability: Selected Essays*. Oxford University Press.

Williams, Bernard & Nagel, Thomas (1976). Moral Luck. *Aristotelian Society Supplementary Volume* 50 (226):115 - 151.

Wolf, Susan (1987). Sanity and the Metaphysics of Responsibility. In Ferdinand David Schoeman (ed.), *Responsibility, Character, and the Emo-*

tions: New Essays in Moral Psychology. Cambridge University Press. pp. 46-62.

Wolf, Susan (1990). *Freedom Within Reason*. Oxford University Press.

Wolf, Susan (2015). Responsibility, Moral and Otherwise. *Inquiry : An Interdisciplinary Journal of Philosophy* 58 (2):127-142.

Zheng, Robin (forthcoming). What is My Role in Changing the System? A New Model of Responsibility for Structural Injustice. *Ethical Theory and Moral Practice*: 1-17.

Zimmerman, Michael J. (1986). Negligence and moral responsibility. *Noûs* 20 (2):199-218.