UNIVERSITY OF CALIFORNIA, SAN DIEGO

The Epistemology of Moral Responsibility

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Philosophy

by

Henry Argetsinger

Committee in charge:

Professor Manuel Vargas, Chair
Professor Matthew Fulkerson
Professor Dana Nelkin
Professor Caren Walker

2023

The Dissertation of Henry Argetsinger is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

University of California San Diego

2023

TABLE OF CONTENTS

LIST OF FIGURES

ACKNOWLEDGEMENTS

The writing of a doctoral dissertation cannot be done alone, despite how it may, on occasion, feel. There have been many hands that helped lighten the load along the way - too many to mention or remember. Still, I will try to give a rough accounting of the people that were most influential in the completion of this dissertation.

For inspiring me to pursue philosophy that "matters" I have Heath Massey and Matt Tedesco to thank. Their guidance as an undergraduate at Beloit College convinced me that asking philosophical questions could have some importance in the world, and that whatever that importance was had nothing to do with institutional prestige or what was in or out of vogue. Indeed, it was their lack of effort to "professionalize" me that I credit with allowing me to exist in the field. I saw that good human beings could pursue philosophy as a way to teach and make change in the world.

For undoing some of their good work, and preparing me to enter the ranks of academia as an actual professional, I have my mentors (and fellow students) at University of Wisconsin Milwaukee to thank. Among them were Luca Ferrero, Andrea Westlund, Kirun Sankuran, Craig Kneply and William Albuquerque. The transition from a small liberal arts college to a cut-throat Masters program was made possible only through their friendship and guidance – especially when I realized that I had no idea what I was talking about, philosophically speaking.

At UC San Diego I had Cory Davia, Matt Braich, Amy Berg, Craig Agule, and Noel Martin to shepherd me into the "big leagues." Knowing that Cory had successfully made the move from UWM made the endeavor seem possible. Knowing that Matt, Craig, and Noel were cool and decent human beings with non-philosophical interests made me feel at home. Of course, Anne Thresher and Leonardo Moauro made me feel quite literally at home, offering up their couch to me multiple nights a week during my years of commuting from Orange County. Will did the same, once he eventually followed from UWM. Their friendship, along with Kathleen Connelly, Andrew Bollhagen, Rosalind Chaplin, Shawn Wang, Daniel Cohen, and Juan Carlos González's sustained me during some of the most difficult years of my life. Above all, Richard Vagnino's friendship has meant more to me over the past six years than any other - and his ideas have undoubtedly made their way into these pages.

Thank you to David Brink and Clinton Tolley for being on earlier constitutions of my committee, and helping me in numerous and invaluable ways to work out some of the basic ideas at the core of my views. And lastly, and most importantly, thank you to my committee members. Caren Walker, for serving as a model of interdisciplinary work (and agreeing to join my committee at the final hour), Dana Nelkin for her unfailing optimism even during the shakiest parts of my journey through the program, Matt Fulkerson for his sage advice about philosophy, life, and how to balance the two, and especially to Manuel Vargas, without whom this would not have been possible. I cannot begin to thank Manuel enough for his willingness to read draft after draft, his guidance and candor about navigating the profession in particular and academia in general, and his modelling of what good teaching and mentorship can look like.

Portions of Chapter 1 appear in publication as "What's the Relationship between the Theory and Practice of Moral Responsibility?", *Humana Mente* 15 (42): 29-62, 2022. This paper was co-authored by the dissertation author and his dissertation chair, Manuel Vargas.

Portions of Chapter 2 have appeared in publication as "Blame for Me and Not for Thee," in *Ethical Theory and Moral Practice* 25: 265-282, 2022. The dissertation author was the primary investigator and author of this paper.

Portions of Chapter 3 have appeared in publication as "Blame for Me and Not for Thee," in *Ethical Theory and Moral Practice* 25: 265-282, 2022. The dissertation author was the primary investigator and author of this paper.

VITA

2010	B. A. in Philosophy *magna cum laude*, Beloit College
2015	M. A. in Philosophy, University of Wisconsin Milwaukee
2023	Ph. D. in Philosophy, University of California San Diego

PUBLICATIONS

Henry Argetsinger, “Blame for Me and Not for Thee”, *Ethical Theory and Moral Practice* 25: 265-282, 2022.

Henry Argetsinger, Manuel Vargas, “What’s the Relationship between the Theory and Practice of Moral Responsibility?”, *Humana Mente* 15 (42): 29-62, 2022

ABSTRACT OF THE DISSERTATION

The Epistemology of Moral Responsibility

by

Henry Argetsinger

Doctor of Philosophy in Philosophy

University of California San Diego, 2023

Professor Manuel Vargas, Chair

In this dissertation I argue that philosophical theorizing about moral responsibility has not paid sufficient attention to the epistemic dimensions of our practices of responsibility. This dissertation asks how we pursue *justification* and *agreement* about moral responsibility in our actual social practices, and, finding those pursuits often less than ideal, further asks how we *ought* to do so. In chapter one I seek to vindicate these claims by reviewing the extant moral responsibility literature, pointing out at various junctures what I call "paths not taken" and "epistemic lacunae." In chapter two I engage with work in contemporary social, cognitive, and moral psychology to sketch a picture of how we actually pursue judgments about responsibility, psychologically speaking. With this "ping-pong" model in hand, I identify some particularly common and robust "epistemic disruptors" that undermine our ability to hold one another responsible in the ways which philosophical theories often imagine.

In particular, I focus on the way in which blame judgments are susceptible to these disruptions.

Given this set up, chapters three and four of the dissertation deal with some practical upshots for the ethics of blame. In chapter three I argue for introducing the norm of Powerful Restraint, which suggests that in contexts where there are large gaps in social power, the powerful should refrain from blaming downward. In chapter four I further motivate this kind of revisionism and asymmetry, focusing on the ways in which our currently constituted practices of responsibility intersect with issues in the contemporary literatures on pragmatic encroachment and cultivated ignorance.

Chapter 1

Moral Responsibility's Epistemic Lacuna

The judgment that a person is morally responsible for her behavior involves—at least to a first approximation—attributing certain powers and capacities to that person, and viewing her behavior as arising (in the right way) from the fact that the person has, and has exercised, these powers and capacities.

Matthew Talbert[1]

[1] Talbert, 2022

1.1 Introduction

This project is about the everyday epistemology of moral responsibility. In particular, it's about the trouble we run into in our attempts to form judgments about agents' capacities for (and instantiations of) morally responsible action. It asks how we pursue justification and agreement about moral responsibility in our actual social practices, and, finding those pursuits often less than ideal, further asks how we *ought* to do so.[2] My focus, then, is squarely on the architecture and "internal" logic of the judgements we make about participants in our responsibility practices, rather than on what further facts might ground our judgements or make them true. My project makes three core claims. First, that, in everyday life, our epistemic practices of determining responsibility are subject to myriad difficulties, distortions, and misfirings. We *can* reliably justify our beliefs about one anothers' responsibility, but it is not an effortless, straightforward task. Most of the time, I claim, we are overconfident about our responsibility ascriptions. We often lack the kind of information and evidence we'd need (by the lights of the best metaphysical theories of responsibility) to make whole-hearted and confident claims about responsibility.

Second, and relatedly, I claim that the epistemology of moral responsibility is fundamentally *interpretive*. That is, gaining knowledge about agents' responsibility requires interpretive work, whether that work is in effortful, conscious reflection, interpersonal (or society-wide) dialogue, or the more or less automatic structuring of our experiences through scripts, schemas, narratives and social cues. The fact that our belief formation requires interpretive work helps explain the many ways in which it is all too easy for those beliefs to become distorted, confused, ideologically driven,

[2] In this sense, I align myself with work now popular in the "social epistemology" literature. Rather than being primarily motivated by traditional epistemic questions concerning the possibility and foundations of knowledge, this work tackles such issues as epistemic injustice, norms of testimony, and peer disagreement. Social approaches take it as a given that epistemology is normative, and that the work of figuring out how agents reason in non-ideal, complex and robustly interactive situations is an important epistemic enterprise. Dotson (2014), Fricker (2007a), and Manne (2018) are representative recent examples of this kind of project.

or biased.

The final claim is that this interpretation is done in irreducibly practice-based, and normatively-laden social settings. Recognizing the social nature of the epistemology of responsibility centers the fact that the practice-based norms structuring our interpretations (and the responses those interpretations are likely to receive) are deeply involved in determinations and ascriptions of responsibility. Whether and how I come to think of you as responsible often has as much to do with interpretation of your character, actions, and history in particular contexts (against the backdrop of thick, contextual social norms), as with your actual psychological or metaphysical capacities.

My aim is to argue for these core claims and work through some of their implications, particularly for our norms and practices of blame. The dissertation is roughly split into two parts. In Chapters One and Two I engage in descriptive work: In the rest of this chapter, I survey the contemporary literature on moral responsibility, arguing that it fails to take seriously the epistemic dimensions I am interested in. In Chapter Two, I advance a model of how attributions of responsibility proceed in day-to-day practice. Based on this model I show that there are many "epistemic disruptions" that contemporary theorists ought to pay more attention to.

The second half of the dissertation deals with some practical and normative upshots of this theorizing. In Chapter Three I argue for a norm of reduced blame from positions of social power. In Chapter Four I further justify this asymmetry in positional blame, and argue that cultivated ignorance, social signalling, and a species of pragmatic encroachment all deserve more attention in our theorizing about blame and moral responsibility. The driving aim of the dissertation as a whole is to work out some of the consequences that taking seriously the epistemic dimensions of responsibility has for the prospects of improving both the theories and practices of moral responsibility.

1.1.1 The Lacuna

The task of this chapter is to trace the shape of a problem for theorists of moral responsibility. I'm going to argue that theorists are missing out on something vital by failing to attend to and adequately theorize about the everyday epistemology of moral responsibility. Contemporary theories of moral responsibility haven't often focused on epistemology – especially not the epistemology of everyday agents grappling with responsibility concepts in day-to-day life. Insofar as epistemology *is* discussed, it is often in the context of the "epistemic" or "knowledge condition" *on* responsibility. This is the idea that there is some sense in which agents must be aware of what they are doing in order to be responsible for their actions. There is a rich literature here which concerns issues such as doxastic responsibility, culpability for ignorance (both moral and non-moral), negligence, and tracing, among much else.[3] Some of this work is indirectly relevant for my purposes, but it does not center the epistemic processes that result in judgements about one anothers' responsibility.

In some sense, this landscape is unsurprising. The recent history of moral responsibility, as much as any other academic conversation, is path-dependent. That is, we can easily imagine a nearby world where the contemporary literature on responsibility focused first on epistemic questions, rather than on metaphysical ones. We can imagine it proceeding for twenty or thirty years, and robust literatures spinning-off and surrounding it. These literatures would tackle questions of interpretation, justification, social scaffolding, normative authority and much else. In this nearby world, we can then imagine a metaphysician of responsibility saying, "this is all well and good, but shouldn't we also ask about the metaphysics of responsibility? We have very good theories of when and how we *judge* someone to be responsible, but wouldn't it be good to articulate the kinds of capacities that would make those

[3] For some representative examples, see Bjornsson (2017), Clarke (2014), Fischer and Tognazzini (2009), Ginet (2000), Guerrero (2017), Haji (1997), Harman (2011), Levy (2007), Mele (2010), Sher (2009), and Vargas (2005).

judgments true?” Surely, this would be a sensible question.

In our actual world, of course, the recent history of work on moral responsibility is a mirror image, charting a course from debates about free-will, to questions of personhood and causal sourcehood, and thence to questions of blame and forgiveness (among much else). The question I pose is: given our very good theories of what it would take for an agent to *be* responsible, wouldn’t it be helpful to articulate a theory of how we come to judge, know, and pursue agreement about whether agents meet those conditions? My point is, in other words, not that path-dependence shows some *fault* in the way the literature has developed – only that certain questions have gone relatively un-examined, and that we can imagine a different path where they might have been more central.

My arguments, then, are friendly to the metaphysics of responsibility: up to a point. I *do* argue that the non-centrality of the epistemology of moral responsibility in the contemporary literature is problematic insofar as it has led to a a kind of epistemic injustice that surrounds our conception(s) of moral responsibility. We have a lack of concepts, language, and resources to discuss the ways in which agents make real-world responsibility judgements, and a lack of tools and theoretical machinery to determine how they might make those judgements in better and worse ways. It is unhelpful to assess our everyday practices as if agents are attempting to do the kind of work moral responsibility theorists do – that is not what we are up to when we apportion blame and praise. As is often the case, such a lack of conceptual machinery leads to downstream injustice as well: the fact that we don’t have robust theories of better and worse responsibility interpretation means that it is hard for us to notice when things are going wrong, and, therefore, hard for us to pursue ameliorative projects.

Before I can convincingly articulate these claims, I’ve got to convince you that there’s a hole where I say there is - and that the hole matters. Have contemporary theorists really neglected to give an account of the epistemology of moral responsibility? And, even if they have, why should we care? Why isn't a focus on metaphysics good enough? In this chapter, I seek to motivate the project by answering these questions. To do so, I probe what I'll call the epistemic "paths not taken" in responsibility theorizing. I do so by looking, primarily, at compatibilist theories of moral responsibility, including those of Brink and Nelkin (2013), Fischer and Ravizza (1998), Fischer (2012), McKenna (2012), Scanlon (2008), and Vargas (2013).

I choose compatibilist theories focused on reasons-responsiveness and quality of will for two reasons. First, I find them to be the most plausible extant theories of responsibility. Second, and much less controversially, they deal most directly with the social, interactive, practice-based nature of responsibility. Libertarian theories, in focusing almost exclusively on the metaphysics of free-will, just won't provide the right kind of friction for my claims. And of course anti-compatibilist Hard Determinists won't think it's *ever* correct to attribute responsibility, so their epistemology will be an error theory. In other words, if an epistemology of responsibility has been explored, we'd expect to find it in the compatibilist literature.

In surveying the literature I'll argue two things: First, that the theories often pay no, or very little attention to the epistemic dimensions of our moral responsibility practices. Second, that what attention *is* paid ignores the significantly *interpretive* nature of that epistemology, instead advancing a kind of implicit "metaphysics first" reading of how we come to judgments of and pursue justification for our claims about responsibility. In doing so, as I've said, I seek out the various epistemic paths that contemporary debates have *not* taken. Rather than viewing such lacuna prodding as criticism of the authors in question, we should view it as an invitation to imagine future paths forward for each family of views.

I'm going to identify the following four areas of under-exploration, rich for opportunity to say more about the epistemic dimensions of responsibility:

1. **Historicism**: Roughly, whether and how an agent's history bears on our ascriptions of her responsibility at a moment.

2. **Contextualism**: Roughly, whether and how the larger context of an agent's situation bears on our ascriptions of her responsibility.

3. **Conversation**: The question of whether responsibility is essentially communicative, and if so, how this conversation relies on interpretations of agent-meaning.

4. **Sociality**: The ways in which the social nature of our responsibility practices affect our understanding of one anothers' responsibility.

Of course, all four of these general areas get *some* uptake in the contemporary literature. Because they are closest to my concerns, they are useful starting points for exploring epistemic paths not taken.

1.1.2 The Anti-Contextualist and the Narrativizer

I said that my first task is to convince you that there *is* a hole, and that it has some motivational oomph for theorizing about responsibility. To trace the hole's edges, consider the following toy example, and an accompanying pair of statements suggesting two very different kinds of reasoning about responsibility.[4]

Murky Matthew:

> Matthew is an eighteen year old, white, male senior at an affluent high-school. He is accused of selling marijuana to several of his friends. If

[4] The following are based, very roughly, on statements by Joe Biden and Judge James Troiano. See: `https://www.vox.com/policy-and-politics/2019/6/20/18677998/joe-biden-1994-crime-bill-law-mass-incarceration`: "The controversial 1994 crime law that Joe Biden helped write, explained," and: `https://www.nytimes.com/2019/07/02/nyregion/judge-james-troiano-rape.html?fbclid=IwAR0Yx6gvveRp5JyxzXh4r5lp2x3m_7UTeXJ6z-T7XN5Zs_t_mrTBo119jTY`: "Teenager Accused of Rape Deserves Leniency Because He's From a 'Good Family,' Judge Says," for the original statements. In order to avoid debates about whether the men were merely grandstanding for political reasons, or whether they are actually concerned with punishment rather than ascribing responsibility, I've made the cases more obvious. However, I include the examples to show that these two kinds of thinking are very common in everyday life.

> the accusations are true, it's clear that he is not a major drug-dealer, but also that he is a non-insignificant source of recreational drugs in his social circle. The event becomes well-known in the community, just as the high-school administrators are trying to decide how to hold Matthew accountable for his actions.

Two parents of teenagers at the high-school have the following reactions:

Sympathetic Narrativizer James:

> James knows that Matthew is a, "good kid;" "He's always treated my daughter like a gentleman! Plus he's supposed to go to Stanford next year. Sure, he ought to get a slap on the wrist, but nothing that would jeopardize his future." Not only that, James notes that: "As I understand it, his dad left a few years back, so he hasn't had a positive male role-model – we ought to be more forgiving because of that." And finally that, "Plus it's not like he's one of these gang-bangers selling crack! It's just a little weed, just some honors students blowing off steam – I don't see what the big deal is."

Anti-Contextual Joe:

> Joe doesn't know much about Matthew, and doesn't care to know. "He's to blame for his own actions - that's all there is to it. He chose to sell drugs at our school, and he knew the consequences. Do I care what made him do it? That his mom and dad split? Not at all! Do I care that he has a scholarship to Stanford? Well, maybe he needs to learn a lesson! None of that other stuff matters; it's not like he was brainwashed or held at gunpoint – he acted like an idiot and he's got no one to blame but himself."

Very clearly, these toy exemplars of responsibility attribution are taking different paths in arriving at judgments of responsibility (and the downstream judgements of punishment they may ground). Whereas James is interested in the larger historical (and forward looking) context of Matthew's case, Joe is not. It's also the case that James is sensitive to details of the case concerning Matthew's social position (in problematic ways!). Joe, on the other hand, is focused on the action(s) in question, to the exclusion of such details.

Which is the right way to think about responsibility? It isn't clear - and I don't propose to answer that question. Instead, I want us to notice the very different epistemic routes suggested by these remarks, and to notice that these are not the only routes available. If one is in James' camp, one wants to find out as much contextual information as possible - and this suggests a set of epistemic pitfalls. James might be far too swayed, for instance, by the fact that Matthew is a white honors student with a privileged future ahead of him – facts that seem irrelevant to whether Matthew is responsible for what he does. Perhaps too there is a kind of underlying misogyny at play that privileges justifications and excuses for men. Caring about context, in other words, is no guarantee of doing so responsibly.

If one is in Joe's camp, on the other hand, there are different kinds of epistemic difficulties to contend with. One might miss crucial contextual elements. *Is* the fact that Matthew's father left relevant in determining his responsibility? Perhaps not in this case, but one can imagine cases where it might be – and Joe is resolute that he won't let such factors influence his thinking. Or, in a kind of reverse of the bias we imagined James falling prey to above, Joe might be the kind of sexist that refuses to acknowledge that women are ever at a disadvantage – a kind of "gender-blind" ideologue who can never see how power asymmetries might be affecting a situation.

Finally, one can easily imagine that both men are unlikely, given what we know about the real world, to be entirely consistent in their epistemic practices. We can certainly imagine James' narratives becoming *unsympathetic* due to explicit and implicit biases, or picture him sliding towards Joe's context-free position when he under-rates the credibility of certain stories, epistemic peers, and sources of evidence that influence his judgment. Joe, of course, is no more likely to be in the clear in his attempts to go context free. In steadfastly attempting to ignore context he may merely be unaware (and unable to become aware) of when it *does* affect his judgement. His local responsibility attributions are in danger of reflecting current social arrangements and their legacies of injustice, with little opportunity to notice the ways something like a history of colonialism, slavery, or gendered oppression might affect a person or situation.

In other words, things are complicated. The upshot is just that, in making determinations of culpability and attributions of responsibility, we can take very different epistemic routes. I assure the reader that I'm not implying that over-narrativizing and under-contextualizing are not the only ways to arrive at responsibility judgements – there are many other routes or processes a responsibility practitioner can take. A critical question this dissertation raises is whether there are principled ways of assessing those routes, such that we can say one was better than the other, that some are always bad, that someone did well or poorly in determining the relevant factors and so on. I argue that, as a starting point, we can at least make clear some of the common ways that epistemic peers like James and Joe go astray, and how this is related to our extant responsibility practices. The toy examples above provide intuitive support for my arguments, whatever we think of their speakers' conclusions. One can read James and Joe's quotes, if one desires, as being about trying to determine the right psychological and metaphysical capabilities of the agents in question. When Joe says that, "I don't want to ask what made them do this," perhaps he means that denying a criminal's responsibility for their actions based on past hardship is to misunderstand their reasons-responsive capacities, which, he assumes, were fully intact during the crimes in question. But I find such a reading strained.

That is, when Joe says that, "I don't want to ask what made them do this," or James says that "He has that scholarship to Stanford," I read them as indicating that it is not the perpetrator's metaphysical capacities or states of mind they are focusing on. Joe doesn't care whether a criminal was reasons-responsive, or has a valid excuse, although these might be background assumptions. He cares about the *action* and its social meaning. And James doesn't care about culpability, he is interested in a larger narrative about the agent's life; a narrative which leads to an interpretation of the individual as a "good young man" capable of great things. The fact that we might think both of these are rather bad interpretations shouldn't cloud the point I'm making, which is just that they *are* interpretations (and that we are inclined to assess them as sub-optimal). So, I'm arguing, we ought to focus on the epistemic processes that have led them to bypass questions of metaphysical responsibility.

1.1.3 Why not Epistemology, Psychology, or Sociology?

Before we get too deep into the weeds, let me address another worry. It might be argued that, if what I'm after is evidence of philosophers treating seriously the epistemic dimensions of moral responsibility, then I am looking in the wrong place. Wouldn't such work naturally find a home in the discipline of epistemology? I will not trouble the reader with a lengthy survey of the extant literature, but instead give a simpler answer. It is true that there are some epistemologists thinking about blame and responsibility. [5] However, we are here presented with the mirror image of the lacuna I've begun tracing. In her discussion of culpable vs. non-culpable ignorance, for example, Miranda Fricker (2007) cites Bernard Williams on the internal/external reasons debate, and no one else (100-104). There is, in other words, no substantive discussion of the contemporary moral responsibility literature. Of course, this is in no way to impugn her project – it is simply to point out the fact that these two literatures have yet to deal with one another in a very substantial way.

Moreover, unless we have a clear sense of how these disciplines connect, there isn't much hope for uptake about the importance of their connection. A similar point can be made about why I'm not pursuing a project in sociology or psychology, for instance. One natural kind of project would be to go out and do some empirical field

[5] Miranda Fricker, for example, is interested in the nature of testimonial justice and injustice, and asks whether, when, and how much those who display the vice of testimonial injustice might be culpable for their lack of virtue. See Fricker (2007b) 100-104, for instance.

work – to try and determine the psychological or social mechanisms that are operative when we hold one another responsible. But of course, in order to do that kind of work, we need to know what concepts are important, what kinds of mechanisms to look for, what questions to ask. I think, given the paucity of attention to the epistemology of moral responsibility, that we don't yet have a clear sense of those concepts, mechanisms and questions. And, I take it, this is why my philosophical approach is worthwhile.

1.2 Reasons-responsive theories, historicism, and contextual control

Now that we have some sense of the aims, motivations, and methodology of my project, I can begin to make good on my claim that there really is something which the contemporary debate is missing. One area where responsibility theorists have made some contact with epistemic concerns is when they discuss the "historical" nature of responsibility. Many theorists hold that responsibility is a historical notion - one that depends not just on a time-slice of a current agent and their actions but on a wider view of the conditions that led an agent to be what they are and do what they do at a given time. However, I'm going to argue the following: first, insofar as responsibility theorists have dealt with history, they have done so in an almost exclusively metaphysical mode. Second, insofar as epistemic concerns are noted, they are often bracketed or put aside as problems to be solved at a later date or in a different area of philosophy.

1.2.1 Historicisim in Fischer and Ravizza

In order to make this argument, I'll begin with "reasons-responsive" conceptions of responsibility, as these theorists have most often claimed that responsibility is an essentially historical notion. John Martin Fischer is probably the most influential theorist of "reasons-responsive" responsibility, so it makes sense to start with Fischer and Mark Ravizza's classic account (1998). After all, if *they* have a meaningful theory of the epistemology of responsibility, then my assertion that there is a lacuna will be straightforwardly incorrect. However, I think it is obvious that they do not. Fischer and Ravizza explicitly discuss epistemology twice in *Responsibility and Control.* The first instance is to let the reader know that their book is *not* going to delve into the epistemic condition for responsibility (focusing instead on the control condition). Clearly, then, we won't find an epistemic theory there. The second discussion concerns what they call "epistemic historicism" (13, and 188-194, respectively). This is more promising, and I will discuss "epistemic historicism" in a moment, as it offers a concrete glimpse of a road not taken. First, a brief reminder of the Fischer and Ravizza view of responsibility.

Fischer and Ravizza argue that their theory is "Strawsonian."[6] Strawson (1989) argued that our practices of holding one another responsible are centered on the "reactive attitudes" – those moral emotions (such as indignation, resentment, and forgiveness) that we cannot help but have towards fellow practitioners. On his account, these emotions are central to the meaning and purpose of responsibility practices, and an agent's aptness to be a target of the reactive attitudes tells us something about her membership in the category "responsible agent." As Fischer argues, "To be morally responsible for one's behavior is to be an apt target for what Peter Strawson called the 'reactive attitudes'—and certain associated practices...[including] moral praise and blame, and reward and punishment" (2006, 106). And being an apt target for such attributions involves guiding one's "behavior in the way characteristic of agents who act freely" (106). This characteristic freedom, on the reasons-responsive view, involves an agent's *relationship to reasons.* Fischer and Ravizza hold that we ought

[6] In some sense I am doubtful that this is a meaningful claim, insofar as I doubt there is any one unified approach that can consistently be found under the banner of Strawsonian-ism.

to think about responsibility in terms of whether an agent acted via a "mechanism" which was sufficiently responsive to reasons.

There are two kinds of relationships an agent's responsibility relevant mechanism might stand in to reasons (moral and otherwise): Receptivity and Reactivity. Receptivity is, as the name implies, our receptiveness to the reasons available to us – it is our ability to *recognize* what reasons exist in a given context. This can be read as a kind of epistemic or knowledge condition on responsibility. If an agent is receptive, we can ask further whether they are *reactive* to reasons: whether they respond to the reasons that they recognize. If an agent is suitably reactive, then they should show a pattern of response to reasons that varies according to the strength of reasons – and such patterns should be intelligible. On Fischer and Ravizza's view, an agent has the kind of control necessary for responsibility if they show the characteristic kind of receptivity and reactivity to reasons that allows them to act in patterns which identify them as apt targets for reactive attitudes.

This set up allows us to specify whether an agent is reasons-responsive by considering nearby counterfactuals. We can show that an agent has the right kind of control and knowledge by arguing that: "holding fixed that mechanism, the agent would presumably choose and act differently in a range of scenarios in which he is presented with good reasons to do so" (187, Deep Control). In particular, Fischer and Ravizza are interested in "moderate reasons-responsiveness" – in our ability to recognize and respond to reasons at least some of the time, in patterns that are understandable and legible to third parties. Let me make clear that nowhere in this picture do we get a story about *how* we determine whether agents meet these criteria, or have these capacities. Again, it may be that Fischer and Ravizza have in mind an implicit epistemology on which we are just natural trackers of the relevant properties, and the capacities reveal themselves to us whenever we observe bits of agential behavior.

As I said, one area in which specifically epistemic concerns *are* mentioned in Fischer and Ravizza (1998) is their discussion of historicism. Fischer and Ravizza hold that responsibility is essentially historical, and see their reasons-responsive view as having an advantage over "mesh" theories (those theories which hold that responsibility obtains when an agent's internal states are related in the right kinds of ways) for this reason.[7] Their argument is that our sense of whether an agent is reasons-responsive can and should include reference to historical factors that would help prove or vitiate the possession of these capacities (1998, 187-188). Historical factors matter even when the current time-slice state of an agent's capacities and mental states are arranged in the right kinds of ways such that we'd normally call such an agent responsible. This is so because historicism allows us to notice factors ranging from the obvious (coercion and incapacitation, for instance), to the fuzzy and contentious: certain facts about one's childhood, moods, or situational factors, for example, which are often relevant in our responsibility practices. Fischer and Ravizza admit that this is complicated and shaky ground. What exactly counts (and when it counts) is up for debate, and, indeed, a large debate has bloomed around these questions of causal sourcehood, manipulation, and tracing (See for instance: Agule (2016), Deery and Nahmias (2017), Mele (2010), and Vargas (2005))

Now for my two arguments. First, the subsequent literature has resolutely focused on the metaphysical implications of historicism. The discussion in the literature on tracing and manipulation concerns whether we can find the right kind of causal source-hood for actions in previous time-slices of the same agent (or whether there are deviant causal chains, or whether the agent's capacities and mental states are such that they could have satisfied a knowledge condition). Again, this kind of path dependence in the literature can mislead us into thinking that the only questions worth asking here are metaphysical ones. *One* interesting kind of question we can ask is what we'd make of a human being whose zygote was designed by a powerful goddess such that it's unclear whether they have the right kind of free will for responsibility

[7] See, for instance, Frankfurt (1969)

(Deery and Nahmias, 2017). But another question we can ask is simply: how are we supposed to parse and take into account historical factors in our attempts to form good judgments about agential responsibility? What role should history play in the development of our beliefs about responsibility?

Once we ask these questions, we can see that their answers depend on epistemic and normative considerations that cannot be answered entirely by metaphysical theorizing. Consider: even if we have a fully fleshed out theory of the metaphysics of responsibility that tells us which capacities must be operative during reasons-responsive agency, when we go to look for evidence of those capacities they will be masked by "defeaters" like historical and situational context. The precise amount and type of history (causal? characterological? narrative?) to take into account will be fixed by interpretive as much as metaphysical considerations. If the metaphysics tells me that agents need situational control, I still need a theory of how to know who has situational control given the messiness and opacity of the actual world.

The second argument is that it's precisely the *bracketing* of the trickier questions of historical context which ought to give us pause as practical epistemologists. If our metaphysical theory isn't going to directly vindicate one account of historicism over another, then we will have to look elsewhere for answers. My claim is that whether and how history matters can be investigated in the epistemic realm without needing to deal with (some) deeper questions of metaphysics. To put it crudely, history matters when it affects our judgments of an agent's responsibility, right now. The basic insight that Fischer and Ravizza get at is that responsibility is historical – but this isn't necessarily true in any deeply metaphysical sense. Mesh theorists can (and do) give us intelligible theories of time-slice responsibility. If we think there is some further sense of responsibility that is important to us, it must be because of something other than metaphysical factors. If we allow that moral responsibility is a historical concept, and that part of it's metaphysical correctness conditions lie in questions of the proximal (and perhaps ultimate) origins of certain causal abilities, psychological states, or levels of capacity, we should see very clearly that we lack epistemic access to the facts that would often settle questions of responsibility. Our judgements just can't typically be reacting (directly, in any case) to the causes of an agent's psychological states that lie in the historical past. And if this is so, it's worth asking what we *are* reacting to. In other words, when we judge that history matters in a particular case, whether this is an epistemic virtue or vice seems to be a matter of epistemic norms, rather than metaphysical facts.

1.2.2 Narrative to the Rescue?

I'll conclude my brief review of Fischer and Ravizza by mentioning that, more recently, Fischer has begun to develop a "narrative" conception of reasons-responsiveness.[8] He is particularly interested in the concept of "taking responsibility," which involves: "seeing oneself in a certain way" such that one is prompted to take control of a situation (and thereby gains the kind of guidance control necessary for responsibility). This more subjective account of responsibility, insofar as it involves the concepts of narrative and "seeing as," looks like fertile ground for exploration of epistemic concerns.

Fischer makes clear that moderate reasons-responsiveness is necessary but insufficient for moral responsibility - for this we need what he terms "guidance control:" moderate reasons-responsiveness as well as "ownership" of the mechanism. This is so because we can imagine an agent who responds to reasons (who is receptive and reactive), but does so in a way which is vulnerable to mesh-theoretic, Frankfurt style counterexamples: the agent, for instance, whose mechanism has been implanted by brainwashing, or who always feels alienated by the decisions and actions that originate from it.

This kind of view might be latent in Fischer and Ravizza (1998). There, they ar-

[8] See, for instance Fischer and Tognazzini (2009).

gue that taking responsibility has three conditions: 1) an agent must see themselves as a source of behavior (see their choices and actions as efficacious in the world), 2) an agent must accept that they are a fair target of reactive attitudes (or that it is fair to be part of certain responsibility practices), and 3) they must base these views appropriately "on the evidence" (210-211). To this earlier sense of what would be required to "take responsibility" the later Fischer explicitly adds a sense of Frankfurtian "wholeness." This can be read, perhaps, simply as a further articulation of the three conditions.[9] One must have evidence that one's mechanism is one's own in the right kind of way such that they are the source of their behavior and it would be fair for others to hold them responsible.

Again, however, we should be clear that this is a metaphysical rather than an epistemic view, despite the discussion of evidence and the appeal to the way an agent "sees" themselves. A very natural reading would be that agents' judgements about the relationship they stand in to their actions, and wider third party judgements about those actions (and the agent's judgements) are what constitute part of the grounds of our reactive attitudes. But Fischer really intends this as a metaphysical claim.

I think this is a missed opportunity - a "path not taken" in the responsibility literature. To see why, return to the issue of "manipulation" very briefly touched on above. I will not have space to give the rich debates in the manipulation literature their due, but attending to recent developments draws out a few key points with regard to the lacuna I'm probing. Classic cases of manipulation go back to Frankfurt (1969) and his attack on the Principle of Alternate Possibilities. A key question to emerge from subsequent discussions is *in what sense* motivations, values, desires and so on must be *one's own* in order for one to be held responsible for acting on them.

Fischer (2012) has argued, not surprisingly, that the key question is whether one's

[9] It also provides an interesting connection with authorship views of action such as Korsgaard (2009) and Velleman (2009)

responsibility relevant mechanism remains under guidance control. He argues that one's mechanism is changed (such that it is no longer under guidance control, thus undermining responsibility) when inputs are "artificially" implanted just prior to action. Why? Because doing so means there is "no reasonable or fair opportunity... to reflect on or critically evaluate the new input in light of... standing dispositions, values, preferences, and so forth" (2012, 197). Fischer clarifies that it is not, primarily, an issue of timing. After all, we often don't have time to reflect on new inputs to our decision making process – we must act in a world where new inputs to our rational capacities constantly impinge on us. Instead, the issues is that manipulative inputs occur in such a way that we are unlikely to be aware of an *opportunity* to reflect. A counterfactual opportunity to pause and reflect on new elements of our "mental economy" is a necessary part of taking responsibility and subjecting our mechanism to "critical scrutiny" (199).

This shifts the question of responsibility to our counterfactual ability to be *aware of an opportunity* to reflect accurately. This seems to me to be a rather unsatisfying response, but one that is unsatisfying for reasons a metaphysically focused responsibility theorist might not appreciate. Fischer has already noted that, in day to day life, we rarely have the opportunity to reflect on our mental inputs before acting – and shifts to a counterfactual ability to be aware of such opportunities. But from an epistemic point of view, this ability is unlikely to ever be *fully* available to us. Indeed, a capacity to be aware of our mental inputs is precisely the kind of thing that much of the recent psychological literature on automaticity and bias can be seen as denying. In cases of situationally cued automatic response, I will not be afforded the relevant opportunity for critical scrutiny - and there is no nearby counterfactual world where I *could* be. This is precisely what it means to act automatically via a situational cue. Furthermore, in cases of non-automatic but affectively valanced or mediated input, even if I were to reflect, my critical evaluation of the input would not "see through" the mediation to some plainer truth about the mental state. It is far more likely in such cases that I will not even be aware of an opportunity for reflection. This kind of commonplace epistemic self-ignorance threatens to blur the differences between manipulation cases and "normal" cognitive life such that there is no clear and bright line between them.

In some sense, Fisher's view might be trying to make compatible two Frankfurtian insights that seem to be in tension: 1) that the sources of our desires and attitudes do not matter so long as we *take* responsibility for them by integrating them into our web of beliefs and attitudes and wholeheartedly accepting that web, and 2) that the sources of our desires and attitudes *do* matter, insofar as implanted desires are not ones we can be wholehearted about. The Fischer view attempts to reduce this tension by showing us that 2 is true precisely because our way of thinking about the etiology of desires, beliefs and attitudes is that *wherever* they come from, we usually have some opportunity to scrutinize them - to subject them to rational reflection. I'm arguing that when we attend to the epistemic abilities of agents on the ground, it appears that Fischer's attempted diffusion is not entirely successful. Assuming *some* kinds of automaticity and affective mediation are commonplace, we are unlikely to ever know the real sources of some elements of our mental economies, eliding the difference between manipulation and non-manipulation cases.

Here I'll pause to acknowledge a very recent contribution to these debates, Al Mele (2019)'s *Manipulated Agents*, in which he argues for an explicitly historical conception of moral responsibility. Mele gives us a spectrum of views that make clear the extent to which a theorist needs to be committed to the idea that the way in which an agent's internal "condition" came into being matters for moral responsibility. For Mele, "An agent's internal condition at a time may be understood as something specified by the collection of all psychological truths about the agent at the time that are silent on how he came to be as he is at that time" (5). What are the options? Without going into detail about them all, they range, roughly, from an unconditional internalism, where, "An agent with condition C (however he arrived in that condition) when he A-s, is morally responsible for A no matter how he came to be in C," to an unconditional externalism, where any internalist view that an agent's internal condition matters for responsibility is false (13-14). The more interesting views are conditionalized versions of each pole. So, conditional internalism says that: "An agent with condition C (arrived at in a suitable way) when he A-s, is morally responsible for A" (13). And Conditional Externalism that: "even if some conditional internalist thesis is true, agents sometimes are morally responsible for A partly because of how they came to be in the internal condition that issues in their A-ing; and, more specifically, in these cases, there is another possible way of having come to be in that internal condition such that if they had come to be in that condition in that way, then, holding everything else fixed..., including the fact that they A-ed, they would have been morally responsible for A" (9).

Why does this matter? Well, if Mele is on to something, then it looks like any plausible view of moral responsibility is going to be committed to some kind of historicism about responsibility (as Fischer acknowledges). And, the view must be historical in a way that is sensitive to the difference between manipulated and non-manipulated agents. Or, in other words, sensitive to the question of how our internal conditions ended up the way they did. I've indicated above that this kind of sensitivity can't amount to checking whether or not agents had the ability to critically scrutinize their mental economies (or internal conditions). Whatever it amounts to, this historicity means that, in determining responsibility ascriptions, our epistemic load is a good deal heavier than one might immediately think. In order to get the story right, we need to react not only to time-slice considerations of agents and their actions, but to much more complex causal factors that occurred in the recent (and perhaps) distant past. The extent to which agents' control over their present situations, the opportunities they had to avoid being in those situations, and the control they have over their internal conditions are thus all live and important epistemic concerns. This issue of epistemic load is one to which I'll return in Chapter

Two.

1.2.3 Fair opportunities and situational control

Fischer's discussion of fair opportunities to scrutinize one's mental economy leads naturally to the work of Brink and Nelkin (2013), who advance a version of reason-responsive compatibilism that focuses both on the capacities of individual agents and on the situations they find themselves in. They hold that agents are responsible insofar as they have normative competence (including capacities for cognitive and volitional control), and situational control – which they gloss as agents having a "fair opportunity" to avoid wrongdoing. As they say: "we think that moral philosophers tend to focus on the capacities involved in responsibility and so tend to ignore the situational element in responsibility recognized in the criminal law literature" (285). Brink and Nelkin's view, then, promises to expand the scope of our responsibility theorizing so that we might better understand the way situational factors partially constitute the responsibility of agents, especially as it relates to issues of excuse, exemption, and justification.

At first glance, this looks like another promising avenue for exploring how we come to judge whether agents are or are not responsible. The business of excuse and justification, for instance, clearly turns on questions of judgment, evidence, and interpretation. However, Brink and Nelkin make several things clear from the start. First, they are resolutely response-independent theorists. They read Strawson as a realist about responsibility, and thus, when considering the practice of blame, for example, do not hold that our attitudes about blameworthiness determine the fittingness of blame. Instead, there are independent facts about the conditions of the appropriateness of blame that determine when someone is blameworthy, and our attitudes ought to (ideally) track these facts.

Now, a strongly realist metaphysics in no way precludes the need for an epistemology. However, it does lend itself to a natural kind of epistemic deflation. It's easy to specify its form and be very ecumenical about its content: the epistemology of responsibility is just whatever reliably gets us towards the realist facts that the right metaphysics picks out. But before we accept this deflationary epistemology, we ought to ask: can the realist metaphysician really get an epistemology so cheaply? Do we, after all, *have* a reliable way of getting at the realist facts? And even if we *could* read relevant metaphysical facts about agents off of the world, would that be the end of the story? Here is the closest we get to a claim about the epistemic processes involved in determining normative competence in Brink and Nelkin (2013). Considering the issue of our epistemic ability to distinguish between "can't and won't" with regard to irresistible urges, Brink and Nelkin write:

> Consider the worry that we cannot reliably distinguish between an inability to overcome and a failure to overcome such obstacles. First of all, this is an evidentiary problem, not a claim about the ingredients of normative competence. Moreover, this evidentiary problem seems no worse than the one for the cognitive dimension of normative competence, which requires us to distinguish between a genuine inability to recognize something as wrong and a failure to form correct normative beliefs or attend to normative information at hand. Making the distinction between can't and won't is a challenge, but not an insurmountable one, in either the cognitive or volitional case. For instance, there are neurophysiological tests for various forms of affective, as well as cognitive, sensitivity, such as electrodermal tests of empathetic responsiveness. (300)

Notice that Brink and Nelkin take two approaches to the complaint. First, they point out that our lack of epistemic access is no complaint about whether theirs is the right metaphysical story. All theories will encounter this hurdle. Second, they assert that we have reliable ways to determine cognitive and volitional capacities, such as "electrodermal tests." I find this proposal overly optimistic, and, more importantly, non-responsive to the kinds of challenge I'm pushing. That is, at best, these methods seem woefully inadequate to our day-to-day epistemology (assuming they are even all that reliable).[10] We do not have the time, resources, or justification to hook up the agents we meet in our lives to electrodermal machines and check them for empathetic responsiveness – we have to use our senses, intuitions, and practical know-how.

Even if we could get a reliable capacity score on agents in real time, what would this tell us about our epistemology? We can, after all, imagine a nearby future where such a thing is possible. Imagine we have a device, the "responsibil-o-meter" which allows for near instantaneous readings of responsibility when we point it at an agent. This would not be the end of the story when it came to forming a judgement about the agent, their actions, and our responses to them. That is, even if my responsibil-o-meter told me that you are 78 percent responsible in a given case, that wouldn't entail any specific way I ought to react to you, or tell me how to hold you accountable. One way to think about this would be that the responsibil-o-meter might fully determine attributability but not accountability. What the electrodermal tests (or their responsibil-o-meter counterparts) tell us may just be the output of a complicated function which combines the amount of control an agent has over a situation with their receptivity and responsiveness to reasons at the current moment. This would be a fascinating number to get, but it wouldn't answer many of the responsibility-relevant questions that interest us.

For example, imagine what initially seems like a clear cut case: I invite you over for dinner and, when you think I'm not looking I see you take some cash I'd left on a bureau in my office. Preparing to summon up some choice reactive attitudes, I point the responsibil-o-meter at you, but it returns a reading of 0% responsibility. I frown and give the meter a whack - surely that can't be right. Again, it reads 0, and a line of text below the number reads: "manipulation case." What should I do?

[10] Lewis (2016) makes a similar argument: "we should note that there is no mechanism through which we can directly find out whether an offender acted out of ill will, and if so, how objectionable his or her quality of will really was. The behavioral sciences, for example, do not give us direct access to the reasons that people acted on in the past. We cannot use fMRI results to discern whether someone convicted of burglary committed the crime out of desperation or simply for the thrill" (168).

To insist that the reading of 0% be respected is, I think, unhelpful – or rather it doesn't tell us very much. First of all, it isn't yet action guiding. That the metaphysical reader tells me you are 0% responsible doesn't yet tell me how I ought to react to you – there is more to say about the nature and purpose of our practices that would start to answer this question. Second, to insist that I now know everything relevant about the case that I'd need to in order to move towards holding you responsible just seems to beg the question about how our epistemic practices operate and what their relationship to the right metaphysics of responsibility is. Even if we strongly suspect that something like a manipulation case is in play, this isn't the end of the epistemic story. To make sense of the case, for example, I might need a larger narrative about the agent. Or there may well still be other good reasons to hold you to account (forward looking, or, practice internal reasons, for instance).

What of situational control? As I've pointed out, this seems like an area ripe for epistemology. In Brink and Nelkin's eyes the crucial questions are ones about concrete states of affairs – whether agents are experiencing coercion or duress, for instance. However, the really tricky questions about situational control involve expanding the scope of consideration to much larger contextual and narrative factors. In other words, coercion and duress seem like the *easy* cases, epistemically speaking. We can all agree that agents who are coerced are not responsible, even if it is challenging to give the right account of coercion. But what about agents who are primed by a situational cue that reminds them of past racist abuse they've suffered at the hands of a relevantly similar group? What about someone who engages in *slightly* risky behavior that doesn't quite rise to the level of negligence? These more complicated questions are the ones involving situational control that it seems like our metaphysics simply won't be able to give precise answers to on its own.

In any case, Brink and Nelkin describe two models of how the metaphysical architecture might work. One is that normative competence and situational control are "individually necessary and jointly sufficient but independent factors in responsibility" (304). They go on to say that: "On this picture, we assess an agent in each area separately. We figure out whether she had the relevant capacities (e.g. were they "normal" or "sufficient"), and then we figure out whether she had the fair opportunity to exercise them" (304). Again, we see clearly that on this idealized picture we read these capacitarian and situational factors directly off of the agent and world, an epistemology that tightly corresponds to a translucent metaphysics.

The other model is that normative competence and situational control are individually necessary and jointly sufficient, "but at least sometimes interacting... On this picture, how much and what sort of capacities one needs can vary according to situational features. ... Such a conception would also imply that the requisite levels of normative competence and situational control are not invariant, but rather context-dependent" (304-305). On this model, things are radically more epistemically complicated. The epistemic load necessary for agents to determine responsibility facts about one another shoots up markedly if it's the case that capacities vary with situations and contextual factors. I now need to keep track of: a) what the relevant situational and contextual features that might affect your reasons-responsive capacities are, b) what I think your baseline or normal responsibility relevant capacities are, and c) what I think the current interaction of context and your capacities means for your ability to be reasons-responsive. If we want to be successful practitioners of responsibility, it looks as if we have to track capacities like situational control and normative competence *as well as* the ways in which they interact with social roles, asymmetries of power and situational contexts.

To lay my cards on the table: I think this is right! These are all things that we more or less implicitly track (or ought to track) when we think about whether you are responsible.[11] But, I don't, therefore, think there is likely to be an easy and

[11] Marina Oshana (2018) makes a similar point, summing up her exploration of the relationship between power and responsibility by saying: "successful practices of responsibility mandate effortful co- awareness of the roles we inhabit and of the configurations of power within which we operate" (86).

natural epistemic transparency of metaphysical facts. We are not, that is, very good at reading all of these conditions off of the world – we have heuristics to help us out, and baseline assumptions, and probably some cognitively sophisticated (if non-conscious) Bayesean updating schemes. But even so, if we *are* getting responsibility right (metaphysically speaking) given all of the above, it seems to be as much by luck, or non-conscious schematizing as it is by skill. This, in any case, is the argument I'm going to make in the next chapter, and I recognize that it's where the action lies, and will require much more support.

Once again, let me stress that this critical reconstruction isn't meant to impugn Brink and Nelkin's view, which I find metaphysically plausible. They are not responding to my concerns precisely because those concerns are *mine*, not theirs. What I'm aiming to show is, as I've said, the contours of a path not taken, an alternate way of looking at things. Whether or not we are realists about the metaphysical conditions of blame and responsibility, we can ask: "How do I come to believe that an agent lacks situational control? What properties in the world am I picking up on when I make this determination?" And, I'm suggesting that when we ask those questions, we are likely to be a little less cautiously optimistic about our ability to hold each other responsibly reliably well.

1.3 Interpretation of Agent Meaning and Conversation

Given what I've said above concerning historicism and situational control it seems clear that responsibility ascription is highly contextual. Perhaps one way we navigate such contextualism is by recognizing that, rather than trying to get at an underlying set of responsibility facts (or perhaps *in addition* to this aim), our practices often aim at moral "conversation." When we blame, for instance, we are trying to *say* something to our fellow responsibility practitioners. I move now to discussion of this possibility - a very live option in the contemporary literature. I'd like to see whether conversational theorists have taken up the question of how it is we come to form responsibility judgments given things like credibility, agent meaning, peer disagreement, the difficulty of interpretation and so on. In the course of my investigation I'll mainly discuss three theorists: T.M. Scanlon, Michael McKenna, and Marina Oshana. As we'll see, each author gives us a glimpse at a path not taken in the contemporary discourse.

Scanlon (2008)'s theory of responsibility stands somewhat apart from those canvassed thus far, so I will discuss it first. He focuses on *permissibility* as the keystone notion of responsibility – and holds that our practices of responsibility primarily concern blame and praise-worthiness based on how permissible an action is. Scanlon points out that our *sense* of how permissible some action is (given our sense of the agent undertaking it) can come apart from its actual permissibility. That is, in assessing responsibility we are often concerned with an agent's quality of will, or their *reasons* or *intentions* for acting. And, the sense we have of an agent's reasons and intentions can come apart from those actual properties in interesting ways - we can be mistaken about what their reasons and intentions really are. Despite all of this interesting contetxtualism about agential meaning, however, Scanlon focuses on the metaphysical notion of permissibility itself. He is quick to point out that it is the general question of what an agent *could have reasonably known* about an action's permissibility that determines an agent's responsibility. This is so because what the agent knows about the permissibility of what they are doing has a profound effect on how we interpret what that action *means.*

Thus, on the one hand, we have what appears to be a normative and epistemic notion grounding our responsibility concepts: it's *our assessment* of what agents ought to reasonably know that determines the blame or praiseworthiness of their actions, rather than (directly) their metaphysical capacities. On the other hand, the idea of "reasonable knowing" is given a highly general and context insensitive gloss which ends up looking a lot like the question of whether or not an agent satisfies typical control and knowledge conditions. In the end, whether the agent is generally aware of and responsive to the right kinds of reasons will matter more than their actual worked out views and intentions.

Yet, the focus on permissibility as tied to the *meaning* of an action is a quite new (and welcome) addition to the concepts discussed thus far. For instance, Scanlon (2008) writes that:

> The meaning of an action in the sense I am concerned with should be distinguished from a different sense in which the meaning of an action is determined by the reactions of others (or by the ways it would be reasonable for them to react). For example, if the only family in the neighborhood that was not invited to the block party is also the only black family in the neighborhood, then that family may reasonably take this as a sign of prejudice, even if they were left out by mistake, or because only people with children were invited and they have none. Effects like their reaction are important and can affect the permissibility of an action. Meaning in this broader sense is not a function of the agent's actual reasons. It is, rather, a matter of what others reasonably or unreasonably take those reasons to be. But meaning in this sense is not the focus of my inquiry, because my concern is with the ways in which an agent's reasons for acting can affect the permissibility of what he or she does. (53- 54)

Here again, is a clear summary of a path not taken. Scanlon rightly notes that the reactions that others are likely to have will have some (and perhaps a significant) effect on our judgements of permissibility – both in the case of thinking through our own reasons for action, and in making sense of the judgements we come to concerning other people. Yet, Scanlon doesn't want to focus on this wider social and epistemic context. Rather, he is interested in the reasons *themselves* and their relationship to permissibility. As he says: "The case for the moral permissibility of doing X under conditions C depends on the reasons that someone in those conditions would have for doing X" (62-63). And, presumably, in an idealized kind of model, the reasons that they would have are disconnected from social pressures involving norms of meaning.

But why should we think that this is the fundamentally interesting question here? To return to Scanlon's own example, we might think that whether a family is the only black one in the neighborhood will be a strikingly important detail in determining agential meaning – and one that a more generalized conception of "reasons that someone in those conditions would have for doing X" will fail to capture. At the very least, we might think that focusing on the social and epistemic factors at play would prompt us to include the fact of being the only black family in an area as a relevant part of the "conditions" at issue.

Despite his clarity of focus, Scanlon's brief remarks indicate that, epistemically speaking, figuring out the reasons agents act on is a complicated matter of interpretation that involves positionality, social norms that guide interpretation, and the background assumptions we come to an interaction with. The meaning of an action implicates our quality of will in a way that is central for ascribing responsibility on many theories. And figuring out an agent's quality of will is fundamentally an interpretive enterprise. We cannot, that is, rely *only* on their stated intentions. As Scanlon writes:

> If someone acts with no regard whatsoever for the interests of another person, then this has a certain meaning — it indicates something significant about his attitude toward that person and about their relationship with each other — whether or not it was his intention to convey this. I have said that the meaning of an action is its significance for certain individuals, because the same action performed for the same reasons can have different significance for different people, depending on their relation to the agent. (54)

This short passage contains two deeply important insights for my purposes. The first, as I've said, is that the stated (even if truly and honestly stated) intentions of an agent do not *solely* determine the meaning of their action. The second is that meaning and significance is positional – it will be a function, partially, of who is conveying the meaning and who is receiving it. There is no Archimedian "meaning" of an action, as it were. Yet, I think perhaps because he foresees these potentially anti-realist pressures butting up against the edges of his view, Scanlon is quick to clarify that his idealized metaphysics of reasons will, as much as possible, try to sidestep the sociality and positionality of meaning and significance. He writes:

> The phrase "the meaning of the action for Jones" could be taken to mean the significance that Jones sees the action as having. This is not, however, what I intend. By the meaning of an action for a person, I mean 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. The significance of your action, for me, is thus something I can be mistaken about. I may regard your action as a betrayal, but I may be mistaken about this either because I misinterpret your reasons for acting as you did or because I have a mistaken idea of what I am entitled to demand of you, given our relationship. (54, my emphasis)

There are two ways to read Scanlon's talk of reasons here - neither of which set us down the path towards a worked out epistemology of responsibility. One way is to say (as Scanlon does - see page 55) that what we need in order to know the meaning of an action is just to know whichever reason was actually operative on an agent and caused them to act. Given that there are a range of reasons that could move an agent to perform a certain action (and given that the agent might see several of them), we need to know which reason (or reasons) was (or were) operative. The second way to read him is to see the relevant epistemology as more general. In this sense we don't care about which reasons a particular agent really did have access to, or which reasons were truly operative (when determining permissibility). What we care about are the *general* set of reasons that such agents *ought* to reasonably be expected to have access to and operate on.

Again - both readings are metaphysically interesting, and even involve their own deflationary epistemic theory. But the path I'm interested in, given Scanlon's contextual insights concerning permissibility, is the one that interrogates the kind of sociality, positionality, and interepretive interplay the quoted passages uncover.

1.3.1 McKenna and agent meaning

Given the even more explicitly "conversational" nature of their views of responsibility, Michael McKenna and Marina Oshana come closer to giving us the kind of full-on accounts of the epistemology of responsibility I'm searching for. Like Scanlon, McKenna focuses on the ways in which the "meaning" of our actions figures into how we hold one another responsible. As we'll see, his particular account of *agent meaning* is firmly situated in the context of his conversational theory of moral responsibility. I don't take this to be a defect of his account – instead I want to show that the points he makes about the importance of interpretation need not be chained to a controversial theory of responsibility, and that there is a great deal of room for expansion. I'll deal first with his (2012) arguments about the conversational nature of responsibility, and then examine his recent turn towards a *socially* mediated account of responsibility, which brings the focus of his arguments much closer to my own.

McKenna (2012), in his *Conversation and Responsibility*, does an admirable job of surveying the state of play in the contemporary responsibility literature, helpfully cutting through many debates to, it seems to me, get at the heart of the matter. I agree with McKenna, for instance, that much of the literature focuses too exclusively on the issue of free will. Even *if* freedom of the will fully captured the control condition for moral responsibility (which I doubt), this would not tell us everything interesting about moral responsibility – it is just one condition on it. Not only this, McKenna is both even-handed and penetrating in discussing the entanglement of the epistemic and metaphysical realms that result from Strawson's admonition that we ought to look at practices of holding responsible as a way to make in-roads into the concept of responsibility itself.

Yet, McKenna (2012) is also perfectly clear that it is the metaphysical realm which has explanatory primacy when it comes to responsibility theorizing – and that the metaphysical is therefore where his primary interests lie. Responsibility, conversational or not, is a matter of the properties and capacities of agents. As he writes:

> Where I part company with Strawson and those following his lead is in the contention that considerations about the nature of holding morally responsible are more basic or fundamental than considerations about the nature of being morally responsible. I think this is wrong. Room in a broadly Strawsonian theory of moral responsibility needs to be made for the thesis that a person's status as a morally responsible agent has a metaphysical standing to which considerations about holding morally responsible must answer. (3)

In some ways, then, McKenna is *less* interested in the epistemic dimension of responsibility than other Strawsonians. After all, he makes plain that any account of holding responsible (which will include, on a Strawsonian view, the having of reactive attitudes themselves), must answer to the metaphysical capacities that vouchsafe *being* responsible.

However, this endorsement of the primacy of being responsible is mainly a concern about grounding, and, one I'm happy to take on board. As I wrote in the introduction, my account is perfectly compatible with the idea that it is the metaphysical capacities and mental properties of agents which primarily *ground* the truth, accuracy, reliability or justification of our responsibility judgements. My claim, once again, is that given a certain *gap* between the required knowledge of the properties on which our practices are grounded and our *actual* knowledge of those properties, a further account of the epistemology of moral responsibility is necessary.

In any case, it will be helpful to outline McKenna's basic view of the nature of responsibility. McKenna holds that responsibility requires three conditions. The first two are familiar: *epistemic* and *control* conditions, of the kind we saw in our survey of Brink and Nelkin and Fischer and Ravizza. As he puts it: “Only a person with the relevant capacities can be a candidate for assessment in terms of praise and blame” (12). However, McKenna recognizes a certain epistemic uncertainty in our attributions of responsibility, given that he views it as a scalar notion, and one which is contextually sensitive (10). To determine whether an agent is responsible, we may need to know about the particular context she is in, and even then, the answer is like to be “patchy.” Given that McKenna agrees that these capacities are context sensitive and scalar, we might naturally ask: how are we tracking them?

Here McKenna relies on a third condition for moral responsibility: “quality of will.” What really matters, in terms of determining responsibility, is what the utilization of an agent’s epistemic and control capacities tells us about her intentions, and thus her quality of will. This is all well and good, but once again, it cries out for an informational theory: how do we get this evidence? And how do we know when we’ve done a better or worse job interpreting it? This is particularly troubling insofar as McKenna is a staunch Davidsonian naturalist. That is, on his account, intentions are an irreducible part of the natural world, and *cause* actions. So access to and assessment of intentions is going to be informationally crucial for determining an agent’s quality of will.

The most McKenna tells us is that quality of will is, “revealed in her reasons for action, her immediate proximal intention in acting as she did, and the attitudes she has for relevant others, as well as the broader motives or plans into which specific reasons and intentions for action are nested [and is] ...also revealed in terms of epistemic considerations regarding her relative state of belief at the time” (18-19). Again, the talk of “revelation” fits the pattern I’ve outlined in the chapter so far. Instead of giving us an account of the epistemic processes by which agents come to find out about things like reasons and intentions, a transparent, deflationary, quietest and reliable process is assumed: we just do it.

Once we have evidence of quality of will, McKenna is interestingly alive to the idea that *holding* an agent morally responsible is not *merely* a matter of theoretical concerns. As he puts it, holding responsible: "*also* involves adopting a *practical* attitude toward the person by which those holding morally responsible are disposed to place certain moral demands upon her, have certain expectations of her, and praise and blame her depending upon how she conducts herself" (22). But this doesn't explain how one arrives at such a practical attitude, or what epistemic processes and considerations result in the taking up of the attitude itself. McKenna is noting that holding responsible is nor merely a matter of judgement, but also, as Strawson reminds us, a certain attitudinal disposition.

Yet, McKenna also wants us to pivot away from Strawson's more epistemic concerns and back towards a capacitarian account. What's his argument for this "reconceptualization?" The best way to summarize it is to present McKenna's critique of R. Jay Wallace's view of moral responsibility. Wallace develops an account on which: "theorizing about moral responsibility. . . [conceptualizes] it 'as essentially a normative debate about the conditions that render it appropriate to hold a person morally responsible... [facilitating] an interpretation of the facts about responsibility that make them dependent – in the right way – on our practices of holding morally responsible" [(Wallace (1994, 85) quoted in (41-42)].[12] To that end, Wallace gives us a Normative Principle (N):

> *S* is morally responsible (for action *X*) if and only if it would be appropriate to hold *S* morally responsible (for action *X*). [Wallace (1994: 91), quoted on (34)]

Now, importantly, as McKenna tells us: "[Wallace's] claim is about a metaphysical thesis, not an epistemic one. The claim is not that we are to seek an interpretation of the facts about moral responsibility that makes our ability to know them depend on our practices of holding responsible. Rather, the facts about responsibility themselves are dependent on our practices; the latter, as Wallace puts it, "fix" the former.

[12] See also Vargas (2004) on this reading of Wallace.

Wallace calls this the "normative interpretation" of the free will and moral responsibility controversy.

McKenna thinks we can avoid the temptation to read Wallace's claim in an epistemic light by making sure not to confuse (N) with a claim about judgements. Holding S responsible does not boil down to judging S to be responsible (that is, it's not merely a matter of belief) – it's a matter of having the right reactive attitudes, amongst other things. In the end, McKenna must reject (N), metaphysical or not, because it grounds our metaphysical natures as responsible beings in our practices of holding one another responsible – putting the cart before the horse (for McKenna).

I don't wish to enter into the debate over metaphysical or epistemic priority.[13] I have already made clear that my view can be consistent with either response dependence or staunchly metaphysically realist views about the nature of responsibility. But, the route not taken by Wallace (at least as re-constructed by McKenna) is much closer to my ambitions. McKenna tells us that Wallace is *not* claiming that we ought to interpret facts about moral responsibility in such a way that it would make our *knowledge* of those facts dependent on our practices. But why shouldn't this be precisely the case, given our understanding of knowledge production? Whatever the metaphysical facts *are*, our practice embedded interpretations of those facts do constitute the extent of our *knowledge* of the facts. The epistemic claim that McKenna thinks Wallace *doesn't* make, is precisely the lacuna I'm trying to explore.

When we learn to let go of some interpretive fear, and accept that interpretation can produce perfectly good and reliable knowledge of the sort we use in everyday life – knowledge that may fall short of reflecting ultimate reality, and may be epistemically limited, confused, mal-adpated or twisted in familiar ways – we can start attending to what our practices and practitioners within them really get up to. And, given further metaphysical commitments about how things really are or ought to be, we

[13] My sympathies lie most with something like McKenna's "no priority" view in attempting to strike a balance between claims about the direction of priority.

can try to correct and improve those practices.

1.3.2 Conversation and the social world

Now, McKenna goes further than most in articulating an interpretive component within his theory of moral responsibility. He writes:

> On the conversational theory, an agent's actions— those that are candidates for blameworthiness or praiseworthiness— are potential bearers of meaning, where meaning is a function of the quality of an agent's will. This meaning is analogous to the meaning a competent speaker conveys when she engages in conversation. I call this agent meaning. Like speaker meaning, agent meaning can be affected by the interpretive framework whereby others interpret the agent. In the case of agent meaning, a moral community assigns saliences to types of actions, and they do so in light of expectations about the cooperative constraints of something analogous to a conversational transaction. (45)

McKenna is quite explicit that, in order for our practices of responsibility to be effective, it has to be the case that there is an interpretive framework against which our actions and reactions are judged (and by which our actions and reactions are modulated). As he says:

> A competent moral agent thus acts in a social context wherein these interpretive pressures are liable to affect her judgments about what does and does not signal good or ill will and so what might or might not be communicated to others who stand prepared to hold her to account by way of praising and blaming practices... In short, just as the actions for which we are accountable have meaning, so too the means of blaming or praising through outward manifestations of reactive emotions have meaning. And in each case, this sort of meaning is shaped in part by a background set of loose interpretive conventions against which instances are evaluated. (45-46)

It is striking then, that the interpretive conventions themselves are not a central focus of McKenna's theory of agent meaning and moral responsibility. It seems clear, as I've been arguing, that the epistemic process of determining whether an agent is responsible is going to be as much a matter of figuring out what interpretive moves are licensed by our responsibility practices as it is a matter of determining relevant metaphysical facts about the agents in question. In a recent paper, McKenna (2018) comes much closer to considering the epistemic issues that are the subject of my project. Recognizing that, for many quality of will and reasons-responsive theorists, "moral responsibility is essentially interpersonal because being responsible is conceptually connected to holding responsible, which in turn is understood in terms of social practices," McKenna begins to articulate some ways in which social power and position might affect one's quality of will and culpability for wrongdoing (38).

As McKenna puts it: the "moral landscape will be affected insofar as even morally virtuous agents will be liable to participate unwittingly in wrongdoing" (2017, 44). Why should this be the case? McKenna begins by noting an important epistemic dimension in our practices of determining the quality of will of other agents. This is that the, "standards for a competent agent's acting from a reasonable quality of will are understood by reference to the expectations of the moral community positioned to hold responsible" (39). This, in turn, leads to a question that has been neglected in the Strawsonian literature, namely: How do power relations and social inequities affect responsibility?

Already, we can see a familiar road not taken. McKenna is going to be interested, in the end, in the metaphysical conditions of responsibility. Recall that, on McKenna's view, there are three conditions for responsibility: a control condition, an epistemic condition, and a quality of will condition. McKenna wants to point out that social dynamics can (and do) affect agential control, an agent's ability to know and be in contact with the right kind of reasons, and, therefore, agents' quality of will. This kind of social turn is of the utmost importance for contemporary debates about responsibility. It is shocking how little attention has been paid to positionality, privilege, and power when thinking about the socially mediated ways in which holding and being responsibility take place.

Yet, once again, the metaphysical expressions of these issues are not the only things we ought to be alive to. We could also ask how the interpretation of an agent's actions, quality of will, and responsibility are affected by our epistemic practices, not merely how power, position and privilege directly affect these metaphysical capacities. Again, McKenna seems aware that this is an issue worth taking up. His own theory, of course, is conversational, and so he recognizes, for instance, that the *meaning* of an agent's quality of will, "can be affected by interpretive and social frameworks," so that salient social meaning are often settled by, "social authority - being positioned to have one's interpretive scheme do the work in settling meaning" (2019, 47).

McKenna rightly notes that this is a moral issue of some importance. Some particular (or some particular classes of) agents in any given society will have the resources and power to set expectations and police departures from that expected behavior. But again, McKenna is interested in this question insofar as it shows something about the culpability of individual participants in socially stratified responsibility practices, rather than in an explicit account of how our reasoning about responsibility may go astray or could be improved.

As he puts it:

> But as it happens, these interpretive schemes, and the conventional meanings they assign to patterns of action, have baked-in forms of bias that serve as the basis for even well- intentioned people to engage with each other. Moreover, since on the conversational theory the interpretive enterprise involves efforts to understand the particularized meaning of an agent's actions, moral agents rely upon the interpretive community being inclined to interpret them well. Those disadvantaged due to significant asymmetries in relations of power then risk alienation by defying or departing from these conventions, both in attempting to act with good will and in reacting to others when holding them accountable. Hence, our moral responsibility practices — our actual practices as they normally

> function — are in a sense morally tainted, or at least they are morally dubious. At the very least, they need to be assessed from a critical distance. (51)

This is, as should be obvious, very close to the kind of critique I advanced against Scanlon's account of agent meaning above. Without a properly descriptive account of the ways in which various biases are "baked in" to our interpretive schemes, we'll fail to really understand the practice we are studying as it operates in everyday life. In this sense, I see McKenna's recent work as a very large advance in the contemporary understanding of responsibility and reactive attitudes as conversational and involving agent meaning. But again, McKenna stops short of my project. In framing the non-ideal nature of our interpretive schemes as a *moral* issue rather than as a straightforwardly epistemic one, we don't get clear about how and where these epistemic deficiencies arise. I agree that there *are* moral issues embedded in our responsibility practices that call for serious reflection and theorizing. But a full response to these issues would say: if we were paying more attention to the robustly interpretive epistemic dimensions of our responsibility practices, we might have better epistemic norms, better ways of reacting to one another and determining the truth, reliability, and goodness of our judgements. To hammer this point home, I turn lastly to the work of Marina Oshana.

1.3.3 Asymmetries in Social Power

Marina Oshana (2018), like McKenna, is concerned that fundamental injustices lie at the heart of our moral responsibility practices which have mostly gone unexamined and uninterrogated. And, like McKenna, her theory is conversational, or, in her words, "interlocutive." She is worried by asymmetries in credibility and recognitive respect afforded to our epistemic peers. The asymmetries often occur due to power imbalances – imbalances which shape how agents see themselves and others as responsible. Such recognitional impairments, she argues, may affect the very capacities that constitute our agency, or, at the very least, impact our responsibility practices in important ways.

Again, these concerns are very nearby to the central features of my project, near enough that there is critical overlap in several areas. Oshana writes that:

> [Asymmetries of power] have gone unnoticed, I think, because those of us who write about responsibility have focused almost exclusively on the status of the presumed responsible agent... For instance, there is little said on the question of who decides upon the standards of evidence for credibility: Who in the interlocutive community determines the worthiness of others to stake claims for accountability and to lay charges of responsibility? ... little if any attention has been paid to the social, economic, and political power dynamics – often asymmetric - that structure the context within which evaluations of responsibility occur. (82)

I agree! These questions are worth taking up, and focusing on the metaphysical make-up of particular agents rather than a larger social and epistemic context has meant that they rarely are. However, Oshana makes clear that her particular way of answering these questions is to focus on a refinement of the epistemic condition on moral responsibility and some of its implications for our practices.

Why think these questions could be answered by cleaning up the epistemic condition on moral responsibility? As Oshana continues on to argue, “who we are and how we have been represented to the world shapes expectations about our agential competence and capabilities as responsible parties” (83). If we think that agents need to have certain kinds of knowledge and knowledge acquiring capacities, then these asymmetries in power, privilege, recognition and credibility affect both the kinds of things we can know (thereby directly affecting the knowledge condition on moral responsibility), and the ways we will be represented as knowers (thereby affecting *indirectly* or representationally the knowledge condition on moral responsibility). These factors are part of the epistemic “ecology” or environment that enables and supports the development and maintenance of our responsibility capacities. And bad ecologies can lead to injustice, as when those without the epistemic resources to play the responsibility game well are represented as lousy players, rather than as suffering from epistemic injustice.[14]

These kinds of concerns help clarify the normative contours of my project. I want to look at common epistemic disruptions and failures for their own sake - so that we can work towards repairing and overcoming them. However, I also want to look at them in service of a more theory driven aim: so that we can identify the more general and schematic structure of our practical epistemology of responsibility. For example, once we note a particularly impaired responsibility ecology, I want to say something about how those conditions will translate to our judgments about each other. And, if we can identify some general features of good epistemic practice, it should be easier to see when and why the kind of break-down's Oshana is concerned with occurred or can be patched up. Oshana's identification of the systemic and structural barriers the less powerful face in responsibility practice are helpful resources for my project: they identify important places of rupture that lay bare the more general epistemic structure I'm interested in exploring. When there are credibility deficits, I ask, what does that show us about how credibility *usually* operates? In what ways are they linked to other kinds of agential interpretation that leads to differentiated responsibility ascription? The work Oshana is engaged in is both important background for my view, then, at the same time as it provides compelling case studies.

Oshana notes that we can always ask the following question concerning knowledge *acquisition*: "What are the right, or best, ways of acquiring knowledge of a sort

[14] In the next two chapters, I'll identify common practical epistemic errors of this kind as "status sensitivity errors" which include, among other things: 1) "Genuflection to Power and Position:" wherein we are much more likely to find *excuses* (and, paradoxically, to view as more generally responsible) for those with power, privilege, and/or proximity to us in social position, and 2) "Under-consideration of the Oppressed:" the flip-side of the above: wherein it is much more likely that we will view as non-responsible (in ways that aid oppression) those who are oppressed, while at the same time holding responsible the oppressed when it maintains that status quo. Take, as an example, the prevalence of trying juvenile people of color as adults in the U.S. criminal justice system, paired with prevalent racist rhetorics claiming they are not as fully rational as their white counterparts.

requisite to being responsible and to holding responsible?" (85). This is a particularly important question (or really, two distinct questions), given the ways in which I argue that the kinds of power asymmetries Oshana is concerned with are endemic. Oshana might identify some particularly politically fraught versions – ones which we ought to take seriously – but everyday life in moral responsibility practice is generally asymmetrical. There are always asymmetries and epistemic defeaters around. In fact, I think it is telling that Oshana sees her question as a single one, failing (at least in this moment) to distinguish between the conditions that allow an agent to *be* responsible, and those that allow them to hold others responsible.

Keeping these questions distinct is important. What kinds of processes of knowledge acquisition allow us to *be* responsible is something I will touch on at points in this dissertation, although my claims must, necessarily, be incomplete and suggestive. As Oshana notes, "successfully representing ourselves as responsible agents dictates that we remain alert to the impressions others have of us, especially in contexts that are fertile ground for bias... successful practices of responsibility mandate effortful co-awareness of the roles we inhabit and of the configurations of power within which we operate" (86). It is in this vein that philosophers like Meghan Griffith (2019) argue that a capacity for *narrative* is a crucial part of becoming a full-fledged responsible moral agent, or that Hechler and Kessler (2018) argue that attributions of responsibility by observers may help correct individual errors and foster agency. This is the heart of Oshana's argument. In the paper she begins by describing, at length, various forms and concrete instances of sexism that female identifying professors and academics are subjected to in their lives as researchers and teachers (87-91). In particular, she focuses in on the ways in which women in the classroom are not accorded the same kinds of respect and recognition as their male professorial peers. From this, she concludes that these are instances where, "responsibility-competence is wrongly discredited owing to stereotype, bias, and marginalization" (91). However, this claim is somewhat opaque. It is true that the professors in question suffer from unfair portrayals, seeings as and representations of their competency. But what does this have to do with those agents' abilities to see their students, for instance, as responsible, and to hold them to account?

Oshana's answer is that:

> it is within dynamics of power that persons come to be considered reliable agents, credible interlocutors, and deserving members of the moral and broader community. It is within dynamics of social power that the normative expectations we have of ourselves and of each other originate. It is within dynamics of social power that persons judge one another as apt participants in the interpersonal relationships that characterize responsible agency. (92)

Once again, then, this seems to be a question of how these experiences form and shape our *capacities* for responsible agency, not directly about how we come to the judgements of responsibility (or how they break down due to the biases and stereotypes) themselves. So, we see a more purely epistemic possible path – although we can (and should) admit that the formation of these capacities in particular social ecologies is metaphysically interesting and important, we can also see how the ecologies and people's experiences in them affect how we think of them, what processes of judgement they come to use (and we come to use on them), and so on. In any case, it's clear that there is a basic claim that both Oshana and I accept: coming to be a responsible agent involves socially mediated (and developed) awareness of what other people think of us, and the contexts that affect these judgments. Given this claim, I want to push in a slightly different direction. If "effortful co-awareness" of the kinds of judgements we are likely to make about one another is an important part of responsibility practice, how do we go about forming those judgements, and how might that process go well or poorly?

A final, concrete, example will help meaningfully differentiate our projects and show why answering this second kind of question matters. Consider Oshana's treatment of the killing of Trayvon Martin by George Zimmerman. This widely publicized case of the killing of an unarmed, black teenager by an armed member of a "neighborhood watch" group was a test case for various intuitions about race, white-supremacy, and state-sanctioned violence in contemporary America. But, again, how does the case relate to responsibility? Oshana is interested both: 1) in the way that the legal system came to deal with Zimmerman, and what this might reveal about asymmetries in power and our ability to hold one another responsible across those positional differences, and 2) "why Martin was not afforded the opportunity to offer an account— even if Zimmerman was not entitled to demand one— of his presence in the neighborhood" (101). In other words, why Martin was not treated as a worthy interlocutor.

Considering the first issue, Oshana writes, "To take seriously civil rights and the rule of law is to take seriously the necessity of responsibility practices that are attentive to the asymmetrical dynamics that inform the activity of ascribing responsibility. It is to be sensitive to asymmetries that are present in being called upon to give an account of one's behavior" (96-97). This last sentence contains a rather slippery argumentative move, and differentiates our goals. Although reasonable people can (and do) disagree, the Zimmerman/Martin case seems to me like an indescribably depressing case of bias – both in an individual and in the law – a way of treating black people and their bodies as less important in the material and legal realms. But, again, what precisely, does it have do with determining ascriptions of responsibility? And what does that have to do with "being called upon to give an account of one's behavior?" Perhaps one issue here is that Oshana seems to be implicitly presuming that responsibility just *is* accountability. However, this seems to preclude the very kinds of things she's interested in: that is, how power asymmetries affect the way we *ascribe* or judge one another to be responsible. Given that we can ascribe responsibility (and perhaps even hold someone responsible) without calling on them to give an account of their behavior, equating the two is going to hamstring our analyses of injustice in these cases.

Oshana's focus is on the important ways that Zimmerman is able to *avoid* being held responsible – the ways in which the laws and social conventions are set up so as to be maximally forgiving to people (usually white men) in Zimmerman's position. But it isn't clear how these systemic or structural points directly relate to interpersonal issues of credibility and recognition respect. As in McKenna's account, what we are primarily given here is an explanation of the ways in which a lack of epistemic credibility and power can lead to structural disadvantage (and of course, downstream interpersonal injustice). Oshana tells us that, "Acknowledging one another as equal participants in the interpersonal relationships definitive of a participatory democracy is essential to sustaining faith in the ideals of democracy. Being taken seriously as a competent interlocutive partner is a prerequisite to satisfying public interaction... this takes on critical urgency when confrontations involving the legally sanctioned and unprovoked use of lethal force are commonplace. Incidents of the sort I have described make a mockery of these ideals" (101-102). Yes - but what does interlocutive good faith look like in the Zimmerman case? We have no way of knowing what the interlocutive context of Martin and Zimmerman's encounter was, although it seems obvious that Martin was not treated fairly. This lack of dialogue does not mean we can't see and address the systemic racism that may have lead to the deadly encounter *and* which let Zimmerman off the hook in the trial, but it is hard to see what it has to do with moral responsibility directly. I'd argue that we should, in particular, resist the claim that it was "interlocutive failure" which lead to Martin's killing.[15] Would Zimmerman's acknowledgement of Martin as an equal participant in a shared democratic project have stopped the murder? Perhaps in some very tangential sense that, were Zimmerman to respect Martin in such a way it's unclear how

[15] A kind of claim recently endorsed by Malcom Gladwell in his latest pop-psychology book *Talking to Strangers*. He makes just such an argument about the Zimmerman case, that if only the two could have figured out how to converse, the killing wouldn't have happened. I'm highly suspicious of this kind of claim – see the following critical review: `https://www.theatlantic.com/ideas/archive/2019/09/when-malcolm-gladwell-says-nothing-at-all/597697/`.

the situation could have arisen at all. But this seems to mistake an interpersonal failure to afford recognition respect for the heart of the matter instead of dealing with the structural inequities that led to the possibility of that lack of respect, and thus, for an unarmed black teenager to be gunned down in the first place.

Oshana gives us compelling cases (that I don't dispute) of racial and gender bias – but we don't yet get a clear story of how those biases affect our judgements of responsibility and interpretations of one another as responsible agents. She may be right that there is a failure to treat one another in the kind of mutually recognitive ways that responsibility practice would ideally demand, but we lack the details to say much about the epistemology of these kinds of cases. In essence then, Oshana and McKenna are asking all of the right questions, and it's my job to scrounge up some answers. McKenna, in fact, helpfully notes that no one seems to have provided the answers up to this point. As he says: "how should social inequities and asymmetrical relations of power affect our responsibility practices and judgments? So far as I can tell, to this normative question, Strawsonians have little to offer, nor do I from the resources of my conversational theory" (54). I see my task, then, as filling in first the descriptive and then the normative lacunas: how *do* inevitable asymmetries and the messiness of our interpersonal lives affect our responsibility judgements? And what *should* we do about it? Our interpretive practices matter a great deal for moral responsibility and so, I think, it is time that someone offered an epistemic theory of those practices. To try to answer these questions is to rightly center, as McKenna and Oshana do, the normative question of how we can make our epistemic practices better. I turn towards the descriptive question in Chapter Two and the normative question in Chapters Three and Four.

1.4 Looking Ahead. Looking Back

In the next chapter, I'll look at contemporary work, primarily in social psychology, about what actually happens, in the real world, when we form judgements of responsibility. Part of what I'm doing in beginning with real-world cases is taking on-board and re-articulating a substantive methodological presumption. Namely, I begin, as many of the contemporary responsibility theorists canvassed above do, with a "Strawsonian," practice-oriented view of responsibility.[16] How do agents go about their interpretive business in day-to-day life? What are the mistakes they make? What might they get right? Embracing a practice-first Strawsonian spirit, I argue that we should look and see.

Why? Many "practice-first" theorists think we can extract some portion of our metaphysics of responsibility from the practices in which we are embedded. By beginning, for instance, with our "reactive attitudes," we can make headway towards determining what agential features responsibility tracks. However, such theorists have rarely taken the same practice-first route with the *epistemology* of responsibility. I propose that we can, and should, extend this methodology to the epistemic realm. By this I mean that we ought to look first at how agents actually do the epistemic work of determining responsibility (whether well or poorly). From this, we can extract some of the structure and content of what that epistemology looks like, and even begin to see what a more idealized epistemology might consist in. For my purposes, however, the more important point is that however far we end up getting in our structure-building, there are more immediate conceptual resources a practice-first examination of the epistemology of responsibility can give us. In particular, when we look at the extant practices we ought to be able to identify various non-ideal distorters or defeaters of our responsibility judgments. That is, even if we don't have a fully fleshed out ideal epistemology, it will still (in many cases) be relatively

[16] See P.F. Strawson (2003)'s seminal "Freedom and Resentment."

clear where agents commonly make mistakes in the non-ideal world.

Again, I think it's likely that many theoreticians of moral responsibility are perfectly aware that something like the lacuna I've spent this chapter tracing exists, although they may not have dealt with it directly. Tognazzini (2015), for instance, in a penetrating essay calling for a return to the original spirit of Strawsonianism, argues that we ought to pay more attention to the way our actual interpersonal relationships function when it comes to blame and responsibility. Paying attention to the moral emotions, their wellsprings, and our psychologies is, he claims, what Strawson wanted us to do – rather than construct metaphysical theories. It's hard to re-read Strawson and argue with that, but once we agree, we are in a strange and wild land. If it turns out that the heart of moral responsibility is about *judgments* and *interpersonal regard*, then we need evidence and theories that speak to *that.* Why is it that I'm prepared to see one friend in a certain light and not another? Why does my relationship with an intimate partner shift the types of excuses I'm willing to put up with in a way that being in the "relationship of humanity" with a stranger does not?

Tognazzini's focus, to repeat our theme, lies elsewhere. In his rich exploration of fictional cases, he is interested in interrogating just what this reciprocal stance amounts to, and how it connects up with theoretical notions like "answerability," "attributability," "accountability," and the "participant stance." The point I keep hammering home is that, although these are questions worth asking, they do not fill in the lacuna. How do we come to have the set of judgements we do about the accountability of an agent? What are the epistemic considerations that move me to take the "objective" rather than participant stance?

Let me summarize where we stand at the end of this chapter: the moral responsibility literature has very seriously engaged with the metaphysics of responsibility and various features of moral psychology which that metaphysics implicates. But, whatever the right view of metaphysics and moral psychology is, there's a different kind of question we can ask: “How do we accurately, fairly, or usefully track and judge that people are responsible?” I’m focusing on that question.

Given all I’ve said thus far, there are still two kinds of projects I might engage in. The first would be to give you a general theory about the epistemology of moral responsibility. In some sense, I’d like to do that, but we’d have to get very deep into metaphysical and epistemic weeds. That is, whatever the right epistemic story of moral responsibility is, it will have some connection with the right metaphysical story. And even *if* the two were largely modular, giving a general theory of the right epistemology will still involve many pitched battles about good epistemology itself. Those are battles that I’ll largely try to avoid.

I think I *can* credibly avoid them while still advancing a robust and interesting view, because there’s a second kind of project, the kind I’m undertaking in this dissertation: it says, “here are a set of practices, conditions, and common facts such that, when they obtain, there are distortions, disruptions and impairments to our judgements of whether and when people are morally responsible.” Notice that, for the second kind of project, one needn’t be committed to a particular view in either the metaphysics *or* the epistemology to see when and how these distortions will crop up, and why such defeaters matter.

Of course, this isn’t to say that I have no views about epistemology or metaphysics writ large. I can’t remain entirely neutral, and my view won’t be compatible with everyone’s favorite theory. The point is just that *my* project involves trying to get clear on where our epistemic practices of moral responsibility break down and what the causes of those break-downs might be. The goal is to think about what’s going to count as such a defeater under a wide range of credible views in epistemology and the metaphysics of responsibility. When it is useful to do so I will try to elucidate connections between particular kinds of theories and my own undertaking that I think can be picked up by those engaging in other parts of the literature.

In some sense, my arguments might be read as a kind of quietist position regarding the metaphysics of responsibility. The idea is just that, whatever account one favors, the metaphysical details will, at least sometimes, be misleading and unfruitful in the epistemic realm – we are unlikely to uncover, for instance, the precise neurological underpinnings of responsibility *and* to make such neurophysiological bases epistemically accessible to everyday practitioners of responsibility (that is, to all of us). Until the arrival of "responsibilo-meters" which we can surreptitiously point at our friends and family, the metaphysical facts simply can't play the kind of grounding role that would matter in a robust way for our epistemic work.

So, I propose that we view the metaphysical facts as a kind of black box. The box exists, and it matters. Into it go the various psychological and capacitarian facts from the the natural world, out of it comes responsibility. Such responsibility is real and it matters to us. But what happens inside the box needn't concern us, so much as whether we are giving plausible interpretations of the contours of the box itself.

Portions of Chapter 1 appear in publication as "What's the Relationship between the Theory and Practice of Moral Responsibility?", *Humana Mente* 15 (42): 29-62, 2022. This paper was co-authored by the dissertation author and his dissertation chair, Manuel Vargas.

Chapter 2

Practical Epistemic Problems for Moral Responsibility Practitioners

> When navigating social environments, we must somehow sort through... potential mentalistic causes, and arrive at the most probable interpretation. These abductive inferences require us to draw on our own background knowledge to fill in the gaps between behavioral observation and mental cause... And this is a point where pernicious social biases can enter into the mindreading process distorting our interpretations of the social world.
>
> Evan Westra[1]

[1] Westra, 2019, 2822

2.1 Introduction

The goal of this chapter is to lay out a model of the descriptive epistemology of moral responsibility. It seeks to model how we judge and attribute responsibility in day-to-day life, and to use that model to identify and explain the defeaters, distortions, and disruptions common to our responsibility practices. To do so, it ties together theoretical and empirical threads in psychology, theory of mind, and moral responsibility.

I've argued that the epistemology of moral responsibility is fundamentally interpretive and socially situated. I now begin to put some meat on the bones of those claims. Furthermore, I identify several ways in which our current practices lead to common errors which our metaphysical theories of responsibility, at best, do not adequately deal with, and at worst, obfuscate. This obfuscation, I go on to argue, can lead to downstream injustices in our practices; injustices closer attention to the epistemology of responsibility can help us combat.

Part of the motivation here is to show responsibility theorists why the mechanics of everyday responsibility ascriptions ought to matter to them. We know from work in moral psychology that there are many (and much studied) cognitive errors, biases, and heuristics that human beings make use of in their reasoning. And we know that, if anything, this is even more pronounced in social spaces and our reasoning about one another's mental states. In other words, it's a widely acknowledged background condition that various irrelevant factors can hijack our moral reactions and reasoning. Yet, responsibility theorists have tended to lay these kinds of biases and heuristics to the side when constructing their metaphysics of responsibility. This seems reasonable insofar as their question is *in virtue of what* someone is responsible – not whether well known distortions occurring due to the Fundamental Attribution Error, physical attractiveness, stereotypes, or seemingly benign Bayesian updating schemas affect our ability to track those features.

Yet, it would be equally mistaken to think that such factors play no role in our ecology of responsibility. If part of what shapes our capacities *is* the social space we inhabit, if our reactive attitudes play a central role in responsibility, or if we are on the hook for the norms and expectations that flow from our social roles, then these kinds of psychological foibles matter. It should be no surprise that our responsibility *practices* involve these kinds of misfirings as much as any other aspect of our cognitive life, even if the misfirings do not penetrate to the deep architecture of responsibility.

Given this, my focus will be on those places were errors seem to penetrate deepest, and occur most frequently. In other words, while it's true that *any* bias or heuristic might be a potential disrupter of responsibility attribution, I try to focus in on those that seem endemic and particularly problematic, and to explain why this is important for moral responsibility theorists.

To do so, I focus on two classes of error that show up particularly forcefully and often in responsibility attributional contexts. These are:

1. **Interpretive over-reach**: distortions, biases, and misfirings having to do with our tendency to interpret beyond what we have full license to claim to know about one another. As I've begun to argue, of course, interprative overreach may be a necessary part of most responsibility attribution. That is, we often lack the kinds of knowledge we'd need to make fully accurate responsibility assessments, and at best whatever knowledge we *do* have still requires us to make interpretive choices. Given this, it's crucial that we are *good* interpreters of responsibility, and so I identify particular ways in which we are likely to falter in everyday responsibility interpretation. These include:

   (a) **Over and under-estimation of causal control**: The amount that we feel an agent causally contributed to an outcome is directly correlated with the amount we attribute responsibility for that outcome to them. Unfortunately, many irrelevant factors seem to affect our judgments of causal contribution and control: estimations of warmth and competence,[2] facial features,[3] information, stereotypes, and biases about race, religion and gender,[4] the severity of outcomes,[5] our own self-image,[6] and so on. This set of issues is particularly important for reasons-responsive theories that postulate a control condition for responsibility.

(b) **Character Judgement**: We often make far-reaching judgments of character and quality of will based on extremely limited evidence (a single action for which we have only one, biased perspective, for example). This is particularly important for quality of will theories which make use of interpretations of actions to extrapolate claims about agential intention and character.[7]

(c) **Narrative framing errors**: Issues with interpreting causal control and agential character are "micro" level interpretive errors. Above and beyond this, there is a kind of narrative over-reach we often engage in that has gone under-noticed and under-theorized. At its most basic level, narrative over-reach is an issue of framing (or mode of presentation) effects. When we tell stories about agents, we activate and ramp up certain aspects of our moral psychology: affect, emotional import, empathetic connection - these are all primed and heightened in a narrative mode. Not only this,

[2] See for instance: Cuddy, Fiske, and Glick (2007), Feigenson (2016), Fiske et al. (2002), Fiske, Cuddy, and Glick (2007), Nadler (2012), Nadler and Mcdonnell (2012), and Rahimi, Hall, and Pychyl (2016).

[3] See for instance: Mazella and Feingold (1994) and Devine and Caughlin (2014).

[4] See for instance: Cuddy, Fiske, and Glick (2007), Ellison and Munro (2008), Fiske et al. (2002), Mazella and Feingold (1994), Mitchell et al. (2005), Sommers and Ellsworth (2000), Suedfeld et al. (1985), Westra (2018a), and Willemsen, Newen, and Kaspar (2018).

[5] See for instance:Alicke et al. (2008), Alicke et al. (2015), Fishbein and Ajzen (1973), Gerstenberg and Lagnado (2012, 2014), and Gerstenberg et al. (2018).

[6] See for instance: Collins (2000) and Smith (2000) in Suls and Wheeler (2000) and Alicke and Sedikides (2009) and Jefferson (2020).

[7] See for instance: Bayles (1982), Brewer (1977), Gailey and Falk (2008), Gawronski (2009), Lagnado and Channon (2008), Nadler (2012), Pizarro and Tannenbaum (2012), Shaver (1985), and Westra (2018b, 2019).

narrative explanation is a powerful tool by which we recognize (accurately or not) certain super-structures, plot types, tropes, character archetypes and so on. Stories have heroes and villains, protagonists, recognizable thematic connections, and, above all else, *make sense* to us in virtue of these arrangements. So, when we narrativize, we are tempted to fit the world into these structures, while at the same time ramping up the effects of framing, affect and empathy. I noted above that we are often over-confident in our ascriptions of responsibility. Our tendency towards narrative structure and explanation helps explain such over-confidence.[8]

2. **Status sensitivity errors:** The second class consists of what I call "status sensitivity" errors. These are more straightforwardly erroneous cognitive biases and heuristics that make too much (or not enough) of certain facts about identity, status, power, and social proximity in reaching attributions of responsibility. In particular, I am interested in the following asymmetry:

    (a) **Genuflection to Power and Position**: We are more likely to find excuses (and, paradoxically, to view as more generally responsible), those with power, privilege, or proximity to us in social position, and:

    (b) **Under-consideration of the Low-Status**: the flip-side of the above: we fail to give due consideration and care to those who have low social status. Interestingly, we are more likely to view them as non-responsible when this aids in their oppression, and more likely to hold them responsible when it maintains that status quo. Take, as an example, the prevalence of trying juvenile people of color as adults in the U.S. criminal justice system, paired with prevalent racist rhetorics claiming they are not as fully

[8] For work on our tendency to narrativize, and the likelihood that we do so in over-reaching ways, see: Gallagher (2006), Goldie (2009, 2012), Griffith (2019), Hutto (2006, 2012), Hutto (2016), Lamarque (2004), Morgan and Wise (2017), Roth (2017), and Schechtman (2011).

rational as their white counterparts. Importantly, status sensitivity errors will almost always involve and be intertwined with the interpretative errors I've noted above.[9]

My task now is to marshal empirical support for these claims, and to get clearer about the kind of epistemic injustice I have in mind. I want to show in more detail and with more evidence, that these kinds of problems *do* occur. Not only that, I hope to show that they are somewhat localized to moral responsibility, in the sense that problems I identify here do not merely reduce to other common kinds of cognitive errors. Justifying these claims will require philosophical as well as empirical tools. It will rely on our identifying the right kind of philosophical concepts – concepts that I claimed in Chapter One have been under-developed, or not noticed at all. Applying these new conceptual tools to the existing empirical literature, I argue, will help us identify the kinds of normative failures, epistemic vices, and errors in reasoning I describe above.

2.2 A model of the attribution of responsibility

In social psychology, one prominent field of research is "attribution" theory - roughly, when, how, and why we attribute actions, traits, intentions, beliefs, and other properties to agents. The foundations of attribution theory in Fritz Heider (1958)'s work involve: "identification of the invariant properties of people and features of the social environment," that is, outlining, "the conditions that will presumably lead a perceiver to decide that a behavior or event of interest was produced by a

[9] Indeed, much of the empirical basis for these errors overlaps with the kinds of resources I've cited above, particularly, the work of Alicke, Cuddy, Fiske, Nadler and Westra. But see also work in the "social comparison" and "social identity literatures: for good reviews: Buunk and Mussweiler (2001), Gerber, Wheeler, and Suls (2018), Suls and Wheeler (2000), and Suls, Martin, and Wheeler (2002) and for specific examples: Barden et al. (2004), Krueger (2000), and Zhao and Rogalin (2017).

dispositional property of the person involved, not by factors in the external environment" (Shaver (1985, 6)). One particular strain is concerned with the attribution of responsibility and blame. This literature serves as a foundational starting point for developing a descriptive account of the epistemology of moral responsibility, and I begin by constructing a core model of our attributions based on it. I do this in order to bring out the epistemic processes that are relevant for moral responsibility theorizing, and to identify when and where epistemic practices are captured by factors that may seem irrelevant from the point of view of the metaphysics of responsibility.

In confronting the literature on how we actually go about forming attributional judgements of responsibility, one is faced with an immediate choice point between three kinds of models (highly idealized for this sketch). The first kind of model is "first-personal," phenomenologically temporal, and effortful. It imagines attributing responsibility as a consciously aware, deliberative *process* – involving step-by-step, rational manipulation of information leading to principled decisions about what to judge. In this sense, such models are also highly normative: they specify how we *ought* to form judgments of responsibility. Call this kind of model "**Decision Theoretic**." The second kind of model is impersonal, non-temporal, and computational. It imagines responsibility judgements occurring automatically, perhaps entirely without conscious input, and models responsibility attributions as outputs of pure informational processing, often as the result of Bayesean updating schemas. This is not to say that such judgements are "cold" or robotic, however; in fact these models allow that much of the process may be affectively mediated and rely on emotional reactions as much as other strains of informational input. Call *this* kind of model "**Computational**." Finally, the third kind of model describes the process of arriving at attributional judgments as involving both first-personal, temporal decision making based on (at least semi) consciously held norms, *and* automatic or spontaneous information processing based on heuristics, biases and affective mediation. Call this a "**Hybrid**" model.

Although there is disagreement about how many factors, how conscious or automatic, how "hot" or "cold" attribution is and so on, there *is* broad agreement that our attributions are affected by many factors traditional philosophical accounts have largely neglected. These are, primarily, the kinds of distortions I listed under "Interpretive over-reach" above, that is: 1) personal proximity (whether the target is an in or out-group member in various ways), 2) outcome sensitivity (the severity, or goodness/badness of an action's outcome), 3) judgements of personal warmth and trustworthiness, and, 4) Stereotyping or categorization based on race, gender, religion, and so on.

However, two other categories of features *have* been discussed to some extent by moral responsibility theorists, and are also commonly thought to affect attributions of responsibility: 5) interpretation of: a) agential intention, b) moral character, and c) the social meaning of actions, and 6) judgments of free will and the ability to do otherwise. Given that only two of the six categories above have received treatment in the philosophical literature, and given that only (6) has received sustained attention from many *different* theoretical perspectives, it should already be obvious that whichever psychological model we pursue, engaging in a descriptive epistemic project can help fill in aspects of the lacuna I described in chapter one.

I'm going to argue, in keeping with the prevailing consensus in the contemporary literature, for a hybrid model of responsibility attribution. So that this does not appear *ad hoc* or unmotivated, let me try to describe a few of the drawbacks of purely Decision Theoretic and Computational models before giving a positive argument for the hybrid strain. Very roughly, both kinds of one-track model suffer from a narrow focus that fails to take into account recent developments in dual-process modelling: our attributions depend not *only* on consciously accessible rational deliberation, non-conscious informational modelling, *or* socially mediated affective pressures, but on a blend of all three.[10] To see why a blend is necessary for good modelling, consider

[10] Guglielmo and Malle (2017) have recently argued that this worry is overstated. Their "Path

one input to our attributional judgements: causal attribution.

How is it that agents actually go about attributing causal contributions in determining responsibility? Psychologists in the attribution literature have, in answering this question, given several kinds of answers. Alicke et al. (2015) identify four distinct "metaphors" in the history of psychology of "reasoner as *x*." First, following Fritz Heider (1958)'s influential model of "personal force" and Harold Kelley (1973)'s work on analysis of variance (ANOVA), we get the metaphor of attributors as "scientists." Scientific reasoners focus on distinguishing environmental factors from agential ones, implicitly using the same kinds of statistical models of variation as ANOVAs. There is something quite obvious about this metaphor - after all, the crucial question in causal cases is often *if* and *how much* an agent contributed to an outcome. However, this kind of analysis was historically dominated by behavioralist thinking, and often discounted "mentalistic" inputs such as intention and motivation. Furthermore, it did very poorly in explaining more complex situations with multiple agents.

The second metaphor is of attributor as "lawyer," using Hart and Honoré (1959)'s model of legal reasoning about responsibility to paint moral reasoners as legalistic scrutinizors. Psychologists like Brewer (1977) describe agents as investigators, factoring in association, causality, foreseeability, and intentionality to assign blame. Although this model does better in including mental inputs like forseeability and intentionality, as well as dealing more directly with causal history, excuses and justifications, it still treats responsibility as something with rather narrow scope. That is, larger issues of disposition and character, as well as non-"material" harms like distress or social disfavor are not well represented.

Model" of blame tests whether responsibility attributions tend to follow a specific informational processing path (they say yes). Yet, as I hope will become clear, this is compatible with a hybrid model. That there is *usually* a well defined informational path does not preclude that in the real world information comes to us in messy fashion, and parts of the path can be skipped, re-routed, looped, and affectively mediated. They seem to admit as much about non-laboratory settings. As they say: "naturalistic search strategies may often proceed in more nuanced ways, rather than strictly or solely along the steps of the Path Model" (968).

The third metaphor is that of attributor as “reconstructor.” Building on pioneering research in the biases and heuristics literature (from, for instance, Kahneman and Tversky (1982) and Bargh and Chartrand (1999), the “rationality” of previous metaphors is downplayed in favor of counterfactual modelling that is sensitive to “abnormalities.” Very roughly, we construct simulations that we expect to match the real world. The more surprising an agent’s actions are, the more “uphill” reasoning we have to do to get our model to match their actions, and the more likely we are to blame them as a result. As an example, we are more likely to blame a driver who strikes a pedestrian when they spontaneously decided to take a new route home, than one who strikes a pedestrian on their daily commute.

Finally, and most recently, we have the metaphor of attributor as “moralist.” This model can be seen in the flourishing of recent experimental philosophy and moral psychology, as in the work of, for instance, Knobe and Fraser (2008) and Knobe (2010), as well as in the revival of interest in moral intuitionism (for instance, in Haidt (2007)). Here we have the full expression of the complexity of moral reasoning, including the fact that it includes more than merely probabilistic or counterfactual reasoning, is sensitive to biases and heuristics, and often involves the affective reaction we have to narratives about agents. Key to this model is the idea that “One of the most important ‘extraneous’ influences on causal perception is the positivity or negativity—or *evaluative tone*—of the agent’s motives, character, or the event’s outcomes” (Alicke et al. (2015, 804)). That is, our spontaneous evaluations of agents on a positive-negative scale has as much to do with our tendencies to blame and praise them as ANOVA modelling, legalistic reconstruction, or counterfactual reasoning.

2.2.1 Temporal, Decision Theoretic Models

These metaphors, again, deal only with the ascription of causality, but they serve as a motivational nudge for our choice-point concerning a Decision Theoretic, Computational, or Hybrid model. The first-personal, temporal, step-wise, decision-theoretic model may appear attractive insofar as it matches up very nicely with the philosophical literature on responsibility and blame. It tries to capture what agents think and feel as they are determining responsibility (although, we must be clear that many of these theorists would happily admit that this is an abstraction and that attributions often occur, at least in part, non-consciously or near-instantaneously). Here is a purely temporal, Decision Theoretic model (somewhat similar to Shaver (1985)'s influential model of attribution):

1. A thing happens.
2. We seek causal explanation for that thing – why did it happen? What caused it?
3. We determine that an agent was (or agents were) involved.
4. If agency is implicated, the following questions are asked and answered:
   (a) To what degree did the agent contribute to causing the event to occur?
   (b) How good or bad was the event?
   (c) What does the agent's action mean? That is:
      i. What did they intend to do?
      ii. What does that intention say about their character?
      iii. What does the outcome say about their character?
      iv. Have we made repeated observations of this person in similar situations? If so, how does this action fit in to a larger pattern?
5. Based on the answers we give to these questions, we arrive at an attribution that:
   (a) Specifies whether and to what extent we think the agent was responsible,

(b) Moves us in the direction of a reactive attitude or moral emotion, and

(c) Updates our expectations about the agent and their character.

However, given the discussion of the history of causal metaphors above, we can already see some problems with the temporal "stage" model. When do these steps occur? And do they really occur in this order? Do we always proceed *to* blaming *from* causal attribution? Or, can reactive attitudes and moral emotions color our perceptions of responsibility itself? There are also difficult questions to answer about the way that each step occurs and how cross-situationally and cross-personally consistent the steps are meant to be. Although a stage model may at first appear attractive, it focuses too much on agents as rational, legalistic reconstructors, searching to develop stable models of agents and their character, and to attribute praise and blame narrowly and without emotion.

In addition, the model is overly prescriptive. That is, it presents a model of optimal reasoning that agents *ought* to follow. One central problem with the prescriptive nature of step-by-step theories, is that it assumes a model of reasoning that may or may not accord with reality. Without getting the descriptive facts about everyday agents right, our prescriptive theories run the risk of being disconnected from our lives, as well as lacking justification for their claims. That is, without further justification we cannot assume that human beings who deviate from the prescriptive model are getting things wrong. Although the model above might be what we'd hope for from a juror in an important trial, it isn't necessarily what we'd see (or want to see) from an everyday reasoner about responsibility.

2.2.2 A-temporal, Computational Models

Given the problems outlined above, one might be tempted to go purely a-temporal or Computational. In particular, since Decision Theoretic models are not sufficiently attentive to the actual cognitive and affective processes that agents undergo when they attribute responsibility, we'll want to carve out space for including those factors. Purely computational models hearken back to the metaphor of attributors as "scientists," as in Brewer (1977) and Kelley (1973), although they may also allow for counter-factual modelling as an important component. Contemporary computational models also exist, such as Gerstenberg et al. (2018), Guglielmo and Malle (2017), Malle, Guglielmo, and Monroe (2014), and Rai and Fiske (2011)'s, although even these tend to be at least somewhat hybrid in nature, as I'll explain below.

Here is a purely a-temporal, informational-processing based Computational model:

1. A thing happens.

2. The brain computes various counter-factual probabilities about agential involvement and what the action says about an agent's moral character.

3. These probabilities are compared (non-consciously) to a pre-existing schema of expectations, which can be richly populated with previous observations about the agent or nearby similar agents, or a mere sketch of likely possibilities.

4. The resulting difference between expectation and event result in some attribution of responsibility (in terms of amount and type).

It should be obvious that this model has serious limitations. It neglects, for example, the very important social mediation that takes place in conversation about responsibility and blame. Of course, a good Bayesian may argue that conversations and other social feedback can merely be folded in to updates of our schemas and models. But, in making distinctively consciously effortful and phenomenologically felt experiences purely computational and "brain-based," such a model both alienates ordinary agents from an intuitive sense of how these processes normally operate and threatens to unnecessarily obfuscate the relationships between our conscious thought and these mathematical models. Another basic problem, one felt even in Heider's day, is that presenting man as a scientist necessarily involves presenting him as a rather *bad* scientist. Human beings are notoriously irrational. So, even if these computational models reflect some part of the story about our attributional processing, they cannot be the whole story, given that ordinary humans just don't stick to the outputs of probabilistic models. If we are lay ANOVA-ists, it is not a hobby where we aim for perfection.

2.2.3 The Hybrid "Ping-Pong" Model

Given the drawbacks of each of the pure models, and following theorists such as Alicke (2000), Alicke et al. (2008), Alicke et al. (2015), and Rahimi, Hall, and Pychyl (2016), among others, I advocate for a hybrid or dual-process model. There are both conscious, deliberative aspects to the attribution of responsibility and non-conscious, "spontaneous" ones. The business of responsibility attribution is fundamentally normative, and normativity is messy. There is, to put it plainly, no clean break or distinction between aspects of the model which are affected by biases, heuristics, and affect, and those which are not.

This is true in three ways. First, the dual processes *can* remain separate, but in many normal cases they will interact. Second, the contextual information that conscious deliberation makes use of often already includes biased, irrational, or morally irrelevant factors. Although spontaneous evaluations can be arrived at without "emotion" (that is, although model updating needn't always be affectively mediated), many aspects of our non-conscious evaluations will already arrive in our conscious mind tinged with affect. Third and finally, as Peter Railton (2014) has argued, non-conscious processes do not need to be "dumb" – they can be rational and responsive. And, on the other hand, conscious and deliberative processes can be "hot" and affectively mediated. As Alicke et al. (2015) remind us, "Research in [contemporary psychology] makes a strong empirical case for the effects of emotions on moral judgments and...causal attributions" (806). There is no guarantee that consciously deliberating about an act's meaning or an agent's quality of will won't be responsive to our emotions or biases.

In summary, although both temporal, step-by-step decision making and automatic informational processing are featured in the hybrid model, we can make clear that agents are not merely reconstructing (abstractly, cooly) what happened in a series of events, *nor* are they reacting purely automatically, emotionally, or nonconsciously. A better model, then, keeps much of the structure from our temporal and a-temporal models, and focuses on the ways these paths combine and interact. It also makes clear that the process will sometimes output a single, more or less unified, attributional verdict, and that at other times the dual processes will result in independent, and perhaps inchoate, judgements that an agent may hold until they realize their incompatibility (although even then they may continue to behave irrationally and affirm both to some degree). This will be clearer with an explicit model and an idealized example to interrogate.

Here's a very simple kind of example to motivate the model: Two friends, Maria and Hans, see a mutual acquaintance, Kat, interact with a young child. Maria thinks she sees Kat sneer at the child, who looks mortified and runs away. Immediately, she forms a negative judgement of Kat's character, somewhat mediated by her previous interactions with and knowledge about Kat as a generally good person. Nevertheless, she feels that Kat is responsible for terrifying the child and turns to Hans ready to blame Kat for treating the child unkindly. She feels a rush of blood to her face and her heart rate is elevated. Hans, she sees, is laughing to himself. "What's the matter with you, why are you laughing?" she asks. Hans, it turns out, saw something quite different: a game that Kat and the child were playing, where Kat pretends to be a monster, and the child runs away to hide. He explains this to Maria, whose emotions cool, and who no longer feels that Kat is blameworthy. Despite this shift in affective tone, she has some heightened sense that Kat is not quite as trustworthy as she once seemed to be.

Notice the many complex and layered deliberative and affective processes that Maria was undergoing and initiating. Notice that her reasoning was not, in particular, focused on Kat's capacities (although she certainly had a background schematic model of them). Notice the interplay of historical information about Kat, judgements of character, real time social feedback and so on. Finally, notice that Kat was able to revise her judgement, and that her emotions played a key role in several of those steps. Responsibility attribution, as our model endeavors to make clear, is complicated. For a visual representation of the model, see Figure 2.1, below.

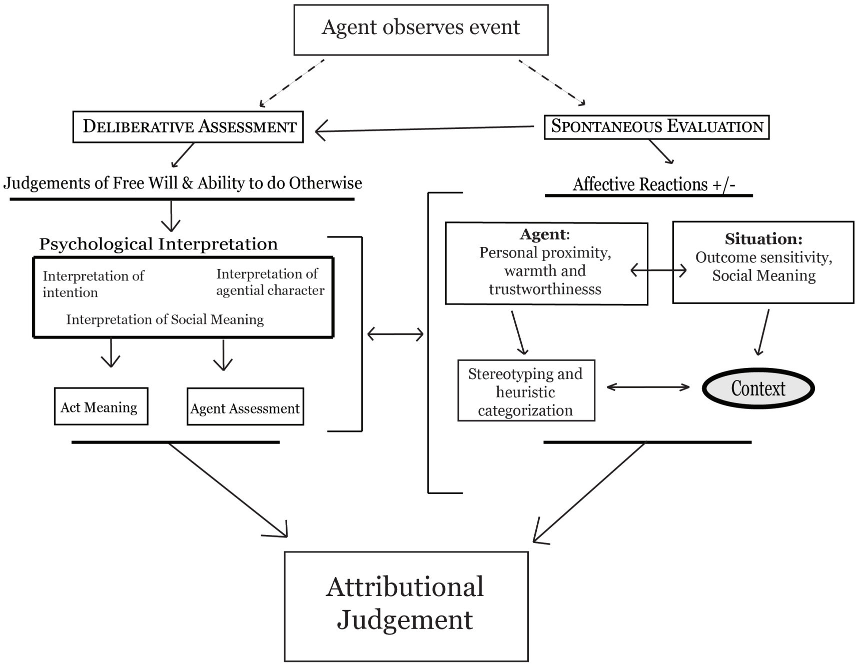
Agent observes event
Deliberative Assessment
Spontaneous Evaluation
Judgements of Free Will & Ability to do Otherwise
Affective Reactions +/-
Psychological Interpretation
Interpretation of intention
Interpretation of agential character
Interpretation of Social Meaning
Act Meaning
Agent Assessment
Agent: Personal proximity, warmth and trustworthinesss
Situation: Outcome sensitivity, Social Meaning
Stereotyping and heuristic categorization
Context
Attributional Judgement

Figure 2.1: The Ping-Pong Model of Responsibility Attribution

Here's a textual re-presentation of the model: we begin at various points on the flow-chart after the observation of an event. That is, as the hybrid-model endeavors to make clear, there is no *singular*, static mode of informational presentation and processing, although there is a more or less “standard one.” This lack of singularity is realized in several ways. First, there are different entry points into the model: Our snap decision to blame someone may be the *beginning* of our reasoning about responsibility, for instance. Second, in any case, once one views a responsibility relevant event, hears a responsibility relevant narrative, considers an agent as responsible, or engages in dialogue about responsibility with others, two streams of evaluation are activated. The first is more or less automatic (Spontaneous Evaluation), and occurs without conscious deliberation. It’s outputs are affective reactions and models of agents and situations. The other (Deliberative Assessment) does involve conscious deliberation. In this sense, it is strictly unnecessary in terms of arriving at an attributional judgment. But we ignore it at our peril. The idea that most judgments of responsibility are fully non-conscious and non-deliberative is surely overstated – especially when we think of responsibility as a socially mediated, dialogic, practice based enterprise.

The key things that this model makes clear are as follows:

1. Once we’ve determined that an agent is involved in a responsibility salient event, both automatic and conscious processes can begin to unfold. There is no guarantee that conscious deliberation occurs. This is unsurprising, as many more salient events pass through the horizon of our experience than we could possibly consciously assess in any given day.

2. Second, neither process needs to run its full course. That is, we might get some affective reaction to an agent, but then become distracted by something else, or begin to assess someone’s causal control, but then decide to move on. Just because the processes start, it doesn’t mean they must end in an output of attributional judgement.

3. Third, on this model, the automatic and deliberative are multiply interactive. Affective reactions, for example, can inform our assessments of agential capac-

ity. So too, schematic models of agency expectation can affect our psychological interpretations of an act's meaning and an agent's character. Crucially, the relationship between non-deliberative and deliberative is bi-directional: it's also the case that our deliberative processes can affect our spontaneous evaluations. What we determine about an agent's reasons-responsiveness or quality of will, for instance, can have an effect on further affective reactions towards them and the situation.

4. Fourth, the kind of broader, agential "context" discussed in Chapter One (agential history, situational factors, conversational influences, local norms, asymmetries of power, facts of identity like race, class, and gender, and so on) influences our modelling on both the deliberative and spontaneous paths. In spontaneous assessment, contextual factors change (and are colored by) our affective reactions and our schematic modelling and updating. On the deliberative side, context affects our assessments of act and agent meaning in ways that should be familiar from my earlier arguments.

5. Fifth, the model makes clear how crucial *interpretation* and *modeling* are to arriving at an attributional output. A deliberator must make many interpretive choices, and our non-conscious models of agents substantively affect our expectations of and reactions to them.

6. Sixth, and finally, the model allows space for the social to be a key driver of attributional outputs. This is true both in the inclusion of social *norms* as inputs for our reactions to and interpretations of act and agent meaning, and in the contextual allowance for literal (and imagined) conversation to affect our judgements.

For these reasons, I call the model of responsibility attribution that emerges the "Ping-Pong Model." Its core insight is that judgements about responsibility start from diverse inputs, and then "ping-pong" back and forth between conscious deliberation and non-conscious (or automatic) reactions, emotions and mental processes. Each "volley," to continue the metaphor, returns a proto-judgment with a certain trajectory on it - spin and direction which affects the next phase of the attributive process until a final judgement is reached. So, for instance, I may judge you harshly upon hearing about something you've done, then see a picture of you and have an affective reaction of great warmth or pity which modifies my initial judgement, then recall things you've done in the past, and so on. Of course, all of this is metaphorical and idealized. This kind of deliberation can be very quick or very drawn out: judgements may take place entirely non-consciously, or entirely through effortful deliberation. Even so, the psychological literature bears out the kind of model I'm sketching here, and it is useful to have such a model to deploy in order to help us notice the kind of distortions I'm seeking to explore.

2.3 The Ping-Pong model in action

Let's consider a couple of examples from recent empirical and theoretical work, to see what the Ping-Pong Model predicts about agential behavior and the output of responsibility judgments.

Example 1: Drunk-Driver Blame

Simon drank three beers at a local pub. Approaching an intersection near his house he looked down to adjust the radio. Upon looking up, he saw a pedestrian crossing the street, but didn't react quickly enough to brake or swerve around him, striking and killing him. Mary reads about this story in a local newspaper. How does our model predict that Mary will react?

First of all, the model predicts that the way in which Mary hears about the event matters. The fact that she reads about it in a newspaper, rather than seeing it first-hand, or hearing it from a friend, colors the affective reactions she is likely to experience. The model predicts that Mary will have a spontaneous reaction towards Simon. Given that she does not know him personally, she will rely on a general schema she has about drunk drivers and their level of responsibility and blameworthiness. Her schema will be updated with the new information from the story. She will also have spontaneous affective reactions to the story itself. It is likely that these will be strongly negative, but not highly arousing, given her disconnection from events. Of course, all of this depends on facts about Mary herself – if, for example, a friend of hers was recently killed by a drunk driver, the affective reaction *is* likely to be highly arousing.

These spontaneous factors go some way towards outputting a model of the kind of agent Mary thinks Simon is. At the same time, they feed into her conscious thought. We can imagine that Mary, more or less indifferently and without much deliberation, considers the evidence in the newspaper report about Simon's capacities, foreknowledge of likely outcomes, and intentions in getting in a car while buzzed. It should be clear that such evidence is scant! Mary doesn't have much to go on here – there isn't likely to be an interview with Simon, for example. Nevertheless, it seems likely that Mary will construct a psychological interpretation of the kind of person Simon is (agent assessment) and what his action represents about him (act meaning). This is largely going to take the shape of the vague outputs suggested by her spontaneous reactions: her schematic models of similar situations and agents, and her strong negative affective reaction to the story.

Here again, there is space for a variety of factors to influence her thinking. Perhaps the newspaper strongly reminds her (although it is unlikely that she needs reminding) about the local norms against drinking and driving. Perhaps her wife is also reading the same story and remarks to her, "what an asshole!" Such social condemnation is likely to influence her own assessments of the act and the agent.

Given that the newspaper doesn't present her with any evidence that there were excuses or justifications for Simon's actions, or situational defeaters that left him unable to do otherwise, Mary arrives at an assessment of his responsibility (with, perhaps, further cascading affective reactions), and, should she continue thinking along these lines, an attributional judgement: Simon is responsible for the death of an innocent pedestrian.

The key thing to notice here is that our model suggests that she can arrive at this kind of judgement without explicitly considering (or even possessing) any of the relevant information that most theories of moral responsibility seem to require us to have. That is, we could further clarify the story above in the following two ways:

(1) Simon was intoxicated enough that his reflexes were slowed, and, perhaps, his rational capacities diminished. Nevertheless, he was capable of reflecting on whether driving home was safe. He decided it was not, but reflected that it would also be inconvenient and expensive to get a rideshare. Because of these factors he chose to risk it (being perfectly aware of the risks).

(2) Simon was coerced into having the third drink by his boss. He was going to leave after his customary two rounds of drinks with co-workers. But his boss showed up and pressured him into doing a shot with her. Afraid he'd be passed up for a promotion if he didn't go along with it, he acquiesced. Not only this, the final drink substantially diminished his rational capacities. He was then incapable of rationally reflecting on whether driving home was safe. He (irrationally) deliberated about other modes of transportation, but decided that it was safe for him to drive.

Now, it's clear that whether (1) or (2) was the case in the actual world *ought* to matter for our assessment of Simon's moral responsibility. But, in almost all cases of attributional judgment, we don't have a story like (1) or (2). We arrive at an interpretation in conditions of scarce evidence. I'll have a lot to say about this situation in later chapters, but the point now is just that we don't *need* such a story to come to what seem like perfectly competent judgments of responsibility.

The worry, for those interested in how often and how seriously our attributions might go wrong, concerns how good those “competent” judgments are likely to be. The looping, dynamic nature of the system means that, once a negative evaluation enters the picture, it is far more likely that the agent will be found responsible, as a way to ensure that it was appropriate to blame theme. Blame is “sticky” in this sense - once it attaches to an agent, it is hard to wash off.

Alicke (2000) calls this a “blame-validation” or “blame-first” mode of cognitive processing, and argues that such processing is rampant in attributional judgment. Returning to our metaphor: negative affect has an outsize effect on the trajectory and “spin” of our judgments. Once a negative evaluation attaches to an agent, it is likely to stick all the way through our judgemental process. Indeed, given enough negative affect, it’s likely that a snap judgement will be arrived at without much processing or deliberation at all.[11] If a serve has enough spin on it, to return to our metaphor, there may not be much of a volley to engage in. Non-metaphorically: if we begin with a blame judgement before considering other interpretive factors, at least the bare affect of unfavorableness is likely to stick. Our emotions will run hotter, we will be oriented towards confirming evidence of “badness,” and we are likely to overestimate causal control and more easily find unfavorable aspects of people’s characters. The stickiness of negative attributions in general, and blame in particular, is something that the Ping-Pong Model is primed to explain and represent.

Example 1a: Speeding John

Here’s a variation of the above example for which we have some experimental results from Alicke (1992) and Nadler (2012) concerning people’s reactions to two slightly different scenarios:

> John was speeding to get home, driving 40 mph in a 30 mph zone. He came to an intersection and applied the brakes but was unable to stop

[11] This is the core of Alicke (2000)’s arguments, but the relationship between emotion, affect, and blame is complex and also explored by, for instance, Feigenson (2016), Fiske et al. (2002), Fiske, Cuddy, and Glick (2007), Nadler (2012), Rahimi, Hall, and Pychyl (2016), and Weiner (2006)

> in time because of an oil spill on the road. John hit another car in the intersection, injuring the other driver. John was speeding home in order to: a) Hide from his parents an anniversary present for them that he had left out in the open, OR b) Hide from his parents a vial of cocaine he had left out in the open. (Nadler (2012, 7–8))[12]

We can ask two questions about the example. First, how does our model explain the shift in people's responsibility attributions given the (a) scenario or the (b) scenario? Second, how do these explanations align with the experimental results presented by Nadler, Alicke and others? To answer the first question, we need to understand that attributions of responsibility, on the Ping-Pong model, are mediated by positive and negative affective reactions at several distinct junctures, and that these effects loop and interact in messy volleys. The upshot is that our assessment of John's intentions and character, and our purely affective reaction to him as a "good" or "bad" person, make quite a difference.

In scenario (a), it is quite likely that, while we may think John is acting irresponsibly as a motorist, we also think that he is being a good son. In scenario (b), there is nothing to distract us from his poor choices as a motorist, *and* at least some of us are also likely to form a generally negative assessment of John's character as a drug user. These kinds of mixed evaluations are prevalent in day-to-day life. The lack of "clean" cases, as it were, means that our responsibility attributions are unlikely to be a simple matter of checking for capacities and quality of will. That is, our very determination of facts *about* the quality of John's will loop back and affect our understanding of his capacities. As Nadler notes: "Not only do people think John is more responsible, but they also think he is more of the cause of the accident when the object he was thinking about hiding was cocaine, rather than a present" (8). Now, Nadler presents this as an "oddity," but our model would predict precisely this kind of looping relationship. Our negative assessment of John in the cocaine condition

[12] Originally presented in Alicke (1992), "Culpable Causation," *Journal of Personality and Social Psychology*, 368-369.

leads us to ascribe more causal efficacy with regard to the negative outcomes of the case. And, given this context, we are likely to give more weight to his responsibility.

It's clear in this kind of case that our judgments of causal control are often a prime mover (and that this is often a major source of error). When we look at our model, there are four distinct inputs at the level of capacity assessment that our immediate affective reactions can color: causal control, volitional control, knowledge of norms, and knowledge of likely outcomes. In both the (a) and (b) scenarios, John's volitional control, and knowledge of norms and likely outcomes don't change. Whether he is preparing to hide cocaine or a present, he has the same understanding of moral (and driving) norms and likely behavioral outcomes, and (assuming we are not considering a certain class of addiction cases) the same volitional control over whether he speeds.

What's doing the work is our tendency to inflate or diminish the kind of causal control John has over the outcome. This is why I identify one major source of errors in the epistemology of responsibility as the *over or under-estimation of causal control*. Before we move on from John, however, we should remind ourselves that mediation by positive and negative affect can occur on any of the four outputs mentioned above, *and* at several other junctures in our core model. Still to be discussed then, are: 1) times when affect influences our estimation of an agent's epistemic capacities or their volitional control, 2) the messier interpretive work of determining something about the meaning of an action or agential character, and, finally, 3) further reflections on the ways in which our initial reactions to someone in terms of warmth and competence, as well as how good or bad the consequence or outcome of an action are can influence our attributions of responsibility.

2.3.1 Informational Processing: No, really - why not a "path" model?

One thing that's become clear so far is that the way and order in which information is processed matters a great deal in terms of an attributional output. For all I've said against Decision Theoretic models, then, wouldn't it be simpler and more accurate to specify an idealized model of responsibility attribution that is a single stage or step-by-step model of informational processing? Guglielmo and Malle (2017), building on Malle, Guglielmo, and Monroe (2014) offer precisely such a model for blame.

According to them: "An information-processing framework of moral judgment would specify the information input that guides people's judgments and the psychological processes that operate on this information to generate the judgments... by [specifying] what information people acquire, and in what order, en route to... [moral] judgments" (2017, 957). If we could specify such a framework, we'd have a clear picture about what information people seek out (and in what order) when they go about making responsibility judgements. If it looked like people followed a similar path enough of the time, then we'd be vindicated in plumping for a single-track model that fit (nearly enough) most canonical cases.

Guglielmo and Malle are particularly interested in developing such a framework for the informational processing pathway of *blame* judgments. Their Path Model of Blame, "asserts that information processing toward blame begins with a social perceiver detecting a norm-violating event. The perceiver then assesses causality, determining who or what caused the event. If the event appears agent-caused, the perceiver determines whether it was intentional or not. If intentional, the degree of blame depends on the agent's reasons for acting; if unintentional, the degree of blame depends on whether the agent could have prevented it" (958). For a visual representation see Figure 2.2:

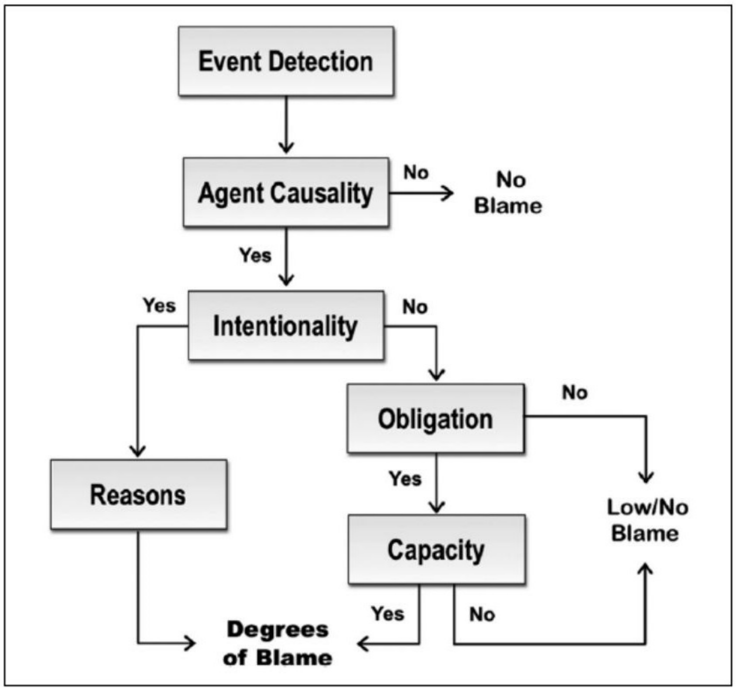

Figure 2.2: Guglielmo, Malle, and Monroe's Path Model (2014)

Although I'm going to take issue with the applicable scope of the model, these empirically confirmed conclusions are incredibly helpful to keep in mind. Blame is social, depends heavily on causation, intentionality, interpretation of reasons, and forseeable prevention. I take all of this on board happily.

So what's the problem? Guglielmo and Malle note that many psychologists dispute the plausibility of this kind of sequential processing. According to the "non-sequentialists" there is no reason to expect any particular path - a belief Guglielmo and Malle set out to dispel. However, I think we should ask if this reconstruction of their opponents is quite right. I take the "non-sequentialist" argument to be something like the following: 1) whatever informational path we follow, it will not be purely deliberative. 2) Given this, such a path will be loopy and multiply interactive. That is, it will go between the deliberative and non-deliberative streams, and blend and return to various levels of the informational picture Guglielmo and Malle paint. 3) Therefore, there is no single *clean* path that informational processing takes. Hot blooded anger may precede a deliberative search for intentionality, and interpretation of reasons may depend on norms about foresight and prevention.

We shouldn't, however, ignore the empirical results Guglielmo and Malle tout. As they say:

> The results demonstrate that the constitutive information components of blame are not on equal footing. People did not simply accept any type of information whenever they could get it. Rather, they showed clear preferences in their information acquisition: causality information had processing priority over intentionality information, which typically had processing priority over reasons and preventability information. These patterns are consistent with the Path Model but not with nonsequential accounts. (961)

Importantly, these results were consistent in settings where agents had plenty of time to deliberate, *and* in those when they had to make very fast decisions (966). But taking all of this into account we can still object that the non-deliberative is given short shrift by Guglielmo and Malle. If there is non-conscious information that an agent receives prior to some cognitive step, and if that colors the cognitive step-by-step process, this isn't "non-sequential" in a way that would change my informational search preferences by *type*. That is, anger might not cause me to want information about intention before information about causation. What it would do is change my interpretation of the causal or intentional information. And, although it seems right that we have some tendency to follow certain evidential paths in our information processing, this doesn't seem to prove anything like necessity or finality. Some derivative pathways *were* chosen at lower frequencies, and not much attention is given to cases where earlier steps were returned to at later stages.

Just as importantly, as Guglielmo and Malle's studies makes clear, the questions people ask are about causality, intentionality, reasons, and preventability; very rarely were they about things like responsibility, controllability or character. Of course, some of this may be do to experimental set up - certainly people ask questions about controllability, agential character, and responsibility in real life. In other words, it's no surprise that the studies in lab-environments don't replicate day-to-day life: real life is messy. It might be true that informational search occurs in some generally standardized ways (and this is actually helpful for us to know), but, in naturalized settings we get information randomly, not at all, in emotionally fraught ways, in the context of narrative vignettes and so on. And Guglielmo and Malle admit as much. As they say:

> Naturalistic search strategies may often proceed in more nuanced ways, rather than strictly or solely along the steps of the Path Model. First, perceivers might receive or acquire information about other blame-relevant features (e.g., about an agent's character or past behavior) or seek to clarify components of the Path Model by obtaining more fine-grained information (e.g., about an agent's effort or planning as clues about the intentionality of the behavior). Second, some negative events afford strong inferences about certain information components (e.g., intentionality) and thereby obviate the need to search explicitly for such information. Assault, for example, is almost certainly intentional, whereas fire blazes are often caused unintentionally... [p]erceivers' own ideological or personal commitments may also guide their information acquisition, leading them to seek or interpret certain information in preferred ways. (968)

I can't think of much that needs to be added. While Guglielmo and Malle are right to point out that we can do more to determine the regular pathways that information travels over (and the mechanisms by which it travels), it's clear that a hybrid model, taking into account the messiness of life, psychology, and responsibility, is what's needed for our philosophical purposes.

Sidebar: Responsibility or blame?

Before we move on, it seems a good time to address something that I'm sure my readers have noticed. One issue that occurs again and again in reading the psychological literature through a philosophical lens is that "responsibility" and "blame" attribution are often used interchangeably. As much contemporary work in philosophy makes clear, moral emotions, capacitarian ascriptions, and accountability practices can (and in many regards, ought to) be kept conceptually distinct. Let's say that Josie has stolen money from JiMin. We might have a negative affective evaluation of Josie's character ("Boo! Bad!"), a rational assessment of her capacities and situation that delivers a responsibility-applicable verdict ("she was reasons-responsive and had fair opportunities, and thus is responsible"), and an overall output in the form of a reactive attitude (i.e., we blame her), and even a further judgement about what that blame might open her up to (i.e., we might feel that she ought to be punished). It's clear that these various aspects of responsibility, although deeply intertwined, can be, in principle (and in practice) separated. We can blame while being unsure of an agent's capacities or without having a particular emotional reaction. We can have a strong emotional reaction and decide (rationally) not to blame, or have a weak emotional reaction, deem an agent has the relevant capacities and opportunities, but decide they ought not to be subject to social or legal sanction. Any set of combinations is possible.

I am sympathetic to the kind of "blame-first," model endorsed by Alicke and others, such that specific reactive attitudes or outputs of responsibility reasoning are themselves also inputs into our responsibility attribution architecture. However, being too loose with this kind of thinking can cause theoretical problems. For one thing, as I'll discuss below, it too easily collapses negative assessments, attributions of responsibility, and *blame* as a particular moral response. One good thing about the ping-pong model is that it makes clear how these distinct features *can* collapse – but also that they needn't. Insisting that blame-validation is a primary feature of our moral reasoning may be going a step too far. There are good reasons for wanting to keep these things conceptually distinct.

What the psychology literature helps bring ought, however, is that this conceptual unhinging of various aspects of responsibility practice is not robustly realized in day-to-day life. Our emotional reactions, senses of capacities and opportunities, and decisions about whether to praise or sanction operate in various *loops.* So, my sense of you as negatively valanced, might lead me to blame you for an action before I ever consider your capacities. This in turn can help me engage in reasoning that is more like to see you as capacitous, and thus more likely to be a candidate for sanction. The way we arrive at responsibility judgments in real life, in other words, does not follow a logically consistent step-by-step process. It is multiply iterated, messy, and full of feedback.

2.4 Theory of Mind: mindreading and mixed approaches

It's clear at this point that our attributions of responsibility have much to do with our initial *impressions* or *perceptions* of people, as well as our downstream *interpretations* of the meanings of their actions. In order to flesh out these concepts it will be helpful to put some contemporary work in philosophy of mind in contact with the psychological literature I've been exploring. Moral responsibility theorists often assume our ability to intuit states of mind such as beliefs, desires, and intentions – even if they admit that this is a complicated process. As we saw above, mindreading isn't easy. Not only are there basic problems of opacity and interpretation, but our attributions of mental states can be biased, are looped into distinctive cognitive processes, and are often affected by our emotional state. As Evan Westra (2019) puts it: "would-be mindreaders face a persistent challenge: behavior is quite often ambiguous, and consistent with many different possible mental causes. A smile from a stranger on the subway, for instance, could be a signal of recognition, an act of flirtation, an absent-minded reverie, or simple politeness. A shout from a neighbor's apartment might be an outburst of rage from a domestic disturbance or excitement at a sudden turn of events in a football game. Inferences from behavioral effects to mental causes are always underdetermined" (2821). It's clear, therefore, that a deeper dive into mindreading is necessary.

In this section I argue for the following three points:

1. Mindreading is a distinct and important psychological ability involving *both* automatic and intentional processes.

2. Mindreading makes use of character modelling to do important cognitive work.

3. Mindreading is irreducibly interpretive, and often relies on cognitively "penetrated" perception.

When we looked at the Drunk-Driver Blame and Speeding John examples, one thing that became clear was that what is under consideration in responsibility attribution is not merely an amorphous agent or isolated incident. Rather, we have a more holistic view of a *person* engaged in intentional action. One aspect of this "person perception" is reasoning about their mental states. Another (related) aspect is reasoning about their character traits. A full story of how, when, and why we develop a folk psychological picture of mind and character is impossible in this chapter. But it is crucial that we understand some of the features and mechanisms by which these things occur, as well as their ubiquity.

My discussion of Guglielmo and Malle's Path model made clear that, in laboratory settings at least, reasoners search for some basic and predictable types of information first. Once we know that an agent with specific intentions is involved, their reasons matter to us, and it is often these that we try to uncover. But, as I've stressed again and again - life is messy - and we often do not (or cannot) have access to those reasons directly. We must infer them. Here is how Westra (2019) describes the situation that moral reasoners encounter:

> When navigating social environments, we must somehow sort through these potential mentalistic causes, and arrive at the most probable interpretation. These abductive inferences require us to draw on our own background knowledge to fill in the gaps between behavioral observation and mental cause. Sometimes, we may fill in these gaps with our knowledge of the mindreading target herself and her individual history: if we know someone well, we are often able to infer what she is thinking quite accurately. But just as often, we interact with complete strangers, about whom we know nothing. In these cases, we may instead fall back on stereotypes about the target's social group membership. And this is a point where pernicious social biases can enter into the mindreading process distorting our interpretations of the social world (2822)

Westra highlights several key aspects that are driving the focus of this chapter. First, that there are informational gaps about responsibility that must be filled in – gaps that we attempt to fill with social and historical context. Second, that this filling leaves the door open for unjust, biased, and ”pernicious” shortcuts and distortions. Third, that the informational processing necessary to arrive at attributions of blame is *inferential*, or, as I've been putting it: "interpretive." Given all of this, how is it that we go about interpreting one another's psyches? I'd like to put to one side debates in theory of mind over whether theory-theory, simulationism, or some hybrid model best explains our general folk psychology. I'll return below (very briefly) to the debate, but I don't think much hinges on it for our purposes. Whatever the right theory is, we'd expect some of the same inputs to matter for our concerns.

First and foremost, there are decades of psychological research that make clear that we, almost immediately, assess and begin to categorize people along two axes: those of *warmth* and *competence*.[13] From the first moment we encounter individuals, we make judgments of their warmth and competence – judgements that appear to be as perceptual as they are cognitive. It is a truism, of course, that first impressions are everything, and this particular truism is born out by psychological research. Warmth

[13] Cuddy and Fiske's work on character, warmth, competence, and stereotyping is particularly instructive here.

and competence assessments loom large as precursors to and crucial ingredients of our downstream responsibility ascriptions.

Cuddy and Fiske (2009 & 2015, cited in Westra (2019), 2825-2826)) recognize four combinations of warmth and competence, relating to four general categories we can sort people into in a rough and ready way (although we ought to keep in mind that these are scalar, potentially fine-grained judgements):

1. High warmth/high competence - these are "social reference groups:" those we seek to emulate and consider our peers.

2. High warmth/low competence - these are agents for whom we have paternalistic stereotypes: those we seek to protect, but do not consider as peers.

3. Low warmth/high competence - these are agents for whom we have "envious stereotypes" (in other words, we see them as high status, but view them as a threat).

4. Low warmth/low competence - these are agents for whom we have "contemptuous stereotypes" (in other words, we view them as low status and also as unthreatening).

These kinds of immediate reactions involve stereotypes, and, as we can already see, also involve character-trait attributions. If I meet someone and judge them to be of low warmth and high competence, I am primed to fit them into a category that gives me an idea of the kind of person they are likely to be and the kinds of attributions that are most fitting for them.[14]

[14] Indeed, the "Social Comparison" and "Social Identity" psychological literatures back up this claim. See Buunk and Mussweiler (2001), Gerber, Wheeler, and Suls (2018), Suls and Wheeler (2000), and Suls, Martin, and Wheeler (2002) for good reviews of these fields and Fiske et al. (2002) for a good review of the literature on judgements of warmth and competence as they relate to stereotypes and social membership. A key set of ideas that emerges is that we sort people in relation to both our *selves* and various social in and out-groups very quickly, and that these judgements are motivated by affect, desire, threats, self-esteem, and our perceptual, valuational, and conceptual schemas of the social world.

As Westra argues, all of these more or less immediate judgments influence what is called “hierarchical predictive coding.” Hierarchical predictive coding makes use of high-level action-predictions about agents which inform the contents of our perceptual, proprioceptive, and introspective representations. We attribute overall goals (she wants to go into the kitchen), and this allows us to predict various sub-goals and individuated intentions (she’s going to grab the door handle), predictions which can be updated to correct our model of the agent in order to get us more and more reliable predictions about them (2831-2832).

The groupings also, if only implicitly, begin to give us a sense of an agent’s *character*. As Westra (2018a) writes, “the more quickly we start to construct a model of a person’s character, the faster we will be able to use that information to predict and interpret their behavior,” a process which begins within milliseconds of encountering another agent, often using seemingly irrelevant and unproductive biases and cues (facial structure, for instance) (1232). So, although, “initial trait attributions based on faces are neither accurate nor particularly informative for predictive purposes,” they do begin the hierarchical predictive coding process that allows for us to make relevant predictions about intentionality downstream (*ibid.*).

Why should a theory bother with character trait attributions at all? Because, as it turns out, human beings are “character sensitive.” In our ordinary attributions of responsibility they emerge as a key inferential link to to mental states, and as a general practice-based posit. Of course, this doesn’t mean that our empirical theories need to treat character traits as a theoretical posit – it’s an open question whether something like character exists. But, as Doris (2002) notes in what is essentially an error theory of character as a metaphysical reality, people certainly talk and act as if those traits exist.

This is at least partially the case because inferring character traits help us solve the intractable problem of interpreting other minds. To use Doris’ definition, character traits are temporally *stable* and cross-situationally *consistent*. This stability and consistency is precisely the kind of information that can help us "code" people and make useful predictions about their behavior. Once we posit character traits, we can use them to make inferences about much less stable mental states like beliefs and desires. As Westra (2019) notes, informational processing here ends up running in the opposite direction one might expect it to. That is:

> If representations of character traits sit towards the top of the action-prediction hierarchy, and have significant downstream effects upon other forms of mental-state attribution, then it would make sense for this information to be prioritized, and processed as rapidly and efficiently as possible. Ironically, this means some of the most rapid inferences that we make about people are about what we take to be their deepest, most stable traits. (2833)

Let me summarize: social psychology tells us that we make judgements of warmth and competence and downstream character trait attributions milliseconds after encountering an agent. These attributions help us sort, code, and group agents. They are also, crucially, themselves influenced by the groupings we initially perceive. These sets of attributions are, in other words, often partially constituted by (and constitutive of) stereotypes – either of highly general warmth and competence categories or of specific kinds of bias. Furthermore, the attributions inform and constrain our action predictions and influence the belief-desire-intention attributions that we make about agents in particular situations.

Not only that, one of the bedrock findings of the social comparison literature is that these judgements prime us to want to "assimilate" towards positively valanced agents and "contrast" ourselves with negatively valanced agents.[15] Given that it's very hard to know what other people's mental states are (or, more pessimistically, it is impossible to know such a thing), here's one kind of solution: make some attributions about character traits, and use those as proxies for mental states. As psychologist Janet Nadler (2012) puts it: "so that a person with a bad character is blamed as if

[15] See Collins (2000) and Smith (2000) in Suls and Wheeler (2000).

he were reckless, whereas a person with a good character is blamed as if he were not reckless” (5).

This fits nicely with Alicke’s blame-validation model. As a reminder, his model posits that we have an immediate negative reaction to an actor who caused harm, which leads to a fast, automatic initial blame judgment. This initial blame judgment then guides subsequent perceptions about the actor’s causal role in producing the harm (13). So, Nadler (2012) writes:

> The blaming process is infused with motivation and emotion, and not dictated solely by individual acts and their consequences. Humans are social beings, and blame is a social process. When we observe a harmful outcome, our first reactions are emotional, and those emotions are informed by our immediate assessment of what kind of person could have caused this harm. On this account, a person with a bad moral character who causes a harmful outcome is a person who disrespects the community’s way of life As observers and community members, we react to such disrespect with moral outrage, and we experience the urge to blame and punish. Conversely, we are more willing to exculpate, at least partially, an otherwise virtuous person who causes harm, because his prior good deeds have in some sense licensed the transgression. (36)

Finally, as the Ping-Pong Model made clear, these effects are looping and multiply interactive. They are interwoven and everywhere massively influenced by a variety of factors. The process of coming to attribute responsibility then, relies on an interwoven series of conscious judgments about causation, intention, belief, desire, character and foresight, as well as being non-consciously influenced by predictive models, affect, emotion, stereotype and bias. All of this adds up to a rich stew from which we pull an interpretation of an agent in a particular situation and react to them in a way that comports with our overall folk-psychological picture of their mental states and physical actions.

The reader will surely have been reminded, during this reconstruction of mindreading and predictive coding, of Daniel Dennett’s famous “intentional stance” folk psychological theory. On his view, our theory of mind is an inferential Theory Theory, one on which we attribute states like beliefs and desires based on our assumption that other people (and we ourselves) are "intentional systems." An intentional system is the kind of thing with mental states that have *aboutness* – semantic contents, and (obviously) intentions. More specifically, it's the kind of thing that behaves rationally – in some normative sense of rationality. Given a set of beliefs, desires, and an evolutionary story plus an environmental niche, it's the kind of thing whose intentions and actions we can *predict* with some general (if not-terribly-specific) accuracy.

Crucially, on this view, whether the agents in questions truly have the beliefs and desires attributed to them is not of much consequence. Indeed, it's not entirely clear what the idea of them really having the beliefs and desires amounts to. Dennett (1987) describes his position as a middle-ground between naive realism and naive relativism. Beliefs (for example) are theoretical posits. They are *real* – real in the sense that other kinds of posits or frameworks are real (centers of gravity, the equator, and so on). Real enough! But, they are not like tables and chairs. We cannot (and Dennett thinks we will never) reduce them to purely physical properties – there will be no neuroscientific reduction of belief just as there will be no physicalistic reduction of centers of gravity. These theoretical posits extend to higher-level attributions of character, personality, and even self-hood.

The upshot of the intentional stance, for our purposes, is the irreducibility of interpretation (to modify a famous phrase), and the modest pluralism of acceptable interpretations this opens up for us. As Dennett puts it, "Not just brute facts, then, but an element of interpretation... must be recognized in any use of the intentional vocabulary" (342). If interpretation is unavoidable, we must ask the key question: how do we interpret? Historically, as Dennett points out, there have been two broad stances in analytic philosophy of mind: (a) some kind of "Normative Principle:" we attribute to a creature the propositional attitudes we think it "ought" to have. Or, (b) a "Projective Principle:" we attribute to a creature the propositional attitudes, "one suppose[s] one would have oneself in those circumstances" (342-343). Of course, divorced from theory and thrust into the real world, it seems obvious that we do some blend of the two when confronted with the task of interpreting one another's mental states. Dennett agrees, saying that the difference between the two principles often boils down to, "at most a matter of emphasis" (344). In either case, Dennett argues, it's fine that there are somewhat conflicting interpretations insofar as these interpretive norms both make clear that there's no "real" propositional attitude there, only better and worse pragmatic cases for interpreting one way or another.

If my arguments in this dissertation convince the reader of nothing else, it will at least be, I hope, that responsibility attribution is an interpretive practice. Interpretation is deeply contextual, and part of that context includes the interpreter themselves. To use a Gadamerian commonplace, there is no such thing as meaning devoid of context – all interpretation takes place from a specific point of view. Before we dive headlong into the waters of post-modern meaninglessness, let me remind the reader that the point is not that there is some irreducibly relativistic problem at the heart of responsibility attribution. Whether or not we are fully Dennetian "quasi-realists" about responsibility properties, hardcore realists about responsibility, or responsibility skeptics, the actual *practice* of responsibility relies on us interpreting one another's actions, their meanings, and what those actions and meanings reflect about one another's wills and character.

We can say at least this much: when it comes to responsibility attribution, the informational gaps and necessity of meaning-making that we are confronted with in our daily lives commit us to a practice of rough and ready informational overreach. In order to arrive at reliable judgements of responsibility, we rely on warmth and competence, inferential predictive coding, the use of character trait ascriptions and stereotypes, and whatever else we can get our hands on. The key question I've been asking then, is just *how* reliable these judgements are likely to be, where they might often go wrong, and how we can might try to make them better.

2.5 Putting it Together: The Limits of Evil

To understand the interplay of narrative structure, emotion, and the interpretation of agential character and capacities, let's turn to a well-worn example in the responsibility literature, that of Robert Harris. Harris' case is presented in several long narrative chunks by Gary Watson (2004) in his article "Responsibility and the Limits of Evil." There, Watson is concerned with examining the Strawsonian reactive attitudes and their ability to deal with formative history as a form of excuse or exemption. Strawson famously discussed taking the "objective attitude" towards those who were outside the bounds of moral discourse – including agents exempted from responsibility by, as Watson puts it: "being a sociopath," and being "unfortunate in formative circumstances." Watson's account is meant to raise deep questions for moral responsibility about this type of exemption. What I want us to focus on, are some points that float around the edges of Watson's account.

Watson is very aware that the way in which we hear the Harris story is going to prime differing reactive attitudes in us – attitudes that lead to what he describes as a feeling of ambivalence. I'm going to challenge this reading of the case, by focusing on what the resources in this chapter can help us understand about our reactions to the case. To do so, I'll try to summarize the Harris example, although anyone who has not recently read it would do well to review it in full – the particularly powerful and changing affective reactions one has while reading the narrative are precisely what I mean to discuss. Watson gives us four vignettes of varying length, all quoted from Miles Corwin's, "Icy Killer's Life Steeped in Violence."[16]

Watson's first vignette involves describing the reactions to Harris of fellow prisoners in San Quentin's death row, as well as the law-enforcement officials who put him there. So, we learn that even other hardened killers find Harris reprehensible and unpleasant to be around, and that county and state attorneys who tried him say: "If

[16] Los Angeles Times, May 16, 1982. Copyright, 1982, Los Angeles Times.

a person like Harris can't be executed under California law and federal procedure, then we should be honest and say we're incapable of handling capital punishment," and, "if this isn't the kind of defendant that justifies the death penalty, is there ever going to be one?" (235)

Notice the way in which this mode of presentation interacts with our model of responsibility attribution. As is so often the case in "naturalistic" responsibility settings, we begin *in media res.* Before we are given any "facts" about the case, direct perception of or interaction with an agent, or information about the responsibility relevant actions in question, we are presented with affectively charged information about *other peoples'* agent assessments. So, we begin in the second tier of deliberative assessment, which also loops back and begins a round of spontaneous evaluation. Of course, this spontaneous evaluation is not purely neutral - it is colored by what we are hearing from the people Watson quotes: our model of Harris as an agent begins with incredibly low warmth. It also includes the context of him as a death-row inmate hated by other death-row inmates: we are primed to have strongly negative spontaneous evaluations of Harris and his character.

Watson's second vignette quotes in grisly detail a description of Harris' crimes. A twenty-five year old Harris is hotwiring a car with his eighteen year old brother Daniel, planning to rob a bank with the stolen vehicle. They see two sixteen year-olds (John Mayeski and Michael Baker) eating a fast food meal, and, having trouble starting the stolen car, Harris decides they will steal Mayeski and Baker's car, which they do at gunpoint.

Harris forces the teens to drive to a canyon and, after telling them about the planned robbery and that they would be safe, shoots Mayeski in the back and chases Baker down a hill, shooting him four times. Corwin writes that: 'Mayeski was still alive when Harris climbed back up the hill, Daniel said. Harris walked over to the boy, knelt down, put the Luger to his head and fired." (236) To drive home the cruelty and apparent total lack of empathy Harris has, his brother Daniel is quoted: "'God, everything started to spin,' Daniel said. 'It was like slow motion. I saw the gun, and then his head exploded like a balloon, . . . I just started running and running.. . . But I heard Robert and turned around. He was swinging the rifle and pistol in the air and laughing. God, that laugh made blood and bone freeze in me"'(ibid). We then get the gut-wrenching detail that an apparently unaffected Harris takes the teens' fast food lunches and happily begins to eat a hamburger only fifteen minutes later.

Again, Watson quotes Corwin:

> Harris was in an almost lighthearted mood. He smiled and told Daniel that it would be amusing if the two of them were to pose as police officers and inform the parents that their sons were killed. Then, for the first time, he turned serious. He thought that somebody might have heard the shots and that police could be searching for the bodies. He told Daniel that they should begin cruising the street near the bodies, and possibly kill some police in the area. [Later, as they prepared to rob the bank] Harris pulled out the Luger, noticed blood stains and remnants of flesh on the barrel as a result of the point-blank shot, and said, "I really blew that guy's brains out." And then, again, he started laughing (237).

Watson concludes the second vignette with a few sparse details from Harris' past: that he'd spent much of the decade prior to the murder in prison, that he "was arrested twice for torturing animals and was convicted of manslaughter for beating a neighbor to death after a dispute" (ibid), and that in prison, "He was an obnoxious presence in the yard and in his cell... He acted like a man who did not care about anything. His cell was filthy... and clothes, trash, tobacco and magazines were scattered on the floor. He wore the same clothes every day and had little interest in showers" (238).

As Watson concludes after the second vignette, "On the face of it, Harris is an 'archetypal candidate' for blame," although it is also obvious that if responsibility is meant to be dialogic, Harris is an inappropriate object for "invitations to dialogue." (238). In some sense, then, this is a very helpful explanation of a typical epistemic process I am endeavoring to make clear and tangible. The way in which Harris' story primes our emotions and affective responses moves us very quickly to blame: we are prone to follow Alicke's path of "blame validation." Yet, as Watson rightly points out, this seems to bypass crucial questions about Harris' capacities. Is he actually an appropriate object of blame?

The narrative we are told about Harris has powerful framing effects that exercise our emotional capacities, and present us with an initial model of Harris' character and agential profile. It's also the case that our responses are affected by the severity of the outcomes of Harris' actions. As I argued above, this means we are more likely to find him causally responsible and that our model of his character (and reaction to him as extremely low-warmth) will be one that is less sympathetic to potential excuses. For example, Watson gives us a hint that his past was not an easy one: we know that he spent much of his youth in juvenile detention. Yet, combined with other descriptions of his torturing of animals, his laughter, and the unaffected way he ate the teens' hamburger, this fact serves only as a *confirmation* of his innate evilness, rather than a potential explanation of its sources.

Watson's third vignette attempts to re-frame all of this, and leave us in, as he argues, a state of ambivalence and confusion. To do so, Watson dives into the details of Harris' childhood. As I explored in Chapter One, agential history is one avenue where philosophers of responsibility have made strides in acknowledging the centrality of the epistemology of responsibility to our practices, and this serves as a compelling example. It also, however, serves as an example of the kind of lacuna I discussed: Watson's conclusion is that something is amiss and confounding when we place the facts of Harris' history alongside with the facts of his monstrous crimes.

The third vignette gives us Harris sister Barbara, who puts, "her palms over her eyes and [says] softly, 'I saw every grain of sweetness, pity and goodness in him destroyed.... It was a long and ugly journey before he reached that point.' (239). Corwin writes that, "Robert Harris' 29 years ... have been dominated by incessant cruelty and profound suffering that he has both experienced and provoked.

Violence presaged his birth, and a violent act is expected to end his life" (ibid). His alcoholic father was physically and emotionally abusive (and sexually abusive to Harris' sisters). His mother was also an alcoholic and was also extremely emotionally abusive towards Robert (although, mirroring his own case, we can *almost* understand why, given the abuse she herself suffered from her husband. Corwin writes that although: "all of the children had monstrous childhoods... even in the Harris family ... the abuse Robert was subjected to was unusual. The pain and permanent injury Robert's mother suffered as a result of the birth, ... and the constant abuse she was subjected to by her husband, turned her against her son. Money was tight, she was overworked and he was her fifth child in just a few years. She began to blame all of her problems on Robert, and she grew to hate the child" (240).

Unsurprisingly, things were no better for Harris outside the home. He:

> Had a learning disability and a speech problem, but there was no money for therapy. When he was at school he felt stupid and classmates teased him, his sister said, and when he was at home he was abused. "He was the most beautiful of all my mother's children; he was an angel," [Barbara] said. "He would just break your heart. He wanted love so bad he would beg for any kind of physical contact. He'd come up to my mother and just try to rub his little hands on her leg or her arm. He just never got touched at all. She'd just push him away or kick him. One time she bloodied his nose when he was trying to get close to her." ...The sad thing is he was the most sensitive of all of us. When he was 10 and we all saw 'Bambi,' he cried and cried when Bambi's mother was shot. Everything was pretty to him as a child; he loved animals. But all that changed; it all changed so much." Robert was too young, and the abuse lasted too long, she said, for him ever to have had a chance to recover. (240-241)

At fourteen, Harris did his first stint in a juvenile detention facility, where, more than once, he was raped and tried to commit suicide. By the time he finally got out at nineteen he began to kill animals, because, as Barbara put it: "The only way he could vent his feelings was to break or kill something... He took out all the frustrations of his life on animals. He had no feeling for life, no sense of remorse. He reached the point where there wasn't that much left of him." (241). We now have a very different sense of Harris than we did after vignettes one and two. What has changed? Watson claims that the story, "in no way undermines the judgments that he is brutal, vicious, heartless, mean. Rather, it provides a kind of explanation for his being so" (242). Accordingly, he wonders why it is that the reactive attitudes ought to be sensitive to this kind of explanation at all.

His analysis is that, "in light of the whole story," we have conflicting responses of sympathy and antipathy – both appropriate, but in conflict in such a way that we are unable to respond to the overall picture of Harris coherently. As Watson puts it, "the ambivalence results from the fact that an overall view simultaneously demands and precludes regarding him as a victim" (244). Finally, Watson ties in the issues of the historical dimension of responsibility and moral luck, writing that, one way to make sense of Harris as responsible is to say that he somehow consented to or "took responsibility" for what he became. Thus, we can make sense of Harris' monstrous actions disclosing something about his self and quality of will insofar as he identifies with his current actions. But, Watson notes, this line of thinking is, "rooted in a picture according to which the fact that Harris became that way proves that he consented," when he had an uncoerced opportunity to do so (250). But, this seems like an unlikely possibility given what we know of his past – given his history of abuse, there doesn't seem to be a likely moment when Harris could have made a decision to be evil that was free of the influences of his terrible upbringing.

I want to resist some of Watson's conclusions here, and reorient us towards the epistemic dimensions that are given short shrift. The model I've built up in this chapter can help us make sense of things. Is it really the case, first of all, that we are met with a kind of ambivalence about Harris' story? Of course, reader reactions will vary – and there is no "right" way to feel about Harris on my view. But, I'd emphasize that our model of responsibility attribution doesn't need to settle on "ambivalence" between conflicting viewpoints of Harris as a final output. Instead, each vignette primes another round of affective response, and another round of deliberation. We are given new evidence about Harris that is relevant to our assessments of the meaning of his actions and the makeup of his character. And the powerful and tragic stories of his past abuse give us access to new empathetic responses and new reactions of warmth. Even if we do end up ambivalent, the model I've presented paints a much clearer picture of *why* this is the case. We can say much more than, "this case is complicated, and it's hard to track Harris' responsibility relevant capacities."

In particular, we can see why narrative framing is so crucial in this kind of case both because of the way it heightens emotionally salient aspects of agents lives, actions and contexts *and* for the ways in which it provides relevant contextual detail. As Corwin writes, "[Another sister said that] if she did not know her brother's past so intimately, she would support his execution without hesitation" (241). The narrative evidence we have access to massively influences our eventual responsibility attributions; not just because of the level of detail it provides us (although this is obviously crucial), but also because of the way it relates us to other agents in a social context. The fact that our reactive attitudes are blunted by learning of Harris' past is not a result of ambivalence but of our model delivering a new output: we understand Harris in a new way, one which makes him a tragic (and perhaps even sympathetic) figure. Watson is right that we can continue to hold that Harris' *actions* are abhorrent, but it doesn't seem right that we are precluded from viewing him as a victim. In fact, at least in my own case, I seem to take something very like the Strawsonian objective attitude towards Harris after reading the third vignette. The fact that he killed the teenagers does not make him *more* of a candidate for responsibility and blame but *less* of one. It shows how far he has become removed from the realm of responsibility and morality, such that one can muster only horror towards and pity for him, but not hot-blooded blame.

Watson gives us a final "postscript" to the story as a fourth vignette. He describes the gruesome execution of Harris in 1992 - how after mouthing, "I'm sorry" to a victim's father, he slowly choked to death in a room full of cyanide gas for almost ten minutes. Furthermore, we are confronted with the fact that present at the execution was a *friend* of Harris' who viewed the scene as "indescribably ugly... nakedly barbaric" (259). And so, as Watson concludes: "One thing is clear from this report that was not obvious at the time of the killings: in his last years Harris either remained, or became once again, capable of friendship and remorse. His crimes were monstrous, but he was not a monster. He was one of us." (259). Again, one of Watson's points is about the difficulties of taking the objective attitude towards someone like Harris, who (if the postscript is to be believed) is, after all, just a human being – one who encountered horrendous formative circumstances. So we are left with unsettling questions. Was Harris a capacitous member of the moral community, an apt target for responsibility all along? Would treating him as capable of moral dialogue have made some difference to his own ability to recognize himself as part of that community? Did he deserve to die (or at least, be punished)? My claim is that these difficult questions of psychopathy, rationality, and punishment point towards precisely the importance of attending to the epistemic dimensions of responsibility.

How we will answer them depends crucially on our access to evidence about Harris that goes beyond a (never available) reading off of his capacities. We need to know what caused his capacities (or lack of them) to be the way they are. And, these judgements are fundamentally interpretive. We cannot know the full causal story of Harris' capacities – nor does there seem likely to be some exact, empirical specification of their make-up. Not every victim of appalling abuse becomes a cold-blooded killer. Not every cold-blooded killer repents and comes back to the moral community. We are doing a lot of interpretive guesswork here - and what details we have matter for that interpretation. Not only this, the Harris case makes vividly clear that the order, tone, and authorial makeup of the details matters almost (or just) as much as the information itself. Without the direct testimony of Harris' family members, without the presentation effects of Watson's vignette ordering, and without the affective reactions we have to Harris' abuse, our reading of his case would be very different.

Who we talk to, when, and what they say matters for our judgements of responsibility. That this has been presented in traditional accounts of responsibility at most as a static claim about informational access to evidence is unfortunate. For instance, it's obvious that our social ties to various interlocutors have a huge power to color the way we view responsibility relevant acts and agents. This is true both at the level of the details we will be privy to and the way in which we will tell stories differently. So, it's no surprise that it is much easier to hold as non-responsible groups of agents we simply aren't talking to, or to find guilty and blameworthy those who we only hear negative things about. In other words, interpretive over-reach is rampant in cases like Harris'. This is clear in our ability to form and attribute responsibility judgements after only the first vignette – in the way in which we over-estimated Harris causal control and made far-reaching judgements about his character and quality of will based on limited evidence. So too, it's evident in our shifting reactions after the third vignette – in the ways in which narrative can prime our affective responses and play with our emotions, and perhaps, in our *under*-estimation of Harris' capacities and causal control.

2.6 Common Errors Refined

Let me try to summarize and bind together the various threads of this chapter. I began by arguing that I'd be able to show a few of the most common, pernicious, and important kinds of errors involved in our responsibility practices as they are currently constituted. Let's begin where we just left off, with interpretive overreach:

Interpretive Overreach

I've begun to show that our judgements and beliefs about responsibility often outstrip our evidence or justification in predictable ways. Any normative conclusions about this fact must await further argumentation. That is, the question of what makes interpretive overreach *overreach* turns out to be incredibly nuanced and complicated. In some Dennettian sense, we are *always* overreaching. An interpretation, after all, even if beholden to facts, goes beyond them – that's its point.

For now, let me draw the distinction between the kind of responsibility judgement we are justified in holding and one which clearly over-reaches in the following way: I pointed out that a plausible model of responsibility attribution involves both a non-deliberative, affective track of reasoning and a more deliberative, conscious track of reasoning. One outcome of this is that our deliberative track is often disrupted by affect, motivated reasoning, and biased informational search. Consider a common kind of case involving the fundamental attribution error. I am cut off in traffic; the emotional responses that cascade through me make it natural for me to interpret the offending driver as *intentionally* offensive. When, on the other hand, *I* do the cutting off, I have all the excuses in the world through which to interpret my behavior as benign and unintentional. The other driver is fully responsible (and blameworthy), while I am excused.

Clearly, I am not reacting, in these cases, to the rights kinds of (or all the right kinds of) evidence. My interpretations overreach the conclusions I am justified in drawing. This is a normative claim, and, as I say, it must be fully explored and argued for at a later date. For now it is enough to say that we have very good psychological evidence that various kinds of interpretive overreach are predictably pervasive in our responsibility attributions, particularly where highly emotionally charged or socially salient events and agents are involved.

Let me summarize each particular kind of interpretive overreach I identified, and be explicit about why they ought to matter to responsibility theorists:

Over and Under Estimation of Causal Control

The evidence from the psychological literature surveyed above was very clear: our estimations of causal contribution and control are quite often influenced by irrelevant down and up-stream information. Whether we like someone, what our current mood is, how bad the consequences are, what someone's particular social role is: all of these factors and more influence our estimations of causal control.

This matters for responsibility theorists who argue that control is an important capacitarian component of responsibility (so, all of them). If the epistemology of moral responsibility is full of errors on this count, than we cannot blithely say that whether agents have certain causal powers, the ability to do otherwise, fair opportunity to avoid wrongdoing, sufficient volitional control, and so on is just something we check (even if indirectly) off our list of responsibility conditions. We are likely to get things wrong about ourselves and other agents - to over or under-state their causal powers. And this kind of error directly affects our estimations of their responsibility. Unless we are confident that there is some empirically verifiable answer to how much someone has causally contributed, and what, in general, their causal powers are, then our judgments about causality are going to be interpretive; and they are going to be colored by emotion, affect, bias, and situational context.

Character Judgement

So too with judgements of character and quality of will. As our examples of drunk-driver blame and John's mishap on the way to his parents' house make clear, our judgments of intention, quality of will, and character are as amenable to informational hijacking as anything else. Again, the psychological literature makes clear that stereotypes and judgements of warmth and competency are hugely influential in our perception of character traits. Most importantly, character, like blame, is "sticky." Once we have some stable character-based models of agents, although they will be updated with new information, it is unlikely that they will be entirely revised. It is easier to model downhill than uphill, in other words: if I think that someone is a liar and a jerk, it is hard for me to overcome this assessment even when presented with evidence to the contrary. Perhaps, for example, the lying jerk is just being nice as a way to set me up for a future con, or perhaps he is only pretending to be nice. This kind of reasoning is common.

Narrative Framing Effects

As the final example of Robert Harris shows, all of these effects are mediated, enhanced, and diminished by their framing in narratives. Narrative presentation activates and ramps up the affective and emotional aspects of our moral psychology, as well as making salient certain streams of information and masking others. Different narratives lead us to different conclusions about agents and events, and therefore, different attributions of responsibility. Finally, it's important to note the social and dialogic nature of narratives, and the effect this has on our reasoning.

Indeed, what we are often responding to is not a direct narrative re-construction of a person's relationship to a particular event, but a *distributed*, diffuse narrative that emerges amongst various strands of popular discourse. By the time I form the judgement, "x has behaved poorly," I may have some inchoate sense of the previous judgements of a large range of interlocutors. The Ping-Pong Model shows how we can enter into the responsibility ascription pathway at various points – and this social dialogue is one such entryway. The opinions and judgements of others shape and modify my own, and are often foundational in regards to what I start out believing about an agent or event. Importantly, this can happen subconsciously, and can also "hijack" non-logical processes that filter my judgements. For example, perhaps I find out a story about something that occurred on the set of a movie that is about to come out, and the story I have heard sheds a negative light on the director. Even if I don't form an explicit judgement that the director is a bad person, or form an explicit vow to dislike the movie – it may be hard for me to feel favorably about it, even if I suspect I kind of like it upon watching. This kind of framing effect, I've argued, is heightened by narrative, particularly when such narratives interact with character, affect, stereotypes and other forms of interpretive overreach.

2.7 Conclusion

In this chapter I've argued that a descriptive epistemology of responsibility can make plain some common errors in our responsibility practices and attributions. To accomplish this, I argued for and presented a core model of responsibility attribution, drawing on contemporary empirical work in social and cognitive psychology. I then put the model in contact with some toy examples and ideas from the philosophy of mind to show how such a descriptive model has explanatory advantages for responsibility theorists. Once we are clear on the complicated nature of everyday ascription, it becomes more apparent why certain cases are difficult, where we are likely to go wrong, and what kinds of questions our epistemology raises for the metaphysics of responsibility. Finally, I began to precisify the most pernicious and common errors that I think our exploration has thus far uncovered. In the following chapter I continue this work, turning to a full discussion of "Status Sensitivity" errors.

Portions of Chapter 2 have appeared in publication as "Blame for Me and Not for Thee," in *Ethical Theory and Moral Practice* 25: 265-282, 2022. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 3

Status Sensitivity

It is within dynamics of power that persons come to be considered reliable agents, credible interlocutors, and deserving members of the moral and broader community. It is within dynamics of social power that the normative expectations we have of ourselves and of each other originate. It is within dynamics of social power that persons judge one another as apt participants in the interpersonal relationships that characterize responsible agency.

Marina Oshana [1]

The point here is that theorizing is not just descriptive but also reality constructing. In the process of describing the social forces producing and excusing dimness to certain kinds of wrongdoing, we are also "making up" persons.

Cheshire Calhoun [2]

[1] Oshana, 2018, 92

[2] Calhoun, 1989, 404

3.1 Introduction

So far I've argued that our everyday epistemology of responsibility is highly contextual and socially scaffolded. In chapter one, I argued that contemporary work in moral responsibility largely puts to the side the interpretive and socially mediated ways we come to judge one another responsible. In chapter two, I presented a model of these epistemic processes, concluding that many metaphysically extraneous (but epistemically central) factors influence who we see as responsible, how much we pay attention to responsibility, how eager we are to bestow praise, how likely we are to forgo blame, and so on. A key point in both chapters was that the epistemic load involved in a process that involves rich agential and contextual interpretation (rather than the direct tracking of properties) is very high.

In particular, agents are often overly epistemically sensitive to markers of "high" and "low" status, and too epistemically *insensitive* to the ways that differences in power affect their responsibility judgements. My claim is that these kinds of distortions can lead to reliably bad responsibility judgements. One conclusion of these lines of argument is that where large gaps in "social power" are present, recognitive, interlocutive, and interpretive errors are reliably likely to occur – what I've called "status sensitivity" errors.

In this chapter I'll fill out these claims about social power and epistemic distortion. And, I'll begin to ask how we might revise and improve our practices, given these errors. In particular, I'll ask what it would look like if we introduced a new set of norms into our responsibility practices - norms which ask us to blame those of low social status less and those of high social status more. The general plan is as follows: I'll first concretize the notion of social power and status I have in mind. Then, I'll give an example of the way in which status distorts the epistemic processes of responsibility attribution. I'll consider whether this epistemic distortion calls for an overall reduction in our tendency to blame one another. To assess this, I ask whether such a general reduction would solve the problems I canvass, and what new problems it might introduce into our practices. In the end, I'll argue that a general prohibition is unhelpful. Instead, I argue that: 1) there is an asymmetry in the acceptability of blame between the powerful and oppressed, 2) the withholding of blame by those with social power is often warranted, and needn't be seen as objectionably paternalistic or disrespectful, given that it is a form of self-critique and epistemic humility, and 3) a meta-norm of blaming those with social power who *don't* withhold blame is the key to successful practice revision.

3.2 Social Power as an Epistemic Distorter

In a recent paper, Michael McKenna (2017) analyzes the problems power-dynamics can cause for moral responsibility in the following way:

> There is something morally suspect about the social conditions facilitating exercises of [agents'] agency when they act in ways that are morally praiseworthy (and also morally blameworthy). This is because, as the conversational theory reveals, quality of will is to be identified and explained by a community of interpreters who take some kinds of actions as indicative of good will and other kinds as indicative of lack of good will. Since some in this community are socially empowered, in contrast with others who are socially disempowered, the conditions for what signals good and ill will are liable to arise from potentially unjust social circumstances. (40)

That is, on his conversational theory of responsibility, quality of will is a function of interpretation – but interpretations are messy, social things (as I've been arguing). They take place in particular contexts and are highly sensitive to social power dynamics. Indeed, this means that, before we even run into the dangers of mis-interpretation, the answer of what constitutes the right interpretation at all is intertwined with power and social identity.

My analysis of social power leans heavily on this interpretational point, and broadens it to include the causal story of judgment formation I dubbed the Ping-Pong Model. Responsibility judgements emerge not only from a recognitive community that forms interpretations through a process of explicit conversation, but also within individuals taking affective, heuristic cues from the social dynamics in which they are embedded. The key idea here is that there is a complicated interaction of *individual* psychological error and *structural* unfairness that we need to correctly diagnose.

One particular kind of epistemic distortion I canvassed in the previous chapter was "status sensitivity." We are often affectively biased, both in our initial reactions, later searches for evidence, and overall interpretation of events, by whether someone has a relatively high (or low) degree of social status, and whether their social position is favorable or dis-favorable to us in various ways. I use the phrase social status somewhat interchangeably with the phrase "social power." In both cases I mean to invoke an intuitive combination of our contextual status in social groups, our social abilities, and our material resources. In other words: our power to act, influence, and gain uptake in the social world. This power is partially constitutive of the social world itself, and is deeply relational.[3] Our positionality affects our agency, our options, our relationships, our status and our values – and all of this is mutable and contextual. In some contexts we may be of high social status and have a great deal of social power, and in other contexts we might be of low social status, or even count as oppressed.

No one denies that social power exists, or that we have typical epistemic processes which result in judgments about moral responsibility. What I'm noting is that the two interact in ways that produce outcomes largely absent from the moral responsibility literature.[4] In particular, attention to the epistemology of responsibility helps

[3] My thinking about social power draws, in large part, on the work of feminist philosophers such as Cudd (2006), Oshana (2018), and Young (1990), as well as recent work by Abizadeh (2021) and Menge (2020).

[4] One interesting historical precursor to my arguments here is the work of Adam Smith. He held

us make sense of cases where those with a great deal of social power seem to dodge responsibility generally, and dodge blame in particular.[5]

In this sense, social power is a particularly reliable and pernicious distorter of the epistemology of responsibility. Because the notion is at one and the same time so intuitive and so complex, let me say a bit more about my conception of social power. Feminist philosophers have done a great deal in the last fifty years to advance our understanding of power beyond traditional philosophical views of "power over" and "power to."[6] Authors like Iris Marion Young have argued that power should not be thought of as a simplistic, atomistic ability or disposition that individual agents have. When we think of certain causal powers, this way of speaking is surely appropriate, but in the moral and political realm, the types of powers we are discussing are often relational, socially mediated, and highly non-static.

Social power, then, encompasses both *material* and *reputational* differences. It is a combination of position in concrete social groups, control of resources, the ability to act as one wishes, and the relation one stands in to others, both as an individual and as a member of various groups. The following three features are operative in my understanding of social power: I. That it is an objective social phenomenon, II. that

that we have an unhealthy obsession with the wealthy and powerful, and an unfortunate inclination to look down on those in destitution. As he says: "Upon this disposition of mankind, to go along with all the passions of the rich and the powerful, is founded the distinction of ranks, and the order of society. Our obsequiousness to our superiors more frequently arises from our admiration for the advantages of their situation, than from any private expectations of benefit from their good-will... Their benefits can extend but to a few; but their fortunes interest almost ever body" Smith (1976, 52).

[5] Consider the infamous, recent case of Brock Turner. Turner, a member of the Stanford Swim team, raped an unconscious classmate, was arrested, and faced prosecution. He blamed his actions on, "a culture of drinking, peer pressure and 'sexual promiscuity."' His father lamented, "that his son's life had been ruined for '20 minutes of action,"' and the judge in the case handed down an extremely lenient sentence. (See: https://www.nytimes.com/2016/06/09/us/brock-turner-blamed-drinking-and-promiscuity-in-sexual-assault-at-stanford.html). Compare Turner's treatment to the way the socially "low-status" are magnets for responsibility and blame ascriptions. Note, also, that there is an interesting asymmetry here: it is not as if the high-status similarly "dodge" ascriptions of praise. However, this asymmetry between praise and blame is beyond the scope of this chapter.

[6] See Allen (2016) for a wonderful overview.

it is partially constitutive of individuals and the social world, and III. that power occurs at individual, relational and structural levels. To clarify:

I. As Ann Cudd (2006) has argued, power is “an objective social phenomenon” (23). That is, it’s reality is manifested in the social world. In discussing social power we move beyond a purely material notion of causal power, and beyond a purely dyadic notion of one individual’s power over another. Consider, for instance, a classic definition of Power from Max Weber: “the probability that one actor within a social relationship will be in a position to carry out his own will despite resistance. . . ” (1978, 53). By defining power as an objective social phenomenon, we also make clear that power is not merely potential or dispositional: it actually shapes what we are able to do, individually and collectively. This means that we can view power as constitutive and structural.

II. By constitutive, I mean that power is both something that partially constitutes individuals and their abilities *and* as that it is a set of relations that partially constitute the social world. The powers that agents have are partially constitutive of their agency itself. Not only this, but as Oshana (2018) reminds us, “It is within dynamics of social power that the persons judge one another as apt participants in the interpersonal relationships that characterize responsible agency” (92). That is, there is no practice of responsibility without the constitutive element of social power and the ways in which it shapes social dynamics. To act in ways that count as responsible, agents need powers manifested by and in the social world.

III. By structural I mean that power is not purely a matter of individual wills or, as I said, dyadic relations. I follow Young (1990) in rejecting a notion of power that views it, primarily, as a resource. Many views of power have been focused on the distribution of power as a driver of oppression and power “over” other individuals. Although a distributive understanding of power as a resource can be a crucial tool in identifying and fighting oppression, it does not fully analyze the relations involved. As Foucault would remind us, power is relational, and needn’t be an explicit exercise of individual “sovereignty.” Instead, we should look for it: “as the effect of often liberal and ‘humane’ practices of education, bureaucratic administration, production and distribution of consumer goods, medicine, and so on. The conscious actions of many individuals daily contribute to maintaining and reproducing oppression, but those people are usually simply doing their jobs or living their lives, and do not understand themselves as agents of oppression” (1990, 41-42).

In future discussions of *praise* I think it would also be important to note a fourth feature: a conception of power as productive – the ability to produce certain interpersonal and social outcomes. This kind of social power – the power to lift others up, for instance – is also important and often neglected. Combined with the other features it also illuminates that power doesn’t always have to occur, contra Weber, as carried out “despite resistance.” Instead, productive social power can also consist in *assisting* one another, as Abizadeh (2021) has recently pointed out (2).

Why does this notion of social power as structural, constitutive of the social world, and deeply relational matter for our purposes? One reason is that it is precisely in the messy context of sorting through group identities, social statuses and individual responsibility that gaps in social power are most in danger of clouding our epistemic situation. We need a good understanding of social power in order to recognize its interactions with group and individual identities, and with moral responsibility itself. Dynamics of social power present clear rupture points where problems in our practices can occur, both in our *interpretations* of one anothers’ responsibility and in our attributions of blame and praise.

Given all of the above, we should understand social power as highly contextual, mutable, and perspectival. Included are features of agents that matter in terms of in-group and out-group status markers, but which are free-floating and contextual in terms of material power. I mean such features as: race, religion, body morphology, skill at various activities, knowledge of particular roles, identities, and many more. These are things that mark us as high and low status in the eyes of others according to *their* views of whatever it is that confers status. Clearly, context is king here. For one person, it may be high status to be white, roll-coal, and wear "thin blue line" printed apparel. For another it may be high status to be brown, cook traditional Zacetecana food, and wear snap-front western-wear shirts. For both, it may be low-status to be Orthodox Jewish. This example is a good place to pause and note that "low-status" here *needn't* imply a racist or otherwise prejudicial view of others. It simply concerns who one is prepared to take most seriously in one's social interactions. Equally obviously, such judgements *can* emerge from (or constitute) outright bigotry.

These kinds of social power-dynamics are very real, and they affect our life-chances, our social roles, and the way others treat us. But there is nothing necessarily fixed about them. I say necessarily because, of course, some dynamics are very entrenched to the point of near-fixity - but it is helpful to remind ourselves that they *needn't* be. Still, power is partially constituted by factors which *are* materially in-arguable: for instance, the wieldable power (in the Weberian sense) that comes with wealth, connections, abilities and so on. Some agents have access to doomsday bunkers to which they can flee during an international pandemic. Others are poised to become the world's first trillionaires. Still others must go to work in a grocery store without personal protective equipment. These are not gaps in power based in any direct way on perception or recognitive respect – they are almost purely material.

With these clarifications of my understanding of social power out of the way, one might press me: why does any of this matter for the epistemology of moral responsibility? That is, why worry about fixing unjust power relations in the context of the epistemology of moral responsibility? *Surely the thing that ought to be done about unjust relations of social power is just whatever the right ameliorative theory of injustice says we ought to do.* Yes, but insofar as those relations are partially constituted by our practices of attributing and holding responsible, part of that work will involve looking at the practices of responsibility themselves and asking, "can we do better?"

Thus far, I've argued that our responsibility practices which operate across gaps in social power will be reliably distorted in consequential ways. The dynamics of social power matter at an individual psychological level - in how we interpret one another and form judgements about one anothers' responsibility - and at a structural level - in how our practices are constituted, and in what the rules for interpretation and norms of blame and praise are. Below, then, I introduce an example where there are *obvious* gaps in power that produce *obvious* epistemic misfirings. With a more concrete understanding of the kinds of distortion caused by social power in mind, and noting that almost all of our responsibility practices involve such gaps in power, we can then tackle the structural question: if these gaps and distortions are endemic to our practices as a whole, what should we do about it?

3.3 An example and a clarification

Let us focus on a particular case in which power and social status straightforwardly distort responsibility judgments. It involves *both* undue deference to those with more social status, and a lack of care towards those with less. On February 23rd, 2020, at around 1pm, Ahmaud Arbery, 25, was shot to death.[7] He was black, and he was killed by two white men who followed him in a pickup truck as he jogged through their neighborhood. Their justification for his killing was that he looked like a robbery suspect who had burglarized several properties in the area. To most outside observers, it is obvious that their judgement of Arbery's responsibility, liability to be punished, and the accompanying extrajudicial punishment they meted out, were horrifically wrong (and almost certainly racially motivated).[8]

[7] See this article for a timeline of events: `https://www.nytimes.com/article/ahmaud-arbery-timeline.html`

[8] Indeed, the justice system also eventually agreed with this reading of the case: https://www.justice.gov/opa/pr/federal-judge-sentences-three-men-convicted-racially-motivated-

However, it is what occurred in the aftermath of the killing that I think provides an even *clearer* example of sensitivity to status and social power affecting responsibility judgements. First, as it turns out, the District Attorney originally in charge of the case knew one of the killers, and instructed police officers to refrain from arresting them. In the end, it took two months for the DA's office to bring charges against the killers. Second, right-wing media dug up a video of Arbery walking through a construction site in the neighborhood where he was killed, and online commenters used this "invasion of property" as a justification for his killing. Both of these after-the-fact developments show how sensitivity to various social power dynamics shape individual and group-level judgements about responsibility.

First of all, why was it that Arbery's killers were not detained until more than two months after the incident? Why, during that time, did three separate prosecutors recuse themselves from the case? As Fabiola Cineas writes in Vox:

> The first two prosecutors in the case recused themselves due to professional ties to Gregory McMichael; the third suggested that the case move to another office, citing resource constraints. But before George E. Barnhill, the second prosecutor, recused himself (his son works at the district attorney's office that Gregory McMichael retired from), he wrote a letter in an effort to exonerate the McMichaels, claiming that Arbery "initiated the fight," which the graphic cellphone video footage later suggested was not the case. He concluded that "Arbery's mental health records & prior convictions help explain his apparent aggressive nature and his possible thought pattern to attack an armed man." [9]

If we start at the most one-to-one, interpersonal level, then, we have someone doing an acquaintance a favor. The first two prosecutors eventually recused themselves, citing their relationships with McMichael as reasons they couldn't charge him. Of course, these are not actually good reasons for why he couldn't be charged or arrested, whether or not those prosecutors could themselves have seen the case through

hate-crimes-connection-killing

[9] https://www.vox.com/identities/2020/5/18/21262100/ahmaud-arbery-breonna-taylor-arrest-police-investigation

to completion. Broadening out a bit, we have someone showing solidarity with a law-enforcement colleague. Barnhill's letter made clear that his professional sympathies lay with McMichael. Perhaps broadening out further, we have someone giving epistemic deference to a member of their own race. It is impossible to delve into the psychology of the decision not to charge or arrest Arbery's killers, but the letter certainly reeks of bias.

In all three cases, whatever the exact psychology of Barnhill and the other prosecutors, it seems clear that Arbery was being treated as both a competent, capacitous adult, guilty of making poor choices, *and* a dangerous, aggressive semi-agent, to be treated with something like Strawson's objective stance. His low-status allows for several, perhaps conflicting, kinds of intrepretational schema to activate concerning his responsibility.

On the other hand, McMichael is seen as less than fully responsible for Arbery's killing, as in possession of plausible excuses, or, at the very least, as a former law enforcement officer in line for deferential treatment. Again, this contextually "high" status proximity to power allows a different kind of interpretive schema for assessing the perpetrators' responsibility.

What are the elements of status and power at play here? I've mentioned obvious in and out-group dynamics ranging from personal relationships to racial solidarity. In fact, in some sense, this is all I need to show to get the argument off the ground. Any time there is a relation of an "us" and a "them," the Ping-Pong model will be primed to exploit emotions, biases, heuristics and stereotypes in ways that will lead to non-ideal judgements. In this particular case, there is also, of course, the structure of state power, as well as Arbery and McMichael's relationships to that structure. McMichael was very much a beneficiary of state power – both in presumption of innocence, legal protection and literal employment. We know the ways in which these material and status based nexuses of power matter for people's life chances, outcomes within the legal system and so on. What I'm arguing is that they also matter in terms of how people are *seen* by various agents, and that this affects our judgments about responsibility. It was hard, for whatever reason, for the first three prosecutors to see McMichael as blameworthy – and the relationships of social power between Arbery, McMichael,and the prosecutors, I'm suggesting, may have been a large part of the reason.

As I've said, however, the online reaction to Arbery's death presents an even clearer case of epistemic distortion. Predictably, Arbery's killing was seen by many as yet another data point that the lives of black men do not matter in America. Equally predictably, to reactionary conservatives, the imposition of another #blacklivesmatter viral moment into an #alllivesmatter world called for a robust defense of either white benevolence and egalitarianism, or the claim of "reverse racism." Either way, some justification for Arbery's killing was needed for such commenters to advance the narrative that the murder and subsequent lack of accountability didn't show anything deeply pernicious about American society. For a few days on social media in mid-May, just such a justification was procured. Ahmaud had been seen walking through a nearby construction site: evidence, supposedly, of his criminality, dodginess, and aptness to be shot on sight. Whether we read this as an obvious case of (racially) motivated reasoning, or as a bad application of moral principles, or as a misunderstanding of what might justify lethal violence, the end result is the same: Arbery was painted as somehow responsible for his own murder, and the men who did it were painted as reasonable; either partially excused, or wholly justified.

These two sides of this same coin often present themselves where overt gaps in social power butt up against responsibility judgements. One side's responsibility is heightened, and the other's is diminished, so that the end result is that the "right" people get what they deserve – those who ought to be punished are, and those who ought to be forgiven are treated kindly.[10] One key thing that I've argued is that cases

[10] Indeed, many in the carceral abolition movement have pointed out that the basic move of a hierarchical, carceral state is to offer protection for "insiders" and punishment for "outsiders."

like Arbery's do not necessarily turn on online racists doing a poor job of tracking the responsibility-relevant capacities and properties in play. And this is so because that's not even something they are usually *trying* to do. In arguing that Arbery is to blame because he walked through a construction project, the commenters are not making an argument about Arbery's responsiveness to reasons, or even his quality of will. Their point is not that walking through a construction site shows that Arbery intended vicious harm or malice necessitating his killing. They are focused on the social structure of fear and power that paints "law-abiding" whites as those in imminent danger, and people of color as threats to the economic and social order. In other words, they already take for granted that there is some justification for the killing, and so they are working backwards to find a narrative that suits this conclusion. They are not looking for properties in the world, but a story that satisfies the right emotional arc.

3.3.1 Explicit Bad Beliefs

At this point, I've argued that the way we go about attributing responsibility to one another involves predictable and systematic distortions of status sensitivity. One plausible response to my line of argument and choice of example is that I have *overstated* the distorting effects of social status and power, and *understated* the effects of explicit "bad" beliefs: beliefs in sexist, racist, and classist propositions, for instance. In other words, a plausible objector might argue: "We can grant that the social world impacts our responsibility judgments and practices, but this is because of explicit beliefs and moral failings. Many of these cases are, for example, centrally explained by a failure to fully recognize one another as moral agents."

Why not think that it's the having of racist *beliefs*, for instance, that result in the pathologizing of Arbery's behavior? In other words, nothing is, technically, "distorted" - instead, our beliefs are operating as intended - it's just that they are morally *bad* beliefs. It's counterpart would be the claim that what is in need of *revision* are those beliefs as well. It's no use revising our moral responsibility practices - what we ought to revise are the beliefs of incorrigible sexists and racists.

This is a very plausible objection, but in the end I think it is unsustainable. The claim that most (or many) of the instances in which people mis-judge responsibility across gaps in power are due to explicit racism, sexism, classism, ageism, out-group animosity and so on, ought to be rejected. The flattening of heuristics, biases, and background ideology to explicit belief alone can't be the whole story about all cases of epistemic misfiring. To be clear, surely explicit beliefs do play a role some of the time. Indeed, this is exactly what the Ping-Pong Model predicts. Our responsibility judgements ping-pong back and forth between implicit and explicit, conscious and unconscious. So, it may well be that someone who sees the video of Arbery in a particular social context updates and reconfirms their conviction that black people are untrustworthy, and forms a judgement explicitly based on this racist belief.

Disentangling perceptions, conceptual schemas, social memory, explicit beliefs and subconscious processes is an almost impossible task in such cases. However, it is the claim that explicit beliefs are the main explanatory factor in most instances, or that this generally explains our difficulties with forming judgements of responsibility that I mean to reject. Consider, for instance the inconsistent application of responsibility relevant features like agential history, excuses, and quality of will to Arbery and the McMichaels. Is explicit racism the best explanation of this inconsistency?

In short: no. Those with racist beliefs are neither: a) explicitly ignoring relevant exonerating information when confronted with a case like Arbery's and explicitly searching for damning evidence in his past, or b) doing the best they can to work out the causal history of Arbery's actions and determine whether his responsibility relevant capacities at the crucial moment trace back to earlier decisions. Instead, the judgements are so wildly inconsistent just because of how *little* explicit thinking they are doing. They are primed to react positively to portrayals of a murdered black man as a thug, both given their social milieus and given the easing of pressure on their inconsistent beliefs about justice, race, and responsibility a quick and mostly unconscious reaction affords them. It would be very hard to consistently endorse and hold in mind a libertarian world view of retributive desert, the fact that (by their own lights) walking through a construction site is not a reason to be murdered, and a belief that white people are equally as oppressed as black people. It is much easier to not think much about these things at all - to react first, and reconstruct, and rationalize later and as necessary.[11]

There is an interesting and related debate about the idea that explicitly dehumanizing beliefs do much of the work in supporting systems of racism, misogyny, and so on. Paul Bloom (2016), for instance, argues that, far from dehumanizing victims, absusers often misuse their capacities for empathy to target them for very human reasons.[12] And Kate Manne (2018)'s recent work on misogyny rejects the claim that misogynists do not view women as fully human. She calls this the "humanist" view – that misogyny is a lack of empathy rooted in a perceptual and conceptual set of beliefs about and "seeings as" that paint women as less fully human than men. As Manne explains this kind of view, when we recognize humanity in another, this is supposed to motivate us to be kind (or at least not to be cruel) (141-142). Why? The idea is, roughly, that treating people extremely poorly requires the ability to see them as less than fully human. This is a kind of descriptive, historical claim – one backed up by powerful psychological evidence and anecdote. Consider the infamous and classic example of the Rwandan genocide, a genocide enabled in part by widespread propaganda that literally claimed Tutsi's were "Cockroaches" - less than

[11] Indeed, this is what we should expect given work on epistemologies of ignorance, which I will discuss in the flowing chapter. As Charles Mills says in his discussion of white ignorance, "the concept is driving the perception, with whites aprioristically intent on denying what is before them. So if Kant famously said that perceptions without concepts are blind, then here it is the blindness of the concept itself that is blocking vision" (2017, 63).

[12] (See also, this overview: `https://www.newyorker.com/magazine/2017/11/27/the-root-of-all-cruelty`.

human.

I have a great deal of sympathy for *something* like the humanist claim. That is, I'm *enough* of a Hegelian to think that mutual recognition is an important component of morality, and I've already argued that the ways in which we interpret and see one another are directly linked to our responsibility judgements. But again, in all but the most extreme cases, why think that a lack of recognition *of humanity* is what drives something like outcomes of racism, sexism, or misogyny? Here's how Manne puts it:

> Under even moderately non-ideal conditions, involving, for example, exhaustible material resources, limited sought-after social positions, or clashing moral and social ideals, the humanity of some is likely to represent a double-edged sword to others. So, when it comes to recognizing someone as a fellow human being, the characteristic human capacities that you share don't just make her relatable; they make her potentially dangerous and threatening in ways only a human being can be— at least relative to our own, distinctively human sensibilities. She may, for example, threaten to undermine you. (148)

I think this gets things exactly right. It is not, for most of us, that seeing someone of another race triggers a set of beliefs and desires centered on their lack of humanity – it is that we are socially conditioned to view people of different races in complex (and often negative) ways *based on* their (supposed) distinctive human qualities.

Consider again the recognitive errors that likely occurred in the Arbery case. They are those involving recognizing one another primarily as members of salient identity groups - rather than recognizing each other as something other than or less than human. The mixing together of biases, affective responses, and stereotypical heuristics about certain identity groups with our explicit beliefs and judgements puts us in an epistemically precarious situation - one not best explained by the humanistic thesis, or indeed, by the general thesis that explicit bad beliefs are all that need revising. So, for instance a recognition that you are black by someone prone to anti-black bias, probably does not deny your humanity – but given the white supremacist history and enduring anti-black structuring of our society, it does bring to bear certain kinds of negative affective cues. The downstream effects of those cues may involve morally pernicious actions towards you, actions that can be rationalized in various ways.[13]

Again, there certainly were explicitly racist things said about Arbery's killing, but the internet commenters I am discussing saw themselves as providing an alternative *non-racist* justification: he shouldn't have been walking in the wrong neighborhood, and he shouldn't have been snooping around a construction site. Things are, as is always the case in real life, messy and complex in these cases. The line between explicit and implicit racism is not a clean and bright one.[14] My arguments don't depend on there being such a clear line. Instead, they depend on the fact that responsibility practitioners are often working at a remove from both relevant metaphysical properties and explicit belief formation – operating instead in the land of narrative, interpretation, dialogue, and social signalling.

3.3.2 So where do we stand?

What, then, is the upshot of my working through the Arbery example and the accompanying objection? I argue that we are reliably bad at forming responsibility judgements across gaps in social power, and that the above example is a clear case of this kind of difficulty. We are likely to let the socially powerful get away with too much, and to be too quick to hold the less-powerful responsible. The question then

[13] Indeed, such rationalization is to be expected, given that, on the one hand: a) enforcing group hierarchies is self-serving for the members in more powerful groups, and, on the other hand: b) there is often widespread societal pressure against straightforwardly bigoted enforcement. As Bicchieri (2017) puts it: "Self-serving biases often occur when a choice can be publicly justified as reasonable, if not optimal for all of the parties. We rationalize our behaviors, but the reasons must pass muster with the relevant audience" (78). The reasons that those (who are not self-described racists) are likely to give to excuse their racist behavior must be rationalized semi-publicly in the kind of case we are examining.

[14] And it is even messier when one considers the evidence from the literature on managed and constructed ignorance that often accompanies race concepts.

is: if we are reliably bad at forming responsibility judgements across gaps in power, what should be done about it?

We are faced with a choice: eliminate our faulty practices, retain them, despite the serious faults, or revise them to make them better. I'm going to argue that revision is the best way forward. What kind of revision? Below, I'll argue for the following principle, which will need much more support:

> **Powerful Restraint**: the socially powerful ought to, in general, refrain from blaming the less powerful in contexts where large gaps in power are prevalent.

Almost immediately, when one considers such a proposal, things get very tricky. First of all, as I've said, social power is highly contextual and non-static. Good luck figuring out who is more "powerful" in many real-world cases. Protests in May of 2020 in Minneapolis provide a good example of this kind of positionality. After the killing of George Floyd, several days of peaceful protests turned into riots, and eventually a police station was burned to the ground. In such a context, would it be fair to say that a protester, burning a police precinct, has more social power than a police officer in the city? I don't think there is a clean answer to this kind of question. In some sense, it is hard to imagine a protester having more material power than a police officer, who is an arm of state sanctioned violence and law enforcement. Yet, many observers were supportive of the protesters, and it is clear they wielded enormous social import and status. They were also able, in this instance, to physically overcome the police force, who they greatly outnumber. Again, one can draw their own conclusions here – the point is that analyzing power, because it is non-static and highly contextual, isn't likely to deliver easy answers for us about who should hold who to account, or who should refrain from engaging in the practices of responsibility.

On the other hand, I began with a near-edge case for a reason. It would be ludicrous to claim that Arbery had more social power than the men who killed him – even if there is some way in which social opinion or sentiment is more sympathetic to Arbery than armed vigilantes. Still, it isn't very helpful here to give a prescription like: "those who obviously have more social power should refrain from forming responsibility judgements against those who have less." First of all, it is unclear whether this prescription is psychologically plausible. There's something to this objection, but we can leave it aside for now. For surely we *sometimes* can refrain from forming such judgements – and, at the very least, surely we can refrain from *acting* on our judgements of responsibility, in many cases. So, perhaps one natural prescription is that our blame and praise ought to be highly tentative in cases where we are working across large gaps in social status and power.

A larger question looms in the background: what kinds of practices, if any, are justified if these gaps are truly disruptive? It is a difficult question to answer in the abstract, because we need a sense of what would vindicate one kind of answer over another. In the end, I'm going to argue that many of our responsibility practices*are* justified, despite the epistemic situation I've been describing. This is so because what licenses them is our stake in enforcing valuable social norms (or dis-incentivising dis-valuable ones). As Cheshire Calhoun famously argued, in cases of injustice which occur at the level of social practice, "The question of blame becomes not just a question about blameworthiness, but more important a question about our entitlement to use moral reproach as a tool for effecting social change" (389). Prescriptions against responsibility judgments must proceed from instrumentally justified premises, in other words. We need to ask what the aims of a responsibility practice are, and what configurations of the practices would best meet those aims. Then it is an open question whether something like blaming across large gaps in power is instrumentally justifiable – and if so, when and how.

3.4 Revisionism, Eliminitivism, and Instrumentalism about Responsibility

Making sense of responsibility judgements in irreducibly practice-based and normatively-laden social settings is, unsurprisingly then, going to hinge on some claims about the nature and justification of those practices. I've presented a rough argument above which leaves us with several stark choice points. If it's true that we are ineffective at forming judgments of responsibility across significant gaps in status and social power, then we can: a) keep our current practices despite this, because we think that they are the best that we can do (or that it would be impossible to revise or replace them), b) eliminate our current practices (either in favor of something radically different, or in favor of nothing at all), or c) revise our current practices in more or less significant ways.

Let me make clear that choosing a path forward involves answering three separate questions: first, a methodological question of how to understand what would justify choosing one way or another, second, a substantive question, given that methodological choice, of what we ought, in fact, to do, and third, a practical question of how to go about affecting the changes we decide to go for. I argue for the following three answers: 1) that we ought to go in for an instrumentalist methodology of justification, 2) that this licenses significant revisionism of our practices, and 3) that the changes can be pursued by carefully exploiting social norms and pressures. I will largely leave aside the third question of practical change for now, but let me briefly say more about the methodological and substantive answers here.

I remind the reader that this dissertation has endeavored to be explicitly theory neutral when it comes to moral responsibility. Before I make substantive and controversial claims about our extant practices, let me note that the arguments I've laid out so far can (and should) be taken up by any number of theorists. As I've said, for instance, nothing about my view entails (or precludes) realism about moral responsibility. I've happily accepted, for instance, that the properties that metaphysical accounts of responsibility claim we ought to track may really exist, and, even that we may sometimes be tracking them. My claim is that, epistemically speaking, we are often bad trackers, and that, in any case, the tracking must be done indirectly - we have no special epistemic access to the metaphysical properties themselves.

My theory is also amenable to free-will skeptics, particularly those like Pereboom (2014) who are worried that folk-philosophical theories of moral responsibility invoke a notion of "basic desert" incompatible with the truth of determinism. I note, for instance, that a theory like McKenna's comes closest to dealing with the complexity of epistemic cases I imagine, and he has argued (convincingly enough for Pereboom (2014, 134) to accept) that his view needn't entail basic desert. And, if one wants to go further and do away with responsibility all-together, everything I've said can be used as grist for the error-theory mill. The fact that, as I've argued, we are often very bad at forming responsibility judgements can serve as further support for eliminitivist, incompatibilist intuitions.

However, this is not my project, and, in any case, the conceptual ground here is rocky. I'm claiming that the epistemic processes of forming responsibility judgments are often hijacked, distorted, and disrupted. This means that, at least some of the time, we are doing a bad job of forming correct judgements. But, one way of reading this claim implies that we could be doing *better*, or that there are non-erroneous judgements to make after all. This is the reading I favor. I don't, therefore, endorse an error theory about moral responsibility, or the kind of incompatibilist position that says that no one is ever responsible. Furthermore, I don't subscribe to a kind of nihilism that would lead us to believe that there is no possibility of revising our practices because we are psychologically incapable of doing so. With reminders of the neutrality of my project up to this point in hand, let's get clear about precisely what is meant by holding a revisionist view or pursuing a revisionary theory.

3.4.1 Revisionism

Vargas (2018a) defines a theory as revisionist when the truth of that theory is in conflict with a commonsense view, *and* where that theory seeks to retain the concept, practice, or term in question. This is contrasted with conventionalism and eliminitivism: retaining the common sense target unchanged, or getting rid of the target altogether.

As Vargas writes, “typically, there is a methodological component to revisionism; the revisionist theory is the ensuing distinctive substantive results that come about from adopting the methodology” (1). In our case, the relevant revisionism is indeed *methodological*, and the methodology involved is *instrumentalism* about the justification of our practices. This kind of revisionism largely leaves folk concepts and definitions untouched. However, given the shift to instrumental justification, there may be some substantive pragmatic revision of norms internal to the practice itself.

Vargas also importantly notes that revisionism about conceptual matters needn’t entail practice revisionism. That is, one might endeavor to redefine a concept, but imagine that the practice utilizing the concept would be untouched by the redefinition. This is perhaps most clear in folk-scientific examples: redefining the precise conception of gravity that physicists hold would not do much to shape the practices of users of the every day concept of gravity.

In a mirror image of this, I’m suggesting that my work is practice revisionary without being concept revisionary. The most I say is that my epistemic considerations constrain, very modestly, the range of theories which are contenders for the correct account of responsibility and its metaphysical conditions – but as I indicated in chapter one, it fits with almost all extent compatibilist views. My theory, in noting that there are under-emphasized (or altogether ignored) epistemic dimensions to responsibility, attempts to re-frame the way we think about and enact responsibility judgements, while holding that how we define those concepts is a largely independent matter.

Why think that revisionism, in this sense, is more appealing than eliminitivism? After all, I've suggested that once the epistemic picture becomes clear, we'll see that we often have no direct way to tell whether we are doing a good job tracking metaphysical properties, and that it is quite clear that we often do a *bad* job of forming responsibility judgments. If we are habitually bad responsibility practitioners, wouldn't it be better to get rid of the practice?

Much here depends on how comparatively bad scrapping a practice would be, how likely one thinks revisionism is to succeed, and how radical the revision necessary to improve the practice would be. Vargas notes that one appeal of revisionism is that, "coherence with folk commitments is a crucial/constitutive desideratum of many philosophical methodologies" (6). This counts in favor of revision, given that my view is arguably *more* coherent with folk commitments since very few of the folk think they are tracking properties like reasons-responsiveness when they make responsibility judgments.

Beyond this, we can note (as Vargas does) that the "prescriptive" question of what kind of theory of responsibility we ought to have is at least somewhat independent from the question of the truth of various folk commitments. This is all to say that, given a strictly neutral standpoint about the truth or falsity of the metaphysical components of responsibility, what will vindicate revisions to our practice will be instrumental concerns about how to best achieve the aims of that practice. This assumes that retaining the practice is more valuable than ridding ourselves of it, and that revising the practice is more valuable than leaving it as it is. I now turn to arguing for these claims.

3.4.2 Instrumentalism

What does being an instrumentalist about responsibility amount to? Here it is important to distinguish between instrumentalism about a practice itself and instrumentalism about moves within a practice. The distinction between practice internal and practice external justifications is part and parcel with the Strawsonian tradition and, in this sense, is an insight that many of the theorists I surveyed in Chapter 1 would happily take up. The distinction helps us because it clears up a possible confusion: that being an instrumentalist about responsibility means that one is a consequentialist about the justification of instances of responsibility ascription. However, this needn't be the case. Even if the justification of a practice appeals to consequentialist grounds, this needn't redound into the internal dynamics of a practice itself.

Take the following example: I might think that the current system of Ultimate disc rules is justified insofar as it leads to the greatest possible enjoyment of the sport amongst ultimate frisbee players overall.[15] I don't thereby commit myself to the view that each individual rule dispute must be settled on consequentialist grounds. Indeed, there are many rules that often lead to a *less* enjoyable game in particular instances, but which are justified because of their overall effect on the sport.

Notice, of course, that I can be an instrumentalist about the rules without being a consequentialist at all. The rules of Ultimate are, in fact, based around a notion of the "Spirit of the Game" - a souped up and self-aware version of sportsmanship. The idea is that adhering to Spirit will: a) keep fair competition firmly in mind, leading to fairer sporting outcomes, b) lead to more pro-social behavior, and c) over-ride other considerations. In particular, according to c) winning (which may lead to more overall utility) is no excuse to disregard spirit. Clearly, these are, at least partially, non-consequentialist, instrumentalist grounds for justifying the current rule-set.

The Strawsonian tradition I take myself to be a part of treats moral responsibility in a similar manner. The practices of moral responsibility have internal standards which *make use of* reactive attitudes and backwards looking considerations involving

[15] The 12th edition, if you're counting: `https://www.usaultimate.org/resources/officiating/rules/2020_2021rules.aspx`

desert, retribution, and moral ledger keeping. And whether someone is responsible, on this view, is a question internal to that system. As I've insisted, this is why I take it that the epistemic challenges I'm raising are not, in and of themselves, refutations of particular metaphysical views about responsibility. Instead, the instrumentalist view I help myself to merely claims that what justifies the *practices* are (at least in part) their forward-looking effects.

This kind of instrumentalism is compatible with a wide range of views about what the goods of the practices under consideration are. What the instrumentalism says is simply that, when we are faced with a choice about how to secure those goods, we ought to prefer the arrangement of practices that does the best job getting them. The substantive theory of "best" here is left vague, but all we need is the sense that some arrangements are better or worse than others. In other words, for any social practice (including moral responsibility) it is an intelligible and informative question to ask: "are there better ways to organize the practice?"

I think it's obvious that this *is* an intelligible question. However, it's fair to put pressure on the idea that we have enough of a clear way of working out an answer to it that the question will be very useful. Another way to put the challenge is that we need some standpoint from which to adjudicate arrangements of the practice in principled ways. Here, however, I think a minimal answer is perfectly sufficient: What we can do is try out various provisional standpoints, and see how they fit with the rest of our commitments. So, for instance, we might say: "suppose we want our practices of praising and blaming to contribute to agency cultivation, how should we arrange the practice then?" Or, "suppose we want our practices to track desert - how ought we to set them up to best do that?"

Even here, we can limit ourselves merely to the sense that we are instrumentalists insofar as we are trying to answer the question of how best to justify and arrange our practices given that they *need* some justification and arrangement. We can be quite ecumenical, in other words, and simply say: here's a range of things that inform what kinds of goods we'd want from our practices and a range of ways we could get them. How radical the revision of a practice can or should be is an open question at this point.

What might those goods be when we think about moral responsibility? One popular instrumentalist answer is that the goods are some kind of improvements to our agential capacities. The idea here is that our practices of responsibility can be justified insofar as they, in some sense, make us better agents overall. There is a rough split amongst instrumentalist views of this kind between those that favor overall "agency cultivation" of responsibility users as a goal, and those that favor agency enhancement in the moment. On the first branch, as Vargas (2013) puts it, the justification depends on, "securing the (putatively valuable) goal of fostering and refining our ability to recognize and respond to moral considerations." On the second branch, as McGreer puts it, the question is, "whether... blaming enhances the wrong-doers ability to be suitably responsive to moral reasons" (McGeer 2013, 2015). I broadly favor Vargas' "rule" instrumentalism over McGeer's "act" instrumentalism, but happily take on board their key contention that our agential capacities are "elastic, socially-scaffolded," and moldable via the inputs of responsibility practice.

The argument, then, is as follows: Our responsibility practices have some positive value in securing the goods of agency cultivation and/or enhancement. Maintaining our practices as they are fails to deal with the epistemic distortion I'm outlining - it is thus a sub-optimal option – we could be doing more and better cultivating. Eliminating our practices all-together is both psychologically implausible and, possibly socially catastrophic – it is a bad option. Replacing our practices whole-sale with some other system is, again, implausible and unconvincing unless the defender of replacement can show that their alternative would lead to as much or more improvement of agency. Revising our practices, then, seems like the only option which will deal with the issues I raise that is plausible and not entirely disruptive or socially revolutionary. Given this, we ought to see what kinds of revisions are possible, and if they are likely to be instrumentally preferable to maintaining our practices as they are.

Justifying Moral Responsibility?

I am spending a good deal of time on theoretical and methodological architecture because it is reasonable to believe that the revisionist about responsibility has a special justifactory burden; a burden that instrumentalism, suitably understood, helps discharge. Given that our practices of responsibility seem very "basic" in some social and psychological sense, and given that revisions to them will have wide ranging real-world effects, *and* be socially difficult to bring about, the revisionist must adequately explain why proceeding with revision is worthwhile.

Not only this, the revisionist has a special burden to explain why they are, in the end, talking about the same things as the folk and other theorists. On Waller (2014)'s view, for instance, to be a revisionist about responsibility is almost certainly to shirk one's duty of answering the *real* questions of moral responsibility for some nearby easier questions. As he puts it:

> [A]lmost all philosophers — along with almost everyone else in Western culture — believe that it is obvious that many claims and ascriptions of moral responsibility are true and justified. Thus, rather than wrestling with the painfully difficult problem of justifying moral responsibility, there is a tendency to turn to a somewhat easier question that is related to the question of moral responsibility, deal with that question instead, and suppose that we have an answer to the hard question of moral responsibility. (36)

So, at least if one wants to satisfy Waller, one must convince one's interlocutors that they haven't pulled a philosophical bait and switch. Can I, in other words, sufficiently revise the relevant responsibility architecture to address the concerns I've laid out without changing the subject? If I can't, then it would be best to follow

Waller into eliminitivsit territory – to admit that there's nothing to be done about responsibility and give up on the "stubborn system."

Why think there's a problem here? After all, I've argued that what I'm after is *practice* revision - not revision to our concepts. But, Waller will not be impressed with this answer. There is still a question of whether the revised practices will be the practices themselves, or some kind of bait and switch. And, insofar as one goes for a practice-internal justification, one is committing the cardinal sin of substituting a harder question for an easier one. As he says: "It is certainly easier to establish that some people are morally responsible according to the rules of our moral responsibility system (it is much easier to answer that *internal* question) than to establish that the moral responsibility system itself is justified" (29).

However, Waller's issue seems only to arise because he wants to keep a tight link between: a) questions of basic desert, b) questions of responsibility, and c) questions of practice justification. For him, a thorough answer to one will be a thorough answer to all three. That is, if I'm claiming that you *really are* responsible, then I mean that you really deserve blame or praise, and that the practice that allows me to so blame or praise you is really justified. If the practices are justified, then people really are responsible for their actions, and so they really do deserve blame and praise. The link is so tight that the questions collapse. As he says, inveighing against those who would give us responsibility "re-defined": "The point here is that the basic question of whether punishment and reward are fair — whether they are justly deserved, apart from all considerations of utility — remains an important question, and it is a question that must be addressed, whatever other uses may be found for moral responsibility" (38).

But, of course, this is just what many compatibilists (and certainly many revisionists) deny. In forging new understandings of responsibility, we needn't hold that ideas a, b, and c above are so tightly intertwined as we may have at first thought. Contemporary theorists, following Watson (1996), commonly hold that attributability and accountability can be disentangled. This is precisely what Waller denies. Rather than letting this ground out in table thumping, I think we can give intuitive reasons why such notions *ought* to be distinguished. The constitutive claim which some have read Strawson as making - that responsibility just *is* proneness to certain reactive attitudes, seems too strong to most. At the very least, the question of whether one is responsible for an action surely can and should be distentangeled from the question of whether one ought to be held accountable or punished for it. In the real world, we often run these conceptions together - moving from attributing responsibility straight to punishment in some contexts. But just as often, we keep them separate: we pause and reflect about whether holding someone to account is worth it, is fair, or would serve our overall goals. The fact that, in the real world, these elements often run together, but that they are also conceptually distinct, is precisely one of the reasons we ought to pay attention to the epistemology of responsibility.

Whether responsibility and *desert* can be distentangled is, obviously, a thornier question. We might think there is a very tight link between being responsible and being *blameworthy* or *praiseworthy* – a link which is much tighter than that between blame and punishment, or praise and reward. Derk Pereboom's concept of basic desert holds that an agent "would deserve to be blamed or praised just because she has performed the action, given an understanding of its moral status, and not, for example merely by virtue of consequentialist or contractualist considerations" (2014, 2). Deservingness is thus, on a view like this, intrinsic to actions - those things that agents do in a morally structured world. It does not depend on further considerations whether they are consequentialist, contractualist, or, as I've been exploring in this chapter, instrumentalist aims of a practice.

So how does my account deal with this idea of basic desert? How can we respond to Waller? I hold that the epistemic issues I've laid out in Chapters 1 and 2 add a further layer of complexity onto the identification of desert relevant features. Again, whether or not something like basic desert exists is a question I am happy to be neutral about. But even if basic desert *is* a real feature of actions, my claim is simply that we have no direct access to knowledge about that feature. What we have are interpretations in a socially and normatively scaffolded world. Unlike Pereboom I do not take the whole notion to be necessarily mistaken, but I do argue that we must often demure to instrumental considerations when deciding who deserves what, given that we can't know the truth of basic desert one way or another.

In this way, we needn't reject the idea of basic desert (although I am sympathetic to rejecting it, and my theory is compatible with its rejection). Nor must we provide the realist conditions under which a notion of basic desert could be met. Nor, again, must we engage in revision of the concept of desert. It is enough to say that, whatever the right conditions for desert are, we have common epistemic distortors that get in the way of our reliably knowing them for candidate actions and agents. It is this sense, once again, in which I am *practice* revisionary, and appeal to instrumental aims to answer questions of whether we ought to hold responsible.

3.5 Responsibility Revisionism: Against Blame?

Now that I've discussed the sense in which I am a revisionist and an instrumentalist, I can return to some specific proposals. If we are likely to let the powerful get away with too much, and to be too quick to hold the less-powerful responsible, a question naturally arises: what should be done about it? I've already said I'll put eliminitivism aside. Perhaps, in the end, retaining our faulty practices is all we can do - but given that we recognize them as faulty, it seems worthwhile to *try* to revise them, to the extent such revision is possible.

But what kind of revision? There is probably no easy, practice-wide answer to this question. As I've said, social power is highly contextual and non-static. Recall the example of protesters burning down a Minneapolis police station. We ought to, as I argued in the previous section, ask what the aims of our responsibility practices are, and what configurations of the practices would best meet those aims. Then, it is an open question whether something like blaming across large gaps in power is instrumentally justifiable – and if so, when and how.

Given the fact that determining precise power relations is difficult, one very natural candidate for revision is the following principle:

> **Blanket Blame Reduction**: Given that gaps in social power are prevalent and commonly cause epistemic distortion, we all ought to be more cautious about our judgements of moral responsibility; in particular we all ought to refrain from blaming across large gaps in social power.

Would the introduction of this kind of norm improve our responsibility practices? I think the norm is misguided, and imagines a silver bullet where, unfortunately, none exists. Its approach to a structural problem in our practices is to treat all the individual members of the practice roughly equally. It asks all of us to refrain from blaming as often as we do (especially when we are aware of gaps in power). There are several problems with this approach. First of all, in asking the less powerful to exercise even *further* humility, it both fails to correct the fundamental imbalance of power that gives rise to the most pernicious problems that I've identified, *and* takes away one of the only tools that oppressed peoples have to fight injustice: social sanction. The socially powerful are often given too *many* free passes. The epistemic errors we make tend to put us on a course towards blaming the powerful too little.

Secondly, blame, on many views, has agency and society improving features. Putting aside questions of basic desert and focusing only on instrumental reasons for blaming, there might be a real cost to reducing the amount of overall blame in the world. How these things shake out, and how we might weigh and balance the costs and benefits are complicated - and I will say more about this issue in chapter four. For now, I simply note that the more course-grained and revisionary the norm being proposed is the more it faces two problems: 1) a greater uncertainty about what its actual effects might be, and at the same time, 2) a greater likelihood of imposing significant social costs.

Not only this, Blanket Blame Reduction merely reproduces the flawed normative landscape we've been discussing at a higher order. The introduction of this norm would also introduce a "metanorm" that says we ought to blame those who fail to sufficiently reduce their blaming tendencies. But, because the scope of this norm would include those with low social power as well as those with high social power, the same distorters will manifest at this meta-level as well. The powerful, in other words, won't get their fair share of blame for failing to reduce their blaming tendencies, while the less-powerful will get too much.

Finally, one might be worried that a general prohibition on holding one another responsible will only increase cultivated ignorance.[16] It is hard to see how active ignorance which leads to epistemic injustices can be overcome by a system that asks us all to refrain from blaming for *fear* of being ignorant. It seems entirely open that groups who rely on strategic ignorance to, for instance, maintain a dominant social position without guilt or moral reckoning will be able to further entrench that position if they can act without fear of reproach (and if they can react to blame by saying, "you are blaming me across a large gap in social power - you ought to have more reduced your blaming tendencies!").

For these reasons, let me suggest a first pass at a more specific norm which retains an asymmetry in blame's acceptability:

> **Powerful Restraint 1.0**: the socially powerful ought to, in general, refrain from blaming the less powerful in contexts where large gaps in power are prevalent.

In order to explain why we ought to favor Powerful Restraint I need to do three things: First, I need to sketch out the sense of blame I am working with. Second, I need to further clarify the scope and grounds of the norm of Powerful Restraint -

[16] See, for instance Charles Mills (2017) work on what he calls "white ignorance."I will say much more about this in the following chapter

to whom, exactly, does it apply? And what grounds this *prima facie* obligation?[17] Second, I need to motivate the idea that forgoing blame in the way it imagines could be socially beneficial at all, given the worries I canvassed above.

3.5.1 Metaphysics, Scope, and Justifying Blame Reduction

I've already argued at length that we can be ecumenical about the notion of responsibility that's in play - and that this ecumenicism is earned. Focusing squarely on our everyday practices, we are asking: "what kinds of things are unnoticed distorters or defeaters of our epistemic abilities to accurately, fairly, or usefully track and judge that people are responsible?" The point is to try to get clear about where our epistemic practices of moral responsibility break down and what the causes of those break-downs might be. To do that, we merely need a sketch of our current responsibility practices, from which we can think about what's going to count as a distortion or defeater under a wide range of credible views in epistemology and the metaphysics of responsibility.

Concerning blame, a similar set of arguments can be run. Again, my aim is to be as theory neutral as possible. Whether blame is a reactive attitude, a cognitive state such as belief, an adjustment to a relationship, or a form of conversational protest doesn't matter for the purposes of my argument. We are all aware of the social reality of blame, of the way it feels (both to blame and to be blamed), and of the many forms it can take. Whether there is a univocal (or pluralist) account that we can give of blame's necessary and sufficient conditions won't matter for the two points I make in this chapter. First, that there is an epistemic problem in our practices of moral responsibility, and second, that this problem leads to our blaming badly, and our doing so reliably. Insofar as I take a stand on the many interesting questions of the nature of blame it is only to say the following: however

[17] Thank you to two anonymous reviewers for pressing me to clarify these important points.

one wishes to precisely explore the contours of blame as an emotional, cognitive, or interactional process, we should be able to recognize that blame has what I'll call an "assessment phase" and an "expression phase." The process of coming to form a judgment of blameworthiness is separate from choosing how we we express (or do not express) that judgment to ourselves, the blameworthy individual, or others. Different theories treat these phases in different ways, and I take no stance on the correct way of thinking about the distinction between blameworthiness and blame - all I need to point out is that a gap between judgment and action exists here.

Next, let me clarify the scope and grounding of the norm of Powerful Restraint. In Chapter two, I canvassed a set of epistemic issues that led me to claim that it's likely that the powerful often blame the less-powerful too frequently. Given this, one might ask of Powerful Restraint: Isn't the norm unnecessarily strong?[18] That is, why think that anyone needs to, in general, reduce their blaming tendencies? Wouldn't a more judicious norm ask them to *raise* the epistemic bar their blame needs to clear, rather than reduce it *tout court*? We can imagine that if the powerful focused on better collection of evidence, on correcting for biases, and on double-checking their accounts of what agents did or why they seem blameworthy, the tendency for blame to be misapplied across gaps in power would be reduced. There are two responses I can give here. The first makes clearer the relationship between the epistemic problem I canvassed in section one and the moral problems that result from it.[19] The second points out that this raising of the epistemic bar is meant to be contained in the structure of the norm itself.

The first answer is that I suspect there is no clean and sharp distinction between the moral and epistemic in the kind of cases I'm considering, and that trying to make a clean distinction will be less helpful than it might at first seem. Nothing I say here precludes the possibility that distinguishing carefully between the downstream moral

[18] Thank you to an anonymous reader for pushing me to clarify this point.

[19] Again, my sincere thanks to an anonymous reader for pointing out that the distinction between whether this is a moral or epistemic matter must be made clearer.

effects of an upstream epistemic problem is the right way to frame the issue. However, my claim is that these two issues are reliably blended in the case of the epistemology of moral responsibility, and that there's a specific structure to that blending that is worth dealing with in itself. How so? We have a set of moral practices which ask us to attend to certain kinds and sets of evidence – our best theories of the metaphysics of responsibility tell us to be responsive and attentive to that evidence, and our best theories of epistemology tell us how we ought to gather and assess it.

However, I've claimed that it's both the case that: 1) actual agents in the practices do not follow these standards reliably well, and 2) that the standards themselves may be suspect or faulty in various ways. This calls for revisions in the practice, as I've argued. And, what's going to govern the revision will be structured both by general features of good epistemic practice, and by the particular role those epistemic practices play in the practice of moral responsibility. We have, in other words, an overlapping structure – there are independent epistemic norms about how to treat evidence and form reliable judgments and there are internal, moral norms about how to be a good participant in the practices of moral responsibility. The epistemic norms can be seen as necessary but non-sufficient conditions on our forming good judgments of responsibility. But there are also distinct necessary conditions which emerge from the moral responsibility practice itself. To know whether I ought to blame you I need to know that I'm tracking the right kind of evidence (and tracking it well), but I also need to know what my blame will *do*; whether it will cause undue harm, whether it will be *fair*, and so on.

Where epistemic problems are likely to occur in our practices, I'm claiming then, moral problems follow close behind. And again, I've argued that these moral problems are not random, but are reliably structured by the practices themselves. If it's the case that the powerful are reliably bad judges of the character of powerless members of society, for instance, then they are reliably likely to blame them unjustly. This leads to a distributional problem – an unjust balance of blame at the societal level. This is a moral issue, but one rooted in a particular epistemic problem. So, Powerful Restraint is moralized - but not haphazardly. It suggests a solution that is partially epistemic and partially moral because the problem it responds to is a downstream moral problem resulting from an upstream epistemic one.

Still, one can press: why not just deal with the upstream problem in isolation? Here the second line of response is called for. Recall that the "in general" clause in Powerful Restraint is meant to indicate that there is no blanket prohibition on blame from the powerful - rather there is normative pressure for more care than they often exhibit before blaming, *as well as* pressure for a general reduction of their blaming tendencies (or, an increase in their hesitancy to blame). In other words, the spirit of this objection is already contained in the formulation of Powerful Restraint.

More importantly, however, I am skeptical of the idea that the biases that may lead to epistemic distortion in these cases can be sufficiently corrected or accounted for upstream of blame, such that blame would once again be (in a general way) appropriate. Instead, one thing the chapter argues for is that since we, in general, cannot know the precise factors our blame judgments are formed by and react to, and since we cannot know the precise features of an agents' metaphysical or characterological make-ups that would vouchsafe blame, we cannot make our blame legitimate *merely* by being more epistemically cautious or working to reduce our biases.

Roughly then, we can say that: a) yes, the powerful (and others) should work to improve their epistemic processes as much as possible - we want to get blame right when we can. But, b) no matter how hard we work, mistakes are possible, and become increasingly likely in the kinds of cases I describe where we operate across large gaps in social power and prestige. Given the moral costs of these mistakes and our inability to eliminate them, I claim that Powerful Restraint is justified.

Finally, why should we believe that forgoing blame would be socially beneficial at all? For two reasons: First, by focusing on the *downstream* products of responsibility judgements, Powerful Restraint is far less revisionary than a view that asks us to do without appraisal in the first place. Recall the worry above that our responsibility practices are psychologically ineliminable. While it's true that our "hot" and non-conscious psychological systems may already be priming us to form judgements before we have a chance to assess evidence or do a great deal of careful cognition, we still often have the chance to reflect before *expressing* those judgements. Forgoing blame, in other words, at least in its *social* or dialogical guises, is far more under our voluntary control than forgoing attributions of responsibility in total.

Second, there is good evidence that forgoing blame can have instrumentally beneficial effects in many important contexts. The work of Hanna Pickard (2013), for instance, argues that such a bifurcation between judgements of responsibility and blame is crucial in clinical psychiatric work. Pickard (2013) begins by noting a uniquely *clinical* conundrum: in institutional settings where service users suffer from disorders of agency (bi-polar disorder, for example), caregivers must hold them responsible for their actions while avoiding blame. This is so because holding responsible is crucial for: a) treatment, and b) respecting service users as agents and persons, while blame, on the other hand, is highly detrimental for treatment (see 1135-1138).

The key point is that, although many service users may have *diminished* amounts of control or conscious awareness (or, we can assume, whatever other properties, capacities and faculties one's theory of responsibility calls for), on any notion where these capacities are graded, most will pass a threshold of responsible agency most of the time. Their excuses for diminished responsibility do not exempt them from the practice wholesale. Exemption, perhaps by taking a Strawsonian objective attitude, would deny them their agency – something which would both be disrespectful and counterproductive to the intended therapeutic interventions.[20]

Yet, it's also the case that, in terms of effective treatment, blame, expressed with a characteristic emotional "sting" is highly detrimental in clinical settings. Pickard thus gives us a clear example of a setting where forward looking concerns (the goal of

[20] The objective attitude is famously discussed by Strawson (2003, 79–83).

proper and effective treatment) shape the way in which the practice of responsibility takes place in a particular social arrangement. Pickard notes that, quite obviously, the large gaps in power between clinician and patient also play a role in making affective blame ineffective. We are offered a practical, instrumentally justified solution to the problem: take responsibility for our own emotions and affective responses, keeping in mind the complicated power dynamics between patient and caregiver.

All of this is to say that holding responsible looks different in different contexts. Insofar as different contexts call for different instrumental justifications of blame, our responsibility practices can involve revisions to the frequency and type of blame we engage in. The revisions to our practices which are licensed depend on what we think those practices are good for, and how we think we can best achieve those goods. When we look at the context of the high-status blaming the low-status, what further reasons do we have for thinking that reducing the expression of blame may be beneficial?

3.5.2 Issues of Incentivization and Positionality

One reason to minimize the flow of blame from high to low power individuals is due to the following concern: those who are low-status are going to be incentivized towards certain instances of "blameworthy" behavior in a way that those of high status are not. Not only this, but those with power and privilege are often partially responsible for the structure of those incentives to begin with, and so may not be in an appropriate position to blame those of low social-status and power. That is, where large gaps in social power occur, it may be that we are in the wrong kind of relationship to hold one another responsible.

How might this argument work? Lewis (2016), has argued that those who are complicit in creating the conditions which lead to blameworthy behavior, do something inappropriate when they blame. This is a rather intuitive idea: I shouldn't blame you for doing something I enabled (and perhaps foresaw as a likely outcome) - or, at the very least, my blame ought to be limited or tempered. Lewis advances this argument in a context which works well for our purposes: the fact that those who commit crimes are often disadvantaged persons in disadvantaged communities who are strongly incentivized to do so.

As Lewis notes: "Because blame is a response to a perception of a morally inappropriate attitude, it might be natural to think that blame is justified when that perception is accurate. But it is also natural to think that there is an important sense in which our actions and attitudes are justified only if we stand in the right epistemic position with respect to them" (158). What does Lewis mean here by the right "epistemic position?"

Lewis argues that are two "limiting conditions" on the appropriateness of blame, one Epistemic and one "Positional:" we are justified in blaming others for their actions only to the extent that we have *evidence* that they acted on a morally objectionable attitude, and only to the extent that we are *in a moral position* to hold them to a standard that attitude fails to meet" (161, my emphasis). What we need, for blame to be appropriate, is to have good evidence of blameworthiness, and to have the right kind of standing to act on that evidence by blaming. The rest of Lewis' argument attempts to show that we do not meet these limiting conditions as often as we think.

The basic insight driving the paper is that those who commit crimes are often incentivized to do so by the conditions in which they find themselves. They either do so because there are strong payoffs (in terms of whatever goods they find valuable), or because they think that there is a high likelihood of living a more overall valuable life if they do so. Importantly, this is *comparatively* true – their incentives are stronger than the advantaged, whatever the ultimate strength of the incentives is. This is important because it blocks an initial objection that *everyone* has some reason to engage in blameworthy behavior for illicit gain. This might be true, but if my incentives to do so are much weaker than social disincentives against committing crimes, then it is obvious that I am in a different kind of position from someone whose incentives are comparatively much higher than those social disincentives.

This comparative claim is especially important in understanding the work that incentives do. That is, an easy objection to Lewis' account says that, quite obviously, committing a crime which involves taking one's own interests (however strong) to be more important than the comparable interests of victims is all the evidence of a bad will we'll ever need. But, the epistemic limiting condition shows us that, in fact, committing a crime is not good evidence of a bad will when that crime is highly incentivized. At the very least, incentivization makes our evidence for a bad will *comparatively* weaker.

The positional move in Lewis' picture is to note that "we" are often partially responsible for the incentive structure that low-status individuals find themselves in. Given its intended audience of academic philosophers, the chapter's inclusive use of "we" is probably justified here. But we can soften the claim and bring it into alignment with my own: those who wield a great deal of social power are partially responsible for the construction and maintenance of the very incentive structures that incentivize low-status individuals to commit crimes. Given this, their blame is (at least partially) inappropriate.

3.6 Objections to Powerful Restraint

I've now argued that we aren't very good at forming responsibility judgements across gaps in social power and position, that this is due to general epistemic difficulties, and that our practices will, therefore, be reliably faulty. Furthermore, I pointed out that our practices widely involve such gaps, and, therefore, ought to be revised. Finally, I've argued that the revision cannot focus *merely* on epistemic matters but must be sensitive to moral and pragmatic questions as well. The revision I've begun to describe involves introducing new norms to our practice, and the particular norm I've outlined asks the powerful to refrain from blaming those with less social status across large gaps in power.[21]

This argument led to the principle of **Powerful Restraint** I introduced above. We ought to keep front of mind that the principle is not meant to depend on whether the powerful are likely to give up blaming the less powerful on their own - it is describing a normative aim. The idea here is to come up with a revisionary principle which can reorient and put normative pressure on everyone involved in our practices. The work of shifting norms, however, is incredibly complex and laborious - I don't mean to sell it short.

Let me try to make more precise the content and spirit of Powerful Restraint by dealing with a few further objections to it. To my mind, two main classes must be dealt with, those concern disrespect, and those concerning asymmetries in the resulting practice. Begin with disrespect:

> **Disrespect**: Declining to blame those of lower social status is straightforwardly disrespectful. It denies them full membership in the moral community - treating them either with the Strawsonian objective stance, or as akin to children.

I think this objection is persuasive, but that it can be met. There is very interesting recent work on the connection between respect and responsibility - work that I find compelling.[22] One central idea is that by choosing to withhold blame, we are denying agents a certain kind of respect. Consider a non-moral case: a well-meaning teacher has a student with a learning disability in their class. Instead of providing the student accommodations that would make the classroom equitable, they simply grade the student less critically – declining to hold them accountable for their errors.

[21] I've also hinted at the fact that the less powerful *ought* to continue to hold the powerful to account, and, indeed, that a second norm we may want to introduce into our practices is that they ought to do so more often, although I do not defend that norm here.

[22] For instance in "Blame and Patronizing" by Kathleen Connelly (forthcoming)

This is disrespectful, whether the teacher meant it to be or not. By denying the student an opportunity to be held accountable for their mistakes (in an environment in which it would have been fair to do so), the teacher is denying them full membership in the academic community, as well as the ability to improve their capacities.

So too in moral cases. By declining to blame those we might view as "less capable," we may be denying them access to our moral community, as well as the respect that goes with it, and the capacity to improve themselves. At a first pass, I think my view has a novel response to this kind of worry: in cases of Powerful Restraint, we are not declining to hold responsible because of a worry about the *agency* or *capacities* of the less powerful. Instead, we are declining to hold them responsible because of a worry about our *own* capacities.

It is harder to see how the charge of disrespect can stick in this situation. Imagine an analogy: if I am a surgeon, and I decline to perform a risky operation on you because I'm worried that I don't know enough about your symptoms to proceed, it is hard to see how this qualifies as disrespectful. Of course, we can imagine edge cases where someone of low-status *demands* that we hold them responsible (or, in the analogy, someone demands that we proceed with the surgery) but this just looks like a case where we have sufficient evidence to override the generality clause in Powerful Restraint – there will still be cases where the evidence is good enough to blame.

Still, there might be a nearby issue about the *perception* of respect which is worth taking seriously here. As always, things are very complicated: whether or not is worth taking seriously that one might be perceived as being disrespectful - and to take it seriously enough to override a competing normative prescription like powerful restraint - is, probably, best handled on a case-by-case basis. But recall the generality constraint embedded in the norm: if one is sure that declining to blame is riskier than blaming, because of issues of respect at play, one can always choose to continue towards blaming.

In other words, although it might be generally impermissible to blame across a large gap in power, it is no problem for the view if there are edge cases where permission *is* granted. I suppose the relevant question is how likely such edge cases are. There is no way to answer this *a priori*, but I can imagine one important set of cases that we should pay attention to. Return to our surgery analogy. One common complaint against our currently constituted practices of medicine in the United States is that they often discount the pain of women – and even more so the pain of women of color - particularly black women. We can imagine a case where a black woman is *sure* that she needs an elective surgery to reduce her pain, but a doctor refuses because of a lack of certainty about whether the procedure is necessary.[23] This class of cases *may* involve the kind of disrespect that the objection was after.

So too in the realm of responsibility. If a low-status community commonly *complains* that those in power fail to respect their agency by holding them responsible, I would take it that this is good evidence that continuing to eschew accountability practices *would* be disrespectful. However, I am not aware of this kind of claim being brought forward with much frequency. Instead, the opposite claim, that the powerful are too quick to hold responsible, blame, and punish the less powerful is common, and precisely the issue I am endeavoring to deal with. So, as long as we are not in the class of cases where communities themselves are demanding to be held responsible, I think the charge of disrespect does not go through.

Before moving on, we should consider a closely related objection: that denying the less powerful the opportunity to be blamed robs them of opportunities for self development.[24] After all, on many plausible models of agency development, part of what helps us become competent practitioners of moral responsibility is our being "in the game," so to speak.[25] We come to understand the relevant norms by coming

[23] See Jada Wiggleton-Little (2023)'s recent dissertation for in-depth discussion of exactly these kinds of cases.

[24] Thanks to an anonymous reader for pressing this point.

[25] For distinct and persuasive models of responsibility as a system of agency cultivation see:

into contact with them – we come to be competent blamers in part by learning when *we* are to blame. Not only this, blame serves a valuable social function as a signal that wrongdoing has occurred, and, perhaps, as a form of moral protest.[26]

We can say (at least) two things in defense of Powerful Restraint against this kind of complaint. First, the low-status are not robbed of opportunities to be blamed full-stop. It is still the case that those of similar (and lower) social positions can and should blame them for blameworthy behavior. It is also still the case that, given that Powerful Restraint will be imperfectly followed, blame may flow down from above. So, it is not as if the low-status will suddenly live in a blameless world. An argument more focused on articulating a specific account of blame itself might also rely here on issues of standing. That is, I find it unlikely that, in many cases, the powerful will be in the best position to have unequivocal standing to blame. I won't argue this point at length here, but suffice it to say that I find it highly plausible that there are likely to be other members of a community who would be in better positions to blame in most cases. There will rarely, in other words, be overriding reasons for the powerful to step in in such cases, given the dangers I've canvassed. And again, if we can imagine a context where the powerful are the only ones able to engage in corrective or agency-enhancing blaming, and where such blame would truly be importantly agency-enhancing, the "in general" clause allows that such blame can be appropriate. It should be obvious, however, that I find it unlikely that this kind of situation will be common in our practices.

3.6.1 Asymmetry Objections

Second, and very briefly, there is also a definitional question of what counts as blame as opposed to nearby forms of corrective or enhancing critique. I have said that I wish to remain ecumenical about the nature of blame, and so I won't have much

McGeer (2012, 2019) and Vargas (2013).

[26] See, for example, Hieronymi (2001).

to say on this matter. On some capacious definitions of blame, calm, dispassionate moral critique will count. It seems obvious that this kind of blame will be less likely to harm than full-throated emotional blame[27] However, I merely point out here that on other accounts of blame, well-meaning (or even friendly) critiques will *not* count as paradigmatic of blaming. It is open, therefore, that some of the kinds of moral critique this objection imagines are still perfectly open to high-status individuals, downstream of some conceptual fights about what does or doesn't count as blaming. In all cases, however, we can ask: is it really the case that high-status individuals will be in the best position or have the best standing to blame?

The final major objection to be dealt with targets the asymmetry in blame's acceptability that I've introduced:

> **Asymmetry**: Aren't the less powerful just as likely to err as the more powerful when it comes to epistemic processes concerning responsibility? And, going further, aren't there likely to be pernicious reasons unique to this context? If the poor want to send the innocent rich to the guillotine merely for existing, does my theory excuse this?

No answer I can give in the remainder of this chapter will be fully satisfying.[28] Here's a sketch of how a longer answer would go: first we'd want to ask to hear more about the likely outcomes of run-away responsibility ascription from low to high status. How likely is it that the rich are really going to the guillotine? Is a more likely outcome the re-distribution of wealth, or loss of opportunities for rich heirs? In any case, it looks like the verdict here is going to depend on instrumental calculations that are outside of my theory. Indeed, much here may hang on the supposed "innocence"

[27] On the other hand, see "Blame Italian Style," Wolf (2011) for a defense of this kind of blame.

[28] And this is true along several dimensions. One thing that should be clear at this point in the chapter is that fully working out the norms of who ought to constrain their blame (and when, and how much) would require a *much* more detailed working out of notion of social power I elaborated at the start. I've indicated at various points that this question is alive in the background of the chapter. Will, for instance, the middle-class (as a fuzzy group) have sufficient power that individual members of that group should refrain from blaming the very poor in many instances? These are difficult questions that deserve further careful treatment, but I leave them aside in this chapter.

of our metaphorical rich man. How likely is it that the billionaire is really blameless for society's ills? And, even if they *are* to blame for some of those ills, will the low-status appropriately constrain their blaming to the causes of those ills themselves, or be likely to condemn the billionaire more generally? Normative and ethical theories are going to guide us here as much as a theory of the epistemology of responsibility. My claim is merely that, given the lack of *power* possessed by the low-power in the first place, erring on the side of leniency is unobjectionable. The powerless should be free to form responsibility judgements and pursue their downstream effects precisely because: a) redistributing power (at least in non-violent ways) is not objectionable in these cases, and b) lacking power to begin with, such judgements are relatively unlikely to cause significant harm.

Second, we can, up to a point, engage in some bullet biting and say: instrumentalist justifications just price in certain kinds of errors. That is, the very point of the instrumental justification is that the practice is overall justified when set up in a certain way, while recognizing that there *will* be cases of error in the system. Combined with the first line of response, we'd say that holding the powerful accountable is unlikely to produce systematically bad results, even if it is occasionally done in error.

A related and equally thorny issue concerns *intra*-group blame. One common kind of claim is that those who are members of the same class or group of people are often their own harshest critics. Given this - shouldn't we recommend that those within social groups blame each other less often? First of all, it is hard to know exactly how true such claims are - indeed, sometimes they seem motivated by sexism or other forms of bigotry. Consider the idea, for instance, that women are "catty" and mean to one another by nature, a claim which is surely false (or at the very least involves a sexist reading of a complex social schema). However, there are two reasons to think that *sometimes* intra-group members really are their own harshest critics - but both reasons, I think, militate against including them in a norm that restrains blame.

The first reason is simply the idea that one knows one's own group best, and is thus often in the best position to criticise it. This is at least sometimes true. However, if this is the reason that group members hold each other accountable more often or more harshly, this would be so because such judgements are *accurate*. In such cases, then, members would *not* be in error, and so there would be no reason for them to restrain their blame *based on my view*.

The second reason is that these intra-group judgements of responsibility may involve a sublimation of the very norms of oppression that the powerful wield. In other words, group members may be quick to blame each other because of internalized misogyny, racism, or cases of adaptive preference. This *would* involve moral and epistemic error, and mean that there *is* a reason for restraint in these kinds of cases.

But, notice two things: first, it is unlikely that those who are acting out of internalized misogyny or adaptive preferences will notice what they are doing, or, insofar as they do, describe it in those terms. So a principle that asks them to change their behavior is unlikely to be effective (and this is putting aside other thorny issues about asking those with adaptive preferences to change them).

Second, the aim of Powerful Restraint is to change the normative landscape. If what we are witnessing is really a "trickle down" moral universe where the oppressed are taking up the norms of their oppressors, then changing the norms at the top will (eventually) put an end to it. I am not, of course, claiming that anything like this is easy. The idea is just that, as more social pressure is put on the powerful to restrain their blame, there may be opportunities for those who are internalizing oppression to come to see what is occurring to them *as* oppressors.

In either case, then, I argue that Powerful Restraint would be a more effective tool than trying to constrain intra-group blame. However, I think it's fair to say that these last few objections point us back towards the much more general upshot that a careful working out of the epistemology of responsibility raises: figuring out how to accurately form judgements about responsibility is hard – harder, at least, than we may have first believed.

All of these kinds of objections, I argue, generalize to a point about the coarse-grained nature of Powerful Restraint 1.0. We can head off some of these worries if we do two things: first, make the generality clause clearer and more forceful, and second, make the principle more obviously scalar and context sensitive, while retaining the spirit of the idea that those who are the most powerful ought to be blaming the least. Here are three sub-principles we can build in to accomplish this:

Confidence: The principle makes clear that blame is still warranted in edge-cases where an agent's confidence in the appropriateness of blame and the lack of likely harm are very high. But it makes equally clear that this condition is related to:

Scalarity: We ought to make clear that Powerful Restraint is not all or nothing. Given that we are all more or less powerful in various social contexts, the norm applies more or less strongly to us at various times. The more powerful we are, and the larger the gap we are blaming across, the more confident we can be in restraining our blame.

Generality of Power: We also want to preserve the idea that some agents are more entrenched in social power than others. The following chapter will deal more directly with defending an asymmetry based on the idea that the more centrally your various identities are tied to power, advantage, and privilege, the less often you should blame.

Given all of this, we can revise our principle to be more precise and fine-grained as follows:

> **Powerful Restraint 2.0:** In general, the socially powerful ought to refrain from blaming the less powerful, keeping in mind:
>
> **(1) Confidence**: Where one is very confident that blame is warranted and unlikely to lead to harm, blame can still be appropriate.
>
> But the epistemic bar to clear is higher given:

> **(2) Scalarity**: In particular instances, the more social power one wields in a given context, and the larger the gap in power is between blamer and blamee, the less acceptable relying *solely* on **Confidence** is, and the more one ought to restrain their blame, and:
>
> **(3) Generality of Power**: the more overall social power an agent has, the less acceptable relying *solely* on **Confidence** is, and the more one ought to restrain their blame.

3.6.2 The psychological and the structural: conclusions

There is one other objection to the arguments in this chapter that is worth a brief discussion – an objection that targets the project as a whole. In dealing with the individual psychological biases and perceptions of agents, am I not advocating for an ineffectual individualist approach to deep, systemic, structural problems? As I've noted at several points, many of the powerful are unlikely to change their judgements of responsibility, and asking them to do so appears to leave the task of social change only to the socially well off. I take it that this is a meaningfully (though subtly) different objection to that of the worry of disrespect or the likelihood of the socially powerless to also err.

Here it is crucial that we refocus on the metanorm of blame I've subtly advocated for: blaming the powerful who fail to refrain from blaming the powerless. Since, I've argued, those without corresponding social power should feel free to blame *up*, as it were, they ought to pay special attention to cases where the powerful continue to blame *down*. The idea here is not to create a society of petty moralizers tsk-tsk-ing the socially powerful's every move. The idea is to increase social pressure on those with social power - which may in turn change their attitudes and actions. It is absolutely imperative that my arguments be taken in the spirit in which they are intended: social change here flows from the bottom-up, not the top-down. The point of these practice revisions is to hold the powerful to account – indeed, to do it doubly. We ought to blame the powerful more than we do, for it is clear that in our present social structure they get away with too much. And, we ought to blame them more, *in particular*, for blaming those of low social status too much. Thus, a bottom up account seeks to balance the scales of an unjust system without relying on those at the top to do the heavy lifting.

However, I recognize that even this explanation will not appear sufficiently materially productive to some readers. I am sympathetic. I do not claim that revising our blaming practices will, by itself, right the social inequities and injustices that plague our world. Far from it! The point is that: a) if there are injustices within our responsibility practices that we can erase or minimize via norm change, we ought to do so, b) such change may be a necessary part of a broader attack on unjust structures. How likely is it that our practices can shift in the ways I'm imagining, and that such shifts would be beneficial to them? Much here depends on the precise instrumental aims we identify *and* very careful empirical and philosophical work that would vindicate the claim that holding the powerful accountable *is* likely to be overall beneficial in ways that our practices would justify. I think that such work is likely to vindicate my claim. If my claims about the epistemic distortions endemic to responsibility are true, a tendency to avoid blaming the most vulnerable among us is a small-scale and necessary revision in the process of moving us towards a more just society.

Portions of Chapter 3 have appeared in publication as “Blame for Me and Not for Thee,” in *Ethical Theory and Moral Practice* 25: 265-282, 2022. The dissertation/thesis author was the primary investigator and author of this paper.

Chapter 4

Cultivated Ignorance, Shifting Norms

For me, the epistemic desideratum is that the naturalizing and socializing of epistemology should have, as a component, the naturalizing and socializing of moral epistemology also and the study of pervasive social patterns of mistaken moral cognition. Thus the idea is that improvements in our cognitive practice should have a practical payoff in heightened sensitivity to social oppression and the attempt to reduce and ultimately eliminate that oppression.

Charles Mills[1]

Where there is an ascendant class, a large portion of the morality of the country emanates from its class interests, and its feelings of class superiority... and sympathies and antipathies which had little or nothing to do with the interests of society, have made themselves felt in the establishment of moralities with quite as great a force. The likings and dislikings of society, or of some powerful portion of it, are thus the main thing which has practically determined the rules laid down for general observance, under the penalties of law or opinion.

John Stuart Mill[2]

[1] Mills, 2017, 58

[2] Mill, 1909, 10–11

4.1 Introduction

At the end of the last chapter, I introduced the following norm as one potential revision to our responsibility practices:

> **Powerful Restraint 2.0:** In general, the socially powerful ought to refrain from blaming the less powerful, keeping in mind:
>
> **(1) Confidence**: Where one is very confident that blame is warranted and unlikely to lead to harm, blame can still be appropriate.
>
> But the epistemic bar to clear is higher given:
>
> **(2) Scalarity**: In particular instances, the more social power one wields in a given context, and the larger the gap in power is between blamer and blamee, the less acceptable relying *solely* on **Confidence** is, and the more one ought to restrain their blame, and:
>
> **(3) Generality of Power**: the more overall social power an agent has, the less acceptable relying *solely* on **Confidence** is, and the more one ought to restrain their blame.

I claimed that such a revision to our practices of responsibility is justified, given all that I've argued in the previous chapters. If my claims about the epistemic disruptions endemic to responsibility ascriptions are true, inculcating a tendency in the powerful to avoid blaming the most vulnerable among us is a small-scale and necessary revision – one that seems both psychologically possible and expedient in moving towards a more just society.

In this final chapter, I have three tasks. First, I want to further bolster my argument that, given the necessity of revisionism, an asymmetry in our blaming practices is justified. Recall that, the asymmetry includes the idea that we ought to blame the powerful *more.* Even if one went along with my claim that we ought to blame the socially dispossessed less often, the other half of the asymmetry may be a hard pill to swallow. Yet, I see holding the powerful to account more often as a key shift in our normative architecture. So, more needs to be said to justify it. I've argued that those in more powerful social positions sometimes maintain "active" ignorance of the effects of their blame (and indeed, of their own blameworthiness). How this relates to the epistemic disruption I've described is a complex matter, and the second task of the chapter is to further explore that relationship. I look at some of the practical issues that arise once we notice the relationship between various kinds of ignorance and the sort of epistemic disruption I've been describing. Discussing these issues puts us back into contact with what is often called the "epistemic condition" in the moral responsibility literature. We will see how the tension between holding the powerful to account and their ability to cultivate ignorance produces some of the architecture of our current practices. The third task of the chapter, then, is to explore how these issues of ignorance, power, and blame connect to recent discussions of pragmatic encroachment in the epistemology literature. I'll begin with pragmatic encroachment in order to motivate the importance of the rest of the claims in the chapter.

4.2 High Stakes Blame and Encroachment

Reasons to be cautious about blame stem from two sources, broadly speaking. One is epistemic: we can be unsure that we have sufficient evidence of blameworthiness, or suspect that we might be making a mistake in our reasoning, or peer disagreement might cause us to re-assess our blame. The other is pragmatic (and sometimes moral): we can worry about the harm that our blame might cause, or fear that blaming will damage a relationship we value, for instance. Although the separateness of these sources is clear in the abstract, in practice, the two kinds of reasons tend to blend together. Where blame flows from high to low power, I've suggested, we might require more caution, both because we worry that our evidence might be tainted by various biases, and because we worry about the higher likelihood of harm. Indeed, the sub-principles of Scalarity and Generality of Power suggested that the evidential burden we'd need to blame can shift without us being any less sure of the evidence itself, and without us revising any general principles we might have about the ethics of blame.

This interaction of moral and epistemic factors can be understand by analogy to (or perhaps as a species of) pragmatic encroachment. In its most basic presentation, pragmatic encroachment is just the idea that non-epistemic factors can influence the epistemic statuses of our beliefs, knowledge, justifications, and so on. There are many different definitions of pragmatic encroachment, and indeed, many different theories of it. I'll have more to say on these matters below, but for now, if we need more precision, we can work with this definition from Nolfi (2018):

> **Pragmatic Encroachment (PE)**: Some of the considerations that help to determine the epistemic status of S's belief that p do so without being truth-relevant (i.e. without affecting the subjective or objective likelihood of p). (36)

In other words, pragmatic considerations can change whether we are objectively warranted in believing what we do, *or* whether we are subjectively confident in our beliefs – and these changes can occur despite no change to our evidence.

Here's an example, similar to those introduced by Fantel and McGrath (2002): you need to take the bus to meet a friend. You have a reason to make sure you've got your transit card in your wallet. But the stakes are pretty low, let's say. You could walk back home and get it if you forgot, and you'd only be around 20 minutes late. In this context giving your pocket a quick pat to check that you've got your wallet as you leave your house might be good enough, in terms of evidence collection for the situation. But say that what you need to catch the bus to do is to pick up your young child from school. If you're late, they will be standing in the cold, alone, confused and probably a bit frightened. Now it might seem that patting your pocket isn't good enough – you should really double-check that the card is in your wallet. What's changed here are the pragmatic stakes of the situation. Making a friend wait for a bit while you catch the next bus isn't such a big deal. Leaving your child alone (although not the end of the world) is a much bigger one. Notice that both what you ought to *do* and how certain you ought to be about your *beliefs* shift with the weight of pragmatic considerations.

Pragmatic Encroachment is controversial, and there are costs to accepting it. It seems, for instance, counter-intuitive that whether I believe P should depend on the practical or moral context I'm in, rather than how good or bad my evidence for P is. Note, however, that it neatly explains the bus example above, especially when we think about what a belief in P licenses us to *do*. Whether I pat my pocket or check my wallet seems like it really ought to have some relation to *why* I'm doing those things. And just so, whether I think my blame is likely to have significant costs seems like it should matter (at least subjectively) in my determination of whether to move from a judgement of blameworthiness to the expression of blame. Basu (2021) uses these intuitions as a springboard to explain how the *weight* of the very same evidence can shift in various contexts. Roughly, in "low-stakes" situations, a piece of evidence might be enough for us to form a belief or a plan of action, while in a "high-stakes" situation, it may be insufficient.[3] This thesis of pragmatic encroachment can be helpfully applied in the ethics of blame to better motivate the stakes of some of my arguments above. In particular, it helps further motivate the principle of Powerful Restraint and the corresponding asymmetric claim that those with low social power needn't be as cautious when blaming upward. This is so, in part, because the stakes of blaming are not constant: they are positional and contextual, and this matters for thinking about the acceptability of blame in a way that does not solely depend on evidence of blameworthiness.

Most importantly, attending to work in the pragmatic encroachment literature can help explain why doubling and tripling down on evidence collection in individual cases isn't our best route out of the the bad epistemic situations the powerful often

[3] The language of high and low stakes draws on Stanley (2005), who argues that the difficulty of knowing things rises as the stakes increase.

find themselves in. Accepting Pragmatic Encroachment as a thesis isn't necessary to understand or accept my other arguments, but if one is sympathetic to it, they will be more easily persuaded of some of the particular dangers I canvass in the sections on Epistemologies of Ignorance below. If one accepts something like Pragmatic Encroachment, then it is easy to see why the high-stakes nature of blame involved in blaming from a sufficiently high degree of power calls for the asymmetric revisions to our responsibility practice norms I argue for.

What can be said, then, to further motivate the thesis? Imagine the following scenario: as a teenager, your little brother steals your diary and reads it – cover-to-cover. You catch him red-handed trying to return the journal to it's secret hiding place in your dresser, and he confesses everything. It seems you have all the evidence you'd ever need that your sibling has done something blameworthy.[4] Should you blame him? Above I canvassed some the further questions it might make sense to ask in this kind of case. Here's a non-exhaustive array: Will the blame be effective? Perhaps blaming him will have little or no effect on his future actions – little brothers are notoriously recalcitrant. What about your standing? Perhaps you recently stole *his* diary, for instance. What about his capacities and quality of will? Perhaps it is rarely acceptable to blame children given their not-fully-developed capacities for moral reasoning, or perhaps he had no ill-will and didn't really know how much his actions were a violation of your privacy. The list could continue. But imagine that everything in this kind of case is "normal." That is, there aren't likely to be out-sized effects that stem from your blaming, it's appropriate to blame children in certain ways, he was at least somewhat aware of the wrongness of his action, and so on. What else would we need to know to say that this is a case where blame (of some kind) is appropriate? Very little, it would seem.

One reason to suspect that this intuition is correct is that, as I've described it,

[4] We are simply stipulating this here - questions of blameworthiness and instrumental reasons to blame can come apart, as I will discuss below. But we are assuming that, in this example, your brother really is blameworthy.

this case seems to be “low-stakes” in Stanley (2005) and Basu and Schroeder (2018)’s parlance. That is, despite the real violation, and the intense, teenage emotions you might be feeling, not much hinges on your blaming your brother. Such family dynamics are common, and, unless things go very wrong, your blame - even if it is passionate and hot-headed - is unlikely to lead to a permanent rupture of your relationship. Indeed, following Wolf (2011) it may even be a very *healthy* part of being in the kind of relationship with your brother that you ought to have. A cycle of blame and contrition may be part of what constitutes your emotional bond. This kind of inter-familial drama is, after all, part of what makes the ties of family so special and strong.

But imagine a nearby variant of the case. The facts of blameworthiness remain the same: your brother stole the diary, he knew what he was doing, and he knew it was wrong. But in this case, somehow, you also have the absolutely certain knowledge that blaming your brother will send him over the edge and towards a psychotic break. In this world of perfect knowledge we can stipulate that you know that the feelings of shame and guilt your blame will engender in him will cause him to spiral for several years, require him to be institutionalized for a time, and leave lasting and deep emotional scars. The case here is fanciful, but instructive. Should you still blame your brother in this situation? Unless you have some incredibly strong views about retributivism, desert, and the role that blame must play in a well-ordered moral universe, I hope that your answer is: no. Although blame is largely a backward-looking practice, it’s partial orientation towards future states of affairs is largely non-controversial. Much of the focus of this chapter will be on those forward-looking effects, and the reasons we ought to take them seriously. So, if your blame will permanently mar your relationship with your brother, cause him intense pain, and lead to lasting psychological issues, it seems right that you ought to balance the benefits of blaming him for stealing your journal against these seriously weighty outcomes.

Hopefully all of this is relatively uncontroversial. Finally, imagine a third variant of the case, closer to the real-world. You don't have perfect knowledge: you are back in the realm of assessing evidence and interpreting intention and action. Say that one piece of evidence you have is that your brother is struggling with his mental health, such that you suspect he is close to having a mental health crisis. And, perhaps you have good evidence that he is particularly sensitive to the high-arousal state that being blamed brings about. We can imagine that you have recently observed him having intensely negative reactions to feelings of guilt and shame. Should you blame him in this variant? Here, although the answer is less clear, I think we feel the pull of some of the same countervailing reasons from case two.

Notice that, in all three cases, your evidence about *blameworthiness* is fixed: your brother has confessed to stealing and reading the diary. So, what changes? The evidence about what your blame will do shift, which changes the *stakes* of your blame. In turn, these changes shift the acceptability of what you ought to do, all things considered. There are two dimensions that matter here in terms of analogizing this situation to that of pragmatic encroachment: what the stakes are, and how certain you are about what your actions will lead to. Given this, we can see two ways in which pragmatic considerations are going to encroach on your beliefs about the acceptability of blaming:

> 1) By shifting the amount of evidence one would need to have in order to feel confident about a particular path of action, and
>
> 2) by introducing a general note of caution or suspension of belief and action into one's epistemic and agential ecology - a note of caution indexed to how high the stakes are.

I've been arguing for just such a suspension of blaming, and now I can clarify why thinking of the stakes as relating to the relevant pragmatic considerations matter for cases of blame. From an individual perspective, it might seem important to collect more and better evidence in order to feel confident. But there is a problem here.

Remember that, in the examples above, the evidence you have of your brother's blameworthiness is fixed. There simply isn't more evidence to collect. In addition to this, focusing too much on evidence collection obscures the nature of the deeper problems I considered in the first two chapters of this dissertation. Very roughly: there are cases where no amount of evidence collection will do the trick - given that we don't have access to the kind of evidence we'd need to vindicate our beliefs.

4.2.1 Epistemic Humility in Risky Blame Cases

The pragmatic encroachment literature gives us a lens into understanding the claim that epistemic humility and the restriction of blame might be required of the socially powerful in their interactions with the socially dispossessed. Let's refocus for a moment on a very general definition of knowledge that is amenable to pragmatic encroachment. Kim and McGrath (2018) helpfully gives us: "S knows that p only if S is justified in taking P for granted in deliberation." I've argued that this connection between the epistemic and the realm of action is crucial for understanding how pragmatic encroachment affects blame. Almost nowhere in this dissertation have I attempted to give a tight definition of "knowledge." This is simply, to some extent, in order to avoid hugely complex fights in the epistemology literature. But on a more basic level it is because I don't think the question of how to distinguish outright knowledge from something like "very high confidence in a belief" is important for the points I'm making. That is, the powerful don't have a reason to restrain their blaming tendencies *iff* they fail to *know* that someone is properly blameworthy because of the stakes of a given situation. If this were the case, properly specifying the threshold of knowledge would be an important part of my project. But the point here is that the powerful have a compelling moral reason to avoid blame – a moral reason that encroaches on their epistemic processes such that they ought to withhold from *acting* in certain ways whatever their relationship to an account of knowledge looks like.

We've seen that "knowing that p" is going to depend on the specific context and situation of agents because those contexts and situations are going to help fix specific epistemic standards or norms we must meet to count as knowing. But more importantly, we've seen that the stakes also fix the standards of what our knowledge licenses us to *do*. Notice that we can substantially weaken the general pragmatic encroachment thesis if we focus on the connection between confidence and action, and omit talk of knowledge (and to some extent, belief). All we need to say is something like the following:

> **PE Action**: S ought to act on reason P only if S is justified in taking P for granted in deliberation.

Some readers will balk here at reasons talk, and insist that things have been made unhelpfully and problematically murkier with this reformulation. But I don't think this is right. All I'm after is the standard sense of a normative reason introduced by Scanlon (1998), where having a reason to Φ is a consideration that counts in favor of Φ-ing. Why does this help? Because we can mostly sidestep questions of the epistemic status of beliefs, and avoid all-together questions of knowledge. The relevant question is not whether an agent *really* knows (whatever that relation amounts to) that a subject is blameworthy given the stakes of blame, the question is whether they ought to take their belief in the agent's blameworthiness as straightforwardly granting them license to do something: in this case, to blame them. In other words, what can't be taken for granted in deliberation is that blameworthiness gives an agent license to blame. And this is unobjectionable – it is something almost everyone in the blame literature will agree to. What attending to the epistemic realm and issues of pragmatic encroachment allows us to see is that this non-licensure is often the case for the interestingly systemic reasons I canvassed in the first half of the dissertation. The patterns of the stakes involved in blaming across large gaps in power, and the likelihood of error introduced by those gaps means that the powerful are given reasons not to blame due to: 1) general epistemic caution given what they know about their own epistemic situation, and 2) the relevance of the high stakes.

Here I follow Cohen (2018) in arguing that pragmatic encroachment can help explain the unintuitive result that we can both know something *and* fail to possess it as a reason for action. This is less counter-intuitive than it seems, simply because there is a difference in what our total evidence licenses us to believe and what that belief (combined with our total evidence, including the evidence about our evidence) licenses us to *do*. Here's how Cohen sketches the issue: It looks like you can know something R, "where R is a reason to Φ, but fail to possess R as a reason to Φ" (104). How? Because of the difference, "between your basic evidence —- your evidence that is not derived from other evidence — and your non-basic evidence." Possessing R is a matter of what evidence you have for it. So, perhaps you have strong evidence that you have a reason to go to the doctor tomorrow. But this doesn't fix whether you ought to actually go to the doctor. Perhaps, for example, you come to doubt all of your evidence, globally, having fallen into a kind of Cartesian skepticism. If this is so, you still possess reason R, but you may not possess R *as* a reason to go to the doctor.

In our case, you might possess a reason to blame (having good evidence of blameworthiness), but fail to possess it *as* a reason to blame once you note the stakes of the situation. To see why, let's consider an analogy. Imagine a property "Coolness" (**C**). Let's stipulate that what (metaphysically) matters for an agent to have property **C** is that they have the further properties: "Never Sweats It" (**NSI**) and "*Je Ne Sais Quoi*" (**JNSQ**). Finally, let's stipulate that **NSI** and **JNSQ** involve clusters of mental relations, properties and states that are accessible only from an agent's internal point of view (and even then, not always, and not accurately) – things like, "not caring what people think," "being genuinely unbothered in risky social situations," "having a knack for understanding what's lame and avoiding it," "never trying too hard," and so on. Again, we can just stipulate that everyone agrees that these features are what it takes to be properly counted as having **C**.

What kind of epistemic access do we have to these things? It looks like we have second-hand, unreliable, mediated access at best. We won't directly perceive, for instance, JNSQ. Instead, we will have evidence by the way someone acts that, perhaps, their seemingly being unbothered is grounded in JNSQ. The point is *not* that this mediation poses a problem for having a concept of Coolness, or for ever identifying who counts as Cool. It may be that we have reliable social practices that pick out Cool people, and intuitive senses of who counts and why. The point is that, however those practices work, they aren't working by directly accessing the metaphysical factors that matter for grounding Coolness. They are going to be social, interpretive, and conversationally built up. Returning to the point: here is a case where, were the stakes to shift, demanding that we simply collect more (or better) evidence is not going to be more helpful. If the kind of evidence of **C** I can have runs through **NSI** and **JNSQ** which run through further clusters of mental properties, the best I can do is get an overall sense of whether you *seem* to exhibit them. Once I've done that, there's just no more evidence to collect.

In low-stakes settings, we feel confident proceeding in our actions despite this. This is partially so because we feel confident that the general *practices* we are enmeshed in are legitimate and worthwhile. So, maybe deciding who is cool is just an important part of human social life, and, even though we can't ever be fully confident we are getting it right, it seems worth engaging in treating some people as cool and others as not-cool anwyway. In low-stakes settings we might feel confident declaring things **C** on rather flimsy evidence because we think that coolness, in general, is part of a worthwhile and legitimate set of social practices. Given that we find the practices themselves valuable, and that, seen from a general vantage point, the practice operates (roughly) in the way we'd like them to, we needn't worry too much about our first-order evidence.

The flip-side of this issue is that, where the stakes are high, there may be more *systemic* types of problems that demand a different kind of response than better, more thorough, or more careful evidence collection. Imagine that the stakes of being declared to have (or lack) property **C** are very high in 8th grade. If one has **C** one will get romantic partners, one will be invited to parties, one will be able to start a band, or hang out with 9th graders, or so on. If one lacks **C** one loses access to these important social goods, and one's well-being is thereby diminished. How should we respond to this, when we are deciding whether to declare that someone does or doesn't have **C**? Not, I hope it is clear, by doubling down on our investigations into their mental properties. Instead, we might want to start thinking about the likely effects of our declarations, whether it seems *worth it* to declare a given agent **C** or not in a given case, and, most importantly, whether *we* are likely to be making certain kinds of errors. Perhaps, looked at from a certain vantage point, we notice that our attributions of **C** follow certain patterns that break down along ethnic or gendered lines, or are sensitive to features we don't, on reflection, take to matter (features like weight, or wealth, for example). We might even, in the end, want to be more generally cautious about our use of the concept, or suspend belief about borderline cases. Maybe the high-stakes nature of these cases even pushes us to question whether we like our Coolness practices after all, or whether they *are* worthwhile and socially legitimate.

Noting the distinction between general caution and more specific demands for evidence then, we might begin to ask whether there are *practice level* reforms we could make that would allow us to be: a) more generally reliable in our judgments, or b) less harmful where we are likely to get things wrong. Given all of this, let's get back to blame. I'm claiming that blame itself can sometimes be a high-stakes situation – and in less fanciful ways than the examples above signal. What do I mean by claiming that blame is high-stakes? One place to begin is with blame's tight connection to punishment. If blame often leads to punishment, by the state or by individual actors, or if blame just *is* a kind of punishment, then the stakes will be high whenever the punishment is likely to be serious. Of course, although we ought to take this connection seriously, I've been clear in this dissertation that blame and punishment are conceptually distinct and shouldn't be run together. It's hard to imagine, for instance, how privately blaming leads to (or counts as) punishment. And, most instances of blame will certainly not lead to *state* sanctioned punishment involving fines, imprisonment, or other penalties. Finally, on models of blame that treat blaming as a kind of conversational opening that allows for the restoration of relationships, it is infelicitous to think of blaming as involving punishment at all. Still, there are many recognizable cases where the penalty blame imposes, directly or as a follow-on, is quite high.

What, then, are the takeaways of this distinction between evidence collection and systemic approaches to correcting for practice-wide patterns? First, it *is* the case that the evidential burden we ought to shoulder when we blame may be greater than we often recognize. And this is so because blame is generally treated as a relatively low-stakes social interaction. One that, to be sure, can be inter-personally costly, or in rare cases, go badly wrong - but, overall, an anodyne (if important) part of our day-to-day moral lives. Indeed, I want to press on the reader that much of our blaming practice *is* like this. But it's also the case that blame can be more high-stakes than we generally recognize – and high-stakes in a way that calls for more care in our evidence collection. Where the stakes are unusually high, and where my real interest lies, are in cases of blame between differentially socially positioned moral agents, and especially where blaming is likely to lead to costly social sanction or punishment. Recall Basu's basic point:

> The moral considerations raise the epistemic standards in these high-stakes cases. We can preserve the thought that the facts don't care about your (or other people's) feelings while also recognizing that whether or not you are justified in believing on the basis of the evidence available to you is a question that is sensitive to non-factual or non-evidential considerations. Whether you have enough evidence to believe varies according to the stakes. (205)

So, we are given a reason why the pragmatic can encroach in an unobjectionable way: when the stakes are high, moral considerations ought to be (normatively) relevant to what we believe. But we've also argued for two other points. First, the real point of interest is in investigating the structural conditions that surround blame and our general epistemology of moral responsibility. By treating blaming practices as more-or-less unified and typically involving relatively similarly positioned moral agents, we are robbed of the ability to see the stakes of our blaming practices *themselves* clearly. Once we recognize that power-dynamics and social position make the stakes of our blame much higher or lower in given cases, and that these kinds of cases can follow recognizable patterns, we should also recognize that some *patterns* of blame will involve very high stakes, and correcting these may not be a matter of individual evidence collection.

Second, I've noted that it's really the expression phase of blame where the rubber meets the road here. We can have the same exact evidence for an agent's blameworthiness, and what raising the stakes asks us to consider is what we ought to *do* given the evidence we have. Epistemologists who cache out beliefs *in terms* of dispositions to act will have little trouble accepting this kind of claim, of course, but this is itself a controversial position. All we need to notice, for our purposes, is that there is at least a very strong dispositional relationship between belief and action, in general. On basic belief-desire accounts of action, the strength of my belief (or credence) in something (combined with the right kinds of desires), ought to increase the likelihood of my acting on that belief. So, if I desire a drink of water, and I am pretty sure the cup in front of me is full of water, I ought to be more likely to take a drink of it than if I'm pretty sure the cup in front of me is empty, or full of poison. This very basic point matters because one thing an encroachment view can show us is that this uncontroversial account of dispositions for action shifting based on changes in evidence also applies to shifts in the stakes of a situation (where our evidence stays the same).

This is why it's important to focus on the blame that the highly socially powerful engage in. Here we have a (somewhat) easily identifiable group with an outsize effect, and we can propose group-level norms that don't rely on the highly specific intricacies of individual cases. Rather than asking the powerful to do better, more thorough, and more complicated work in each individual instance, we can simply put normative pressure on them to do *less* by blaming less. Thus, the proposal both avoids thorny issues about how individual agents ought to do risk-assessment and evidence collection in given cases and is more likely to succeed. Instead of expecting the powerful to *care more* about what their blame does, we ask them to avoid blaming in order to avoid social sanction. We needn't pretend that this normative shift is going to be perfectly effective in order to argue that it is likely to be more effective than the alternative. With the high-stakes nature of blame in hand, I turn now to addressing several lingering objetions from the previous chapter.

4.3 Asymmetries in moral responsibility practices: further justification

One reason that an asymmetry in blame may appear problematic, to re-state a previous objection, is that seems unlikely that the powerless are in any better of an epistemic position than the powerful when it comes to ascertaining the basic metaphysical features which ground responsibility. Indeed, one would have to accept a very strong (and I think mistaken) kind of standpoint epistemology to think that existing social hierarchies would make the less powerful better able to make judgements of responsibility in general. Such a view is what Wylie (2003) calls an "automatic privelege" view: the view that particular epistemic vantage points automatically grant us effortlessly privileged viewpoints on the world. But this kind of extreme position isn't what's plausibly needed for my view. Following Tilton (2023), we could modulate the claim and instead say that the less powerful are at a "strong epistemic advantage." This kind of thesis says that those in less privileged positions are at an epistemic advantage when it comes to knowing certain things. But even this "strong" thesis is more than we need. A more moderate approach might suggest only that the less powerful are *sometimes* in a better position to notice features of responsibility that their more powerful counterparts will sometimes miss, and that this is so in virtue of their social position. Weakened to this extent, such a thesis can begin to look trivially true, and basically uninformative. If it is right, then the less powerful will sometimes be in a better position to blame, and the more powerful a worse position – but can it avoid triviality? I'll argue that it can, once we understand its connection to the issues of epistemic disruption I've been working through. I will call this line of argumentation the Standpoint Epistemic Justification for asymmetrical blame, and begin to work through it below. A second line of argument, suggested briefly in the last chapter, is that even if there are many individual cases where the powerless are no epistemically better off, there may still be instrumental, forward-looking reasons to encourage them to blame upward more frequently. This second line of argumentation I will call the Social Signalling justification. Taken together, my claim is that the Standpoint Epistemic and Social Signalling justifications vindicate the overall asymmetry.

4.3.1 The Standpoint Epistemic Justification of Asymmetry

To get a sense of the Standpoint Epistemic justification in action, one can look at the recent work of Ciurria (2020a, 2020b). There they argue that blame and praise can be "emancipatory" narratives – sites of resistance to oppression which fight against the "disappearance" of powerful perpetrators from the public view. Let's focus on the idea of "disappearance." Ciurria's claim is that the agency of the powerful (and sometimes the powerful themselves) often disappear from social narratives of and about their actions.

Take, as an example, the way in which the news media often parrots the language of police public relations departments in reporting on "officer involved" shootings. The use of the passive voice, neutral descriptions of action, and lack of clear causal explanations serve to mitigate our reactive attitudes and, in general, obscure what actually took place. Reports and headlines such as the following are common: "The suspect then ran towards the officers still armed with the sword and an officer-involved-shooting occurred," or "Officer-involved shooting: Butler Twp. chief says woman killed was known to police,"[5] or, consider: "Baby boy killed during attempted arrest in Mississippi, police say."[6]

Here's another striking example journalist Radley Balko writes about in the Washington Post:

> While responding to reports of a stabbing, LASD deputies shot and killed 30-year-old John Winkler. In an initial press release, the department said Winkler "aggressed the deputies and a deputy-involved shooting occurred." Note that Winkler's actions were put in the active voice, while the officers' actions were put in the passive.
>
> As it turns out, Winkler was innocent. He hadn't "aggressed" the officers at all. Rather, he and another victim, both of whom had been stabbed, were running toward the police to escape their assailant. (The deputies shot the other victim, too.) The press release incorrectly assigned criminal culpability to an innocent stabbing victim, but carefully avoided prematurely assigning responsibility to the deputies who shot him.

The epistemic result of cases like these are that our judgements of who was involved, how they were involved, and who is responsible are routed in ways that are full of (by now familiar) distortions. The actions of police officers are presented as

[5] For these examples and many others see: https://www.washingtonpost.com/news/the-watch/wp/2014/07/14/the-curious-grammar-of-police-shootings/ and https://www.cjr.org/analysis/officer-involved-shooting.php

[6] https://www.nbcnews.com/news/us-news/baby-boy-killed-during-deadly-arrest-attempt-mississippi-police-say-n1266320?cid=sm_npd_nn_tw_ma

neutral happenings in the world, and their agency "disappears," in Ciurria's words. The actions of the victims of the shootings, on the other hand, are presented in ways that maginfy their agency, minimize their innocence, make them seem untrustworthy, and fire up cycles of negative affect. The impact of this has been explored in Chapters II and III of this dissertation. But there are two further results of this kind of language we should notice. First, that "disappearing" agency and responsibility in this way is a powerful social tool for manufacturing consent, and second, that it is more likely to be noticed (and thus resisted) by those who are in positions to notice (or care to notice) that something isn't right with the "official" narrative. Let me consider each of these points in turn.

First, a status quo of social power is, in part, upheld by the this kind of uncritical parroting of powerful voices and systems. Police unions and administrations, for instance, have a nearly unbelievable amount of power in contemporary civic life. Their budgets are often by far the largest source of major city's spending, and such is their stranglehold over politicians that they rarely, if ever, receive meaningful push-back from the Mayors and city councils who are, ostensibly, their bosses.[7]

Not only this, but, as I've been exploring, this kind of power with very limited checks can create feedback loops. In this case, we get a loop where much of the general public is exceedingly deferential towards police officers, unions, and administrators. One concrete outcome of this is that information about police activity is tightly controlled and dispersed. For the average member of the public, there is no way to verify what police officers are up to. The press, supposedly, mitigates this epistemic lacuna, but since the press often simply repeats the talking points of police reports, briefings, and bureaucratic mouthpieces as plain facts, an independent check on police activity often remains elusive. The point here (for our purposes) isn't just

[7] See, for instance, this breakdown of the Los Angeles city budget: https://budget.lacontroller.io/ - the Police Department accounts for over twenty-five percent of spending, more than double the next largest expenditure. For the problems of enforcing meaningufl political oversight, see Vitale (2018)

that this is one way police departments become and remain corrupt, but rather that it is a way in which a certain epistemic situation is constructed by using elements of social power. The police have the material and reputational power to make their point of view definitive, and so, for many members of the public, it is definitive.

One argument I've made is that our responsibility practices, overall, often function in an analogous way. That is, we have no real way of verifying, in the vast majority of responsibility cases, the competencies and situational factors involved. We must either take the word of an "official" mouthpiece, or form our own interpretation. Where power and privilege are involved, I've argued, we are *more likely* to accept the "official" account, *and* for there to be epistemic ignorance and a lack of awareness that this is what is happening.

The very mild version of a standpoint epistemic thesis we are working with posits that some agents are more likely than others to notice, understand, or be in a position to explain various aspects of the world because of their social position. All one needs to accept in order to find this claim compelling is some very mild form of epistemic contextualism. Indeed, I will explore below the idea that this isn't even a 'standpoint' theory, since the agents in question haven't done anything intentional to achieve a certain way of seeing or knowing about the world. What we know (or have a chance to know) depends, at least in part, on where we stand in the web of knowledge. This kind of claim, in other words, doesn't depend on knowledge (or truth) being relative - it just acknowledges that knowing (and our access to truth) is socially mediated. What I have the opportunity to know depends on who, where, and when I am.

Ever so slightly more controversially, one could also add in that these things depend (again, in part) on *why* and *how* one aims to know them. This needn't be a robust claim about pragmatic encroachment, although my arguments above show that this is also a live possibility. All I mean is that epistemology, being normative, is dependent on local norms and agential values. So, what I seek to find out, and how satisfied I am with my processes for doing so, will also be relative to a social context or standpoint. Again, a mild form of this claim seems unobjectionable. It simply states that I am more likely to gain knowledge about things I am motivated to look into, and less likely to gain knowledge about things I am unmotivated (or disincentivized) to care about.

These unobjectionable claims are all one needs to conclude that those with less social power are sometimes in a position to be better able to accurately notice the transgressions of the more powerful. Stick with our example of the way in which police public relations departments often "disappear" the agency of officers involved in shootings. If an agent's situatedness as a knower is such that they have little contact with the police, a general feeling that police officers are helpful and protective, and no real motivation to dig deeper into the facts of a case, then it's no surprise that they would take police talking points at their word. Clearly, however, there are times where forming beliefs in this way will be reliably inaccurate: hence, the existence of PR firms. If, on the other hand, one is a member of a group often targeted for police interaction (much less abuse), or who is intimately familiar with patterns of police violence and repression, or, if one is a member of the affected community motivated to learn more about the details of a particular case, one is more likely to (accurately) notice lies of omission, or distortions of the truth in police statements.

Finally, we ought to note that the reverse of this is often true for blame running in the direction of high to low power. It is a classic point in the philosophy of race, for instance, that while racialized minorities are forced to learn about the dominant white culture as a price of social acceptance, this is usually a one way street. That is, white people in the United States may have few incentives to learn about minority cultures – including their norms, practices, languages, and forms of self-expression.[8] These are not hard and fast rules, of course, but it does make it more likely that a materially powerful, socially high-status white person is in a poor position to assess the actions of a low-status, materially disempowered member of a

[8] For defenses of this kind of claim, see McIntosh (1988), Mills (1997), and Yancy (2012).

minority group. These arguments add up to the Standpoint Epistemic Justification of blame's asymmetry. Consent manufacturing (through disappearance and other tools) and situated or contextual epistemic pressures mean that the less powerful often are in a (relatively) better position to judge the powerful blameworthy than the other way around. I've begun to sketch a picture of the ways in which one's situatedness can lead to a greater chance that one will gain true beliefs in some contexts, and remain ignorant in others. In order to vindicate this picture and the Standpoint Epistemic Justification, it's time to deal with ignorance more directly.

4.3.2 Epistemologies of Ignorance

For the last several decades, philosophers like Charles Mills, Linda Martin Alcoff and Kristie Dotson have argued that we ought to pay more theoretical attention to "epistemologies of ignorance." There are several ways to work out what an epistemology of ignorance might amount to, but the basic idea is that ignorance is not always (merely) the result of a misapplication of positive epistemic norms. One might have thought, for example, that to be ignorant about something is merely to lack knowledge about that thing. Instead, these authors have argued, ignorance is sometimes *constructed* or positively produced and reproduced. Of course, as all of these authors are quick to point out, this is not exactly a new idea. It is, in fact, a central tenant of critical theory (especially critical theory drawing on the Marxist tradition) that much of social life is captured by various kinds of "ideology." But, as Mills (2017) points out in the introduction of "White Ignorance," "The concepts of domination, hegemony, ideology, mystification, exploitation, and so on that are part of the *lingua franca* of radicals find little or no place [in mainstream analytic philosophy]" (51).

With the slow (and incomplete) acceptance of standpoint epistemology and feminist philosophy into the mainstream, however, projects that take seriously the idea that idealized, individualistic epistemology might radically misrepresent the state of affairs on the ground are more common. What might the "construction" and reproduction of ignorance look like? First, it can occur at an individual level: for example, to protect oneself from harmful self-knowledge that would lead to guilt and shame, one might work to remain ignorant of certain guilt-inducing facts. Or, it might occur at a communal level: for instance, when an identity group relies on not knowing about certain topics in order to make sense of key beliefs about themselves or others. Finally, it might occur at a structural level: for example, when power structures reproduce themselves by (in part) making knowledge about how to dismantle or oppose them verboten.

All of these kinds of constructed ignorances are compatible with one another, and often work in lockstep. Mills argues, for instance, that white people have communal, positive reasons to misconstrue the world and their place in it. These reasons emerge at or from a structural framework, but they have individual expression and value. A white person can avoid guilt or shame about their privilege, feel positively about their white group identity, and reinforce structures of oppression simply by declining to learn about, for instance, the ways in which black people are systemically oppressed. The basic fact that Mills challenges us to acknowledge is that widespread ignorance at these all of these levels is endemic in our current cultural context. As he argues, in mainstream analytic epistemology, it is often the case that:

> US political culture is conceptualized as essentially egalitarian and inclusive, with the long actual history of systemic gender and racial subordination being relegated to the status of a minor "deviation" from the norm. Obviously, such a starting point crucially handicaps any realistic social epistemology since in effect it turns things upside-down. Sexism and racism, patriarchy and white supremacy, have not been the exception but the norm. (53-54)

If, as I agree, sexism, racism, patriarchy and white supremacy have been endemic both in our social and political cultures and in our practices of philosophy, how should we confront this fact when we theorize the social epistemology of moral responsibility and its corresponding ethics of blame? In two ways. First, by correctly pricing in this kind of ignorance as background to theory of how moral responsibility ascriptions and blame operate in our actual practices, and, second, by theorizing ways we can counteract and overcome these ignorances. I've been arguing that carefully attending to the epistemology of moral responsibility can be valuable to theorists of responsibility in various ways. One of the ways I identified at the outset was that it can help us identify and work to overcome epistemic injustices. I've also identified some interesting kinds of ignorance embedded in our responsibility practices: those in power have reason to avoid learning and knowing about the less powerful as full members of their moral community. So, it is also worth asking the question: do our responsibility practices, as currently constituted, help support an epistemology of ignorance?

The short answer is: yes. In the last chapter I presented two claims that might seem to be in some tension with one another. One was that power is highly contextual. We will all, in some contexts, be powerful blamers, and in other contexts be powerless blamees. The other was that power gaps are reliable enough that we can track likely epistemic disruptions across relevant cases, group-interactions, and social strata. I suggested that one way to reconcile these two claims it that although social *status* is highly contextual and mutable, material social power is much more rigid and concrete. So, although the materially impoverished may, on occasion, have a higher social status than the wealthy or well-connected, it is unlikely that they, in general, have more overall social *power*.[9] This lets us hang on to the important

[9] One thing to note is that most blame occurs locally - that is within social dynamics and groups where levels of power are probably roughly similar. However, there are two ways one can think about this. The first is that the locality of blame means that the arguments I'm making here have less application than you might initially think. If most blame is occurring on roughly equal footing - then worries about blame from high to low power will only go live occasionally. The rarer those contexts are, we might think, the more appropriate restraint will be, given the nature of such instances of blame as edge cases. The other way one could see the situation, however, is that the locality of blame means that hyper-local 'micro' relations of power come into play. It's true that, for instance, an office worker may run in roughly the same strata of social power as their boss, but

contextual point that power is socially constituted and mutable, while not falling prey to a kind of relativism that says we can never analyze power relations and thus must rely on a norm like Blanket Blame Reduction.

Alcoff (2007) makes a similar point about ignorance and epistemology. She argues that an adequate epistemology must include an analysis of ignorance that goes beyond the conceptual and reaches instead to concrete economic systems. As she says, we need to look further than, "the general conditions of epistemic situatedness, the epistemic resources distributed differently across social locations, or the structural contexts that organize and reproduce oppression; to truly understand the cause of the problem of ignorance, we also need to make epistemology reflexively aware and critical of its location within an economic system" (57).

What this means is that when we notice the way that economic considerations affect our epistemic systems, we can see that ignorance is produced both at a first-order level (individual subjects are worse knowers when they are unaware of the forces of economics that structure and work on their lives) and at a second-order level (we are foreclosed against the possibility of asking certain kinds of questions about why people believe what they do, or how they come to have certain judgements). I agree, and although I think the economic situatedness of epistemology is perhaps its biggest lacuna, I'd extend the point (an extension I think Alcoff would agree with) to the other systems of power and oppression I've discussed: race, gender, sexuality, social credit, and so on.

What power does when it operates covertly is reproduce a status quo that keeps power largely hidden. That this should reproduce itself in our systems of *responsibility*, epistemically and otherwise, is unsurprising. Return to the Mill (1909) quote with which I opened this chapter: "Where there is an ascendant class, a large portion of the morality of the country emanates from its class interests... The likings and dislikings of society, or of some powerful portion of it, are thus the main thing

in the day-to-day context of their work, the gaps in their local power may be enormous.

which has practically determined the rules laid down for general observance, under the penalties of law or opinion" (10-11). One thing which Mill is pointing out is that the reproduction of power within a society occurs directly through the maintenance of class interests – including the "likings and dislikings" of the powerful. It is no surprise then, if, more directly, power dynamics reproduce themselves precisely by constructing forms of ignorance that keep the less powerful directed away from blaming the more powerful and the more powerful comfortable with blaming those with less power. Notice that this needn't occur in a conspiratorial way, with explicit top-down direction. It is simply a fact that no one likes to be blamed. Because of this, if it is in your power to deflect blame away from yourself - we shouldn't be surprised if that power is taken advantage of. This is especially true where "taking advantage" can occur largely (or completely) unconsciously and with little effort. All a very powerful person needs to do is let it be known that they don't like being blamed, and their power and the fear of it that others have will push the work onto their inferiors. Indeed, they don't even need to advertise this, since, as I said above, we all know that no one likes to be blamed! It seems that, if one is sufficiently powerful, all one has to do to keep blame pointed away from oneself is decline to learn unsavory facts about oneself or one's place in the world.

If what I've been arguing about our practices is correct, then it is accurate to say that our moral responsibility practices both reflect and help partially constitute an epistemology of ignorance. Gaps in power distort the ways in which we view more and less powerful agents as responsible, and the way in which our reactive attitudes are primed towards those agents. The asymmetric flow of blame from top to bottom further serves to entrench those power dynamics, and to keep those with more material power ignorant about the effects of their blame.[10] We are now in a position to complete the argument for the Standpoint Epistemic justification: because our

[10] Or, as one Atlantic article title glibly puts it, "Power Causes Brain Damage:" https://www.theatlantic.com/magazine/archive/2017/07/power-causes-brain-damage/528711/

practices are enmeshed in an epistemology of ignorance which is actively constructed, and because this epistemology operates, in particular, on the materially powerful in our society, it is likely that the less powerful are sometimes in a relatively better position to be accurate judges of responsibility relevant features of the powerful, and that the powerful are often in a relatively worse position to be accurate judges of responsibility relevant features of the powerless. If this is true, the powerless have justification for blaming the powerful more than they often do, and the powerful lack the justification to blame the powerless as often as they do.

Ignorance as an excuse?

All this talk of ignorance might remind the reader that there is a related discussion to be had about whether ignorance excuses. I said at the outset that this is one area of epistemology that moral responsibility theorists *have* dealt with - discussed in Chapter I as the "epistemic condition."[11] This first order question of moral responsibility re-emerges here at the level of theorizing about the norms of a good epistemology of responsibility. If we are often unjustified in our blaming because we don't have the right facts, is our ignorance excused? Is it right to blame *us* for mis or over-blaming? Of course, the previous section makes clear that a satisfactory answer here is complicated. And this is so even putting aside the complex issues of moral versus factual ignorance, which I'll return to below. Ignorance is not a good excuse, for instance, for white people to *remain* ignorant about racism if part of how that racism operates is by constructing white ignorance. Instead, white people are required to overcome that ignorance, and are blameworthy to the extent that they can and do not. Is the situation similar with regards to the epistemology of responsibility? Again, the simple answer is: yes. That is, the powerful are blameworthy insofar as they continue to blame the powerless if the origins of that blame are largely due to constructed ignorance which keeps them in power.

[11] See, for instance: Bradford (2017), Harman (2015), Nelkin (2014), and Smiley (2016)

Whatever the general prospects of ignorance excusing, it is the particular kind of ignorance that the socially powerful have claim to that we ought to resist as an excuse. To get a handle on the landscape of excuses, we can picture a four box matrix:

	Active	**Passive**
Culpable	Active and Culpable Ignorance	Passive and Culpable Ignorance
Non-culpable	Active and Non-culpable Ignorance	Passive and Non-culpable Ignorance

Begin with the passive ignorance column. I take it as a given that there are many cases of passive and non-culpable ignorance - indeed, that this is perhaps the most common kind of ignorance: if you simply didn't know better, that can often be an at least partially mitigating excuse for avoiding blame. Still, there may be cases of culpable passive ignorance, as explored above. These would be cases of the kind of negligence Alcoff was after: where declining to investigate one's privilege and position in society evinces a lack of care about the way one relates to and moves through the world. Such cases are worth exploring, but I am primarily interested in the active column. The basic claim of the literature on constructed ignorance is meant to push us towards recognizing that not all ignorance that at first appears passive really is. Actively constructed ignorance is harder to excuse. Still - it may be non-culpable - and whether and when it is is precisely what I'm exploring.

Tilton (2023), for instance, argues that there is a widespread mis-use of standpoint theory as a kind of excuse for remaining ignorant about the oppression faced by various kinds of groups. What she calls the "Strong Epistemic Disadvantage Thesis" (or SEDT), claims that "dominant social positions impose strong, substantive limits on what the socially dominant can know about the oppression of others. These limits are strong in the sense that the socially dominant cannot break free of them;

their ignorance is the inescapable result of their dominant social positions. The limitations are substantive in the sense that the socially dominant aren't just missing minor or trivial details; their social positions doom them to ignorance regarding matters of importance." I've already said that, in advancing the claim that our moral responsibility practices often perpetuate and help constitute of an epistemology of ignorance, I am *not* arguing for a version of the SEDT. That is, I am not claiming that such ignorance is inescapable or substantive. Indeed the use of the term Standpoint in the Standpoint Epistemic Justification may be slightly misleading here, as I briefly mentioned above. As Tilton persuasively argues, the "standpoints" discussed in standpoint epistemology are the result of conscious and effortful *work*. They are things to be achieved. What I'm discussing are epistemic differences between differently socially situated knowers. These are probably better technically described as differences in perspective, rather than standpoint. I don't want to muddy the water here or make too much of a fine-grained terminological distinction. The reason this matters is, once again, so that we can be clear about what I am and am not claiming the social powerful and socially disposessed ought to do. Given the histories and experiences of the powerful, some of them will *not* lack the kind of perspective that would help them empathize with the powerless. But many of them will. And so it is not that the socially powerful cannot achieve the standpoint necessary to understand why their blame might be problematic - it is that their perspective makes it less likely that they will, and more effortful for them to do so.

The flip side is that the perspective of being socially down-trodden and unempowered will not automatically grant one a standpoint of clarity about responsibility and blame. As Olúfemi Táíwò (2020) argues, "Contra the old expression, pain – whether borne of oppression or not – is a poor teacher. Suffering is partial, short-sighted, and self-absorbed. We shouldn't have a politics that expects different: oppression is not a prep school." Being powerless is no gaurantee of achieving a liberatory standpoint. It is not even a guarantee of having a more canny or street-smart perspective. But, just as the point for the powerful was that their situatedness makes it, in general, harder to achieve a standpoint that recognizes the plight of the powerless, the perspective of *low*-power may provide a more fertile starting point in which to achieve the right kind of standpoint from which to blame.

Refocusing on the kind of cases we are interested in, then, we wan to consider ones where ignorance is a useful shield that those in power, more or less consciously, choose to take up for various kinds of self-protection. This is a crucial distinction. The powerful may want to use their ignorance of the oppression of dominated groups as an excuse for failing to act better, more justly, or improving the conditions dominated groups are in. They may want, in the case of blame, to use their ignorance as an excuse for blamelessly blaming those in worse off positions. And, if something like the SEDT were true, then the powerful would always remain ignorant. As Tilton writes: "If the insights of marginalized people are uniquely theirs, in the sense that the socially dominant cannot understand or make use of those insights, then not only do the socially dominant not know — they can't know. Thus, the SEDT is at odds with work that argues that the ignorance of the socially dominant... is actively cultivated rather than a mere passive occurrence" (10).As Tilton writes:

> If the insights of marginalized people are uniquely theirs, in the sense that the socially dominant cannot understand or make use of those insights, then not only do the socially dominant not know — they can't know. Thus, the SEDT is at odds with work that argues that the ignorance of the socially dominant... is actively cultivated rather than a mere passive occurrence." (*ibid.*)

They could simply say, "well, I couldn't have known any better," and it will always be true. Their ignorance will remain, and the excuse will always be available, and it will seem, therefore that we have no way of breaking out of the cycle of epistemic and material injustice that the ignorance contributes to.

Thus, we precisely want to deny a thesis like that of the SEDT: being in an oppressive or dominant social position does not excuse one from working to try to understand the realities of the social world. And one isn't excused precisely because coming to know the things that those in less powerful positions know is, in principle, possible. In other words, we want to avoid our theorizing having as an outcome that those in power are able to use the language of marginalization and social justice to further entrench their domination. This calls to mind a distinction from Calhoun (1989) concerning the difference between moral ignorance in "normal" and "abnormal" contexts. On her view, an "abnormal" context is one in which we are at the "frontiers" of moral knowledge. Discussing this in terms of feminist theorizing, Calhoun has in mind situations where concepts, language, or norms are not yet wide-spread in moral discourse, such that ignorance of certain kinds of wrongdoing is the norm. So, for instance, before the invention of specific language and norms concerning workplace sexual harassment, there might be widespread ignorance of some kinds of moral wrongs in the workplace (to use Fricker (2007a)'s famous example). An abnormal moral context is particularly burdensome for those who recognize wrongdoings because they may experience testimonial injustice and be unable to explain why the things that are happening constitute wrongdoings. A "normal" moral context, on the other hand, is just one where ignorance is not widespread: where most members of a moral community agree on and are aware of the moral norms (whether or not they agree on the details/applications/enforcement of those norms).

For our purposes, the interesting question here is whether the socially powerful would be in an abnormal context when they appeal to their moral ignorance. Does a wealthy CEO have, as an excuse to appeal to, ignorance of the relevant harms that might be involved in blaming downward towards a low-ranked hourly worker at her company? My answer is that, to the extent that this is a legitimate excuse, it cannot be because of something like the SEDT. The extent to which the powerful are in an abnormal moral context is just the extent to which the epistemology of moral responsibility remains opaque to everyone. But the rich and powerful do not gain a second type of excusing ignorance of the harms of blame because of their power. Actively constructed ignorance is something that, absent some further explanation, rules out ignorance as a justifiable excuse. As Tilton puts it: "by acknowledging the supposed limitations imposed by their dominant social positions, the socially dominant reap the benefits that (in some contexts) come along with publicly signalling a raised social consciousness. The SEDT is, then, a 'no risk, all reward' strategy for the socially dominant. The costs of the strategy are instead borne by the marginalized" (13). Actively working to maintain and construct ignorance about the harms one is culpable for, under the guise of saying, "I recognize my limitations as a knower," shows precisely that one has had sufficient opportunity to learn about the ways their actions might be resulting in harm.

Insofar as blame of the socially powerful is inappropriate in light of their ignorance, then, it must be because of some other reason than being in an abnormal moral context – it must be related to whatever general reasons there are for ignorance to excuse. Finally, even where such excuses might be legitimate, there might be good general reasons to resist them, or not take them too seriously. Recall that I claimed in the last chapter that, although we ought to expend some effort and hold out some hope for improvement, we shouldn't focus on imploring the powerful to be better blamers as a cornerstone of our revisionism. This claim might seem to introduce an odd tension with the fact that I argued for the norm of Powerful Restraint. But recall that the thrust of my arguments was that, although we need this norm because the powerful produce an out-sized amount of harmful blame which we ought to reduce, the enforcement of the norm would probably need to rely on concerted efforts at blaming upwards: the meta-norm of blaming the powerful when *they* blame badly. Indeed, I have tried to make robustly clear that blame of the socially powerful is often generally appropriate here for forward looking and instrumental reasons. The powerful *can* improve along epistemic and moral lines, they can become better at directing and considering the effects of their attributions of responsibility, but it's that they are unlikely to do so without meaningful social pressure. Our blaming of the powerful signals that this is what we expect from them. Of course, whether blame is likely to inspire self-betterment is a highly contested question. I only mean to point out that the reasons for blame here are not primarily backward looking or oriented around desert.

The general phenomenon I'm tracing is a structural and social kind of ignorance. Although it is constructed, it isn't necessarily the case that individual, socially powerful agents are deeply invested in or consciously working to (re)construct the conditions of ignorance that protect them. Instead, the ignorance arises out of general facts about our psychologies and the set up of our practices. Should we be blamed for erring in this context? The instrumental and forward looking response will be to say: it depends! It depends on whether doing so will make us better moral agents, whether doing so will improve the epistemic set up of our practice going forward, and so on. One particular way in which blaming even the ignorant may do this is by signalling the norms we are committed to - and it is to a discussion of this kind of reason I now turn.

4.4 Social Signalling Justifications of Asymmetry

The second kind of justification of the asymmetry of blame is one of "social signaling." This justification is intimately related to the arguments above. If our responsibility practices are enmeshed in, and help to bulwark an epistemology of ignorance, then correcting this injustice will involve signalling our commitment to opposing it. Return to Ciurria's point about narratives. What social signalling allows us to do, in part, is construct a new set of meta-narratives within and about our practices. Not only can signalling directly strengthen (and perhaps even enforce) existent norms in a practice by making others aware of our commitments, it can be used to shift the normative landscape. This is particularly useful when blame flows from low to high power. As Ciurria (2020a) writes, "When people blame perpetrators who don't recognize their moral authority, their blame can play a valuable (and ameliorative) role in interpersonal networks outside of the victim-perpetrator relationship. It can... educate third parties, motivate third parties to protest wrongdoing, speak to the moral dignity of victims, and advance other interpersonal aims" (54-55). That is, because the high status are unlikely to be directly moved or genuinely influenced by low-status blame (much less notice it), the more relevant use of blame in such cases is to influence the way other moral agents view an action, state of affairs, or person.

Much here depends on the nature of the relationships involved, and speaking only schematically of low-high power relations won't cover all of the relevant cases we're interested in. Still, developing such a schema is a helpful start. We should take each piece in turn, saying more about blame as a social signal, about the possibility of signalling to restructure the normative landscape, and about the efficacy of meta-narratives to do so.

Begin with blame as a social signal. Several theorists have recently argued that one way to resolve intractable conceptual and definitional fights about blame is to see the overall functional role of blame as one of social signalling. Shoemaker and Vargas (2019), for instance, argue that blame is a costly social signal. Contrasting their functional account with those that attempt to define blame based on its distinctive content, they write:

> On content-based accounts, blame is a distinctive attitude or activity. The practice of blame — that is, blaming — is then understood in terms of where and how that attitude or activity occurs. Our proposal inverts that relationship: it is the function of blame—a signal about our normative commitments—that determines which attitudes and activities are instances of blame when they are... In the typical case, an agent's internalized norms and his or her disposition to enforce those norms are what generate the signal's reliability" (7).

In other words, blame signals what kind of practical agent I am by letting others know what norms I am committed to, and this, in turn, helps articulate and defend a set of practice-wide normative orientations. For Shoemaker and Vargas, then, blaming signals: 1) normative competence, 2) commitment, 3) self-disclosure, 4) various demands and conversational invitations, and 5) beliefs about voluntariness and intention (or the lack thereof). Importantly, the signal, although it may have a univocal source, can also say different things to different audiences.

One can see how this view of blame fits nicely with my own arguments about the epistemology of responsibility as interpretive, socially mediated, and conversational. Whether one believes that the sole, primary, or distinctive function of blame is social signalling, the key point stands: blame *can* and often *does* signal normative commitments. It is useful as a way of signalling moral demands. I take it that this point stands as well even if one is convinced by, for instance, Brink and Nelkin (2022), who have recently argued that functional accounts of blame may fail insofar as blame is, for example, not a *good* or productive way of actually enforcing norms or shaping the normative landscape. As a signal of normative commitments blame is unmistakable, whether or not it is always productive.[12] At the very least, I think we can follow Wang (2021, 2022) in viewing a context sensitive, functional role of communication as one distinctive aspect of blame compatible with an overall pluralist concept of blame.

Of course, the degree to which moral demands will be taken up largely depends on the kind of social status one has within an interpretive community. The social signalling of a respected local elder may have far more weight than the social signalling of a town drunk. We may worry, then, that those with low social power will find that their blame is an ineffective social signal. There are some contexts and senses in which this is correct - but they needn't overly concern us. It is true, for instance, that the town drunk's blame will not be likely to taken up communally, and we can easily imagine ways in which this means their blame might be unfairly discounted.

[12] Although, see, for instance Scaife et al. (2020) as a useful counterpoint to the claim that blame is counterproductive.

We could then extend that kind of analysis to those with low-social power in general. But this is just one dimension of social signalling.

It's true that the ways in which those with low social power are able to directly influence the behavior of those they blame might be limited. But I take it that this is not the primary upshot of the signalling view. Instead, what we are largely up to is signalling normative commitments to our peers. It may be that a factory worker blaming a billionaire CEO is unlikely to lead to much material change on its own. But as a social signal it can have huge benefits. First, it can alter and influence behavior in ways that aid in overall social cooperation and norm enforcement. This is so both in the way it can commit the *blamer* to certain norms, and, obviously, in the ways that it signals to others the kinds of norms you are prepared to enforce. Relatedly, it functions as a kind of signalling (and enacting) of solidarity. It can show that you are committed to the same norms as your fellows - which has huge normative stakes for your relationships with those in your social strata, quite apart to what those who are vastly more or less powerful than you make of it. And finally, it has the possibility to create new moral conversations. The fact, then, that social signaling might come apart from any kind of direct payoff in token instances shouldn't worry us. Whether the signalling works for any token instance is less important than if it plays some overall role in altering behavior which helps us more easily cooperate and enforce norms.

Why then, does the social signalling view support asymmetric blaming? Because the normative landscape is already tilted in favor or (and largely created by) the powerful, they have little need of shaping moral conversations via blaming as a social signal. So, if they restrain their blame, they suffer comparatively little moral cost. Of course, this is no guarantee that they will do so, and the high costs of their signalling are precisely why I argue for their restraint. If Elon Musk forgoes publicly blaming his workers for a delay in constructing a pointless, ineffective, and redundant "futuristic subway system", he suffers no moral injury.[13] Insofar as they have cost him time or money, he can set in place new rules or systems that will prevent similar future incidents without also subjecting them to social sanction.

On the other hand, if a Tesla worker is injured in a workplace accident because Mr. Musk doesn't like the sound that forklifts make when they back up, public blame may be their most effective recourse at, not only maintaining their dignity, but making sure that similar accidents do not occur again in the future.[14] The worker can do this by signalling that what Musk has done is wrong, in the hopes that others will latch on to the signal and amplify it. By asking the broader moral community to boost the signal, they may be able to change workplace norms. Such normative re-shaping is likely to be time-consuming, risky, and without guarantee of success, but large power imbalances serve as partial justification for it nonetheless.

To return to the original objection: what if the worker is wrong about Mr. Musk's blameworthiness? Shouldn't they be equally cautious about blaming upwards as he would be about blaming downwards? Assume that the worker is incorrect that the reason the forklifts in the factory do not beep when they go in reverse is because Mr. Musk dislikes the noise. What will be the likely harm of their blaming the CEO? It seems unlikely to me that Mr. Musk will suffer any real reputational damage. Indeed, in the intervening years between the publication of that fact and the present day, he grew vastly more wealthy and powerful. However, the signalling of the injustice of unfair working conditions, and the laying of blame at the feet of the CEO of the company *was*, to some extent, picked up in the media. Again, let's assume that the story is false, and that the real reason the forklifts don't beep in reverse is some kind of manufacturing decision by the forklift maker. Mr. Musk, we have said, suffers no

[13] See the "Hyper Loop" system that he has pitched to various cities and constructed in Las Vegas: `https://www.lvcva.com/loop/`. Passengers go underground to ride single file in individual Teslas, a vastly inferior subway system.

[14] See the NY Magazine Intelligencer article, https://nymag.com/intelligencer/2018/04/tesla-workers-getting-hurt-because-elon-musk-hates-yellow.html

real harm.[15] But if OSHA is putting more scrutiny on the safety of his factories, he may be motivated to contact the manufacturer and get the forklifts to beep. Thus, the blame, and the social pickup of the signal can lead to a positive outcome, even where factual errors occur. This is why I claim that the asymmetry of blame is justified by social signalling.

Of course, this is a rather easy case. Mr. Musk is a billionaire, and almost everyone can agree that workers deserve safe working conditions. Can this kind of argument really work in other kinds of cases, where, for instance, the power imbalance is less extreme, and the likelihood of harm from misplaced blame is higher? Again, much depends on the details of a given case. The reason I have focused on large power imbalances is that these seem to be the least risky when blaming upwards (and the most risky when blaming downwards). I will give a brief answer to these questions of ecological complexity and risk assessment in the following, concluding section.

4.4.1 The Ecology of Blame: Conclusions

A core part of the social signalling view canvassed above is that our responsibility practices, including our blaming practices, help make possible social coordination and cooperation at scale. Blame is one mechanism we have to signal and uphold norm enforcement. A natural worry then, is whether asking the powerful to blame less undercuts coordination and cooperation. I've identified some unintended and pernicious costs to robust norm enforcement, in other words, but the costs of forgoing that norm enforcement might be just as weighty. The trade-offs here are likely to be extremely complicated and difficult to predict. Just what the consequences to increasing and decreasing blame in various parts of our responsibility practices

[15] Very recently, Mr. Musk has begun to suffer real reputational damage, but this appears to have far more to do with his bungled purchase and running of Twitter - i.e. something that lost him money, as opposed to any weighty ethical complaints.

might be is an empirical question. Authors like Pickard (2011, 2013), Pereboom (2001, 2014), and Shaw, Caruso, and Pereboom (2019) have canvassed reasons to think that a world with less blame (or less blame of a certain kind) might be a better one. But this is highly conjectural, and as Bicchieri (2017) in her excellent *Norms in the Wild* is at pains to argue, norm change is a difficult and inexact business. As she points out, even if we have reasons to change a norm or set of norms, and even if we know it would be good to change, we still have to be sure that we are acting collectively – that others are also changing to a new normative schema (107-112). This makes norm change a risky proposition. Deviating from a social norm is very costly, and in case of either abandonment of old norms or the creation of new ones, we face a situation where we must overcome a large scale social coordination problem through collective action. This requires having shared reasons to change norms, collective social expectations, and some coordination of action. Here I merely want to note how difficult norm change is, and that I would recognize it as a real cost if we lost out on the goods that norm enforcement and cooperation and control can get us. These arguments, then, are conditional on the hypothesis that introducing a new kind of asymmetry in blaming practices would, a) be possible, and b) not fundamentally (or drastically) break our existing social arrangements.

Still, I can say a bit more about why I have confidence that this practical set up would not be society breaking. We've seen that the Standpoint Epistemic and Social Signalling justifications may be enough to justify, *prima facie*, an asymmetric model of the acceptability of blame in our moral responsibility practices. But the previous section once again raised the issue of the vast complexities involved in making sense of the power dynamics within our interpersonal and communal relationships with one another. We ought to say more, then, about the way my view makes sense of the overall landscape of our practices as an "ecology." To triangulate, let's begin with an extreme view. Reis-Dennis (2021), has recently argued that Strawsonian resentment is best understood on a social, ecological model. By this he means that blame is properly understood to be about social imbalances of power under conditions of wrong-doing. In other words, it is proper for me to resent and blame you when: a) you have done something wrong, and b) there is a power imbalance between us and you are more powerful. Wrongdoing is thus necessary, but not sufficient for blame. It is improper for me to resent you for wrongdoing if you are of such low social status that my blame would merely concretize this imbalanced relationship.[16]

There are important similarities between Reis-Dennis' view and my own. We both take ecological imbalances to be important and significantly explanatory of both: (a) the psychological mechanisms that are actually operative in everyday blame, and (b) the normative structure of what makes blame and resentment a proper or fitting attitude. And, accordingly, we both hold that some instances of blame by the powerful towards the downtrodden may be inappropriate, despite the satisfaction of the criteria of standard theories of moral responsibility. As he writes:

> Doesn't the ecological view imply, for instance, that oppressed people may find themselves properly exempted from resentment? The answer is yes: the social power view does imply that marginalized people will sometimes be un-resentable, and this state of affairs is indeed a threat to their dignity and self-respect. But the problem here is with our social systems rather than the psychology of resentment. To force ourselves, in the name of morality, to resent those who do no damage, or to see the weak as excessively powerful, would be to resent on the basis of considerations extraneous to those that determine the attitude's appropriateness. (18)

I agree with Reis-Dennis that what we ought to notice here is that the set up of our social systems are where things have gone wrong. However, Reis-Dennis and I depart at an early stage of motivation, for two reasons. First, he seems to hold that in cases of power-imbalances, we are often motivated to blame downward for reasons of upholding "morality" and/or the dignity and self-respect of the targets of our

[16] The same is true of the power-imbalance itself: it is necessary but not sufficient. Resenting without wrongdoing is something like envy, jealousy, or hatred: not resentment at all.

resentment. I'm extremely doubtful that this is psychogically accurate. Indeed, if it were, the world I've been describing in this dissertation would be completely turned on its head. If the powerful have been blaming the dispossessed for second-order reasons based on respect, we could run a "wrong-kind-of-reason" argument and ask that they refrain from doing so. The task of balancing the ecology of blame would be very simple if it turned out, after all, that the powerful don't really *want* to blame the less-powerful, but have been doing so merely because they have had to force themselves to resent in the name of morality.

Second, Reis-Dennis does too much to decouple the link between conditions of agency and responsibility that almost all theorists hew to: the idea that agential capacities such as "reasons-responsiveness, rational control, self-expression, evaluative capacity, and so on" are necessarily linked to the proper resentability of a given agent (17). In explaining why it is improper to resent children, for instance, he argues that:

> [A] lack of social power, not moral power understood in terms of the capacities required for moral agency, is the crucial barrier to adult-child resentment. Children's lack of development is, of course, one of the reasons for their lack of relative social strength, but this is, in a way, a red herring. Their social weakness... is what limits the amount of damage they can do, and thereby prevents them from being fitting targets of adults' resentment... resentment is not only, or even primarily, about the attitudes, evaluative claims, and judgments of moral reasoners, but rather the concrete social effects of ill will and wrongdoing. (7)

But here, Reis-Dennis proves too much, and arrives at an unworkable position. If agency is no barrier to resentment, I can properly resent a very powerful gust of wind - but this kind of reactive attitude serves no social purpose. Reis-Dennis sees resentment as a moralized attitude that emerges in everyday life because of deep-seated psychological (perhaps evolutionary) pressures having to do with fairness and desert. He also sees resentment as a social tool: as something which is utilized by the powerful to maintain their privilege and, in extreme cases, uphold oppression. Fair enough. But, Reis-Dennis himself added in a necessary condition on resentment that it properly track wrongdoing (or ill-will). Indeed, he ends the above quote by talking about the social effects of *ill will and wrongdoing.* But how, we might ask, is this compatible with his claim that agential capacities are unnecessary for determining the conditions of proper resentment? In what sense can I tell that you have done something out of ill will if I am not interpreting your attitudes? The answer, in short, is that I cannot.

Instead of biting the bullet about agency and its link to responsibility, Reis-Dennis could follow my example and bite elsewhere: as he admits, agential capacities *are* properly linked to blame. But, as I've argued at length, we simply don't have reliable, universal, or constant access to facts about those capacities. And, the very ecology of power and privilege he notes serves to further distort our ability to reliably gain that access. So, it's not that we can properly blame the powerful whether or not they have agential control, it's that, given our general lack of epistemic access to facts about agential control, social status is a better guide for the appropriateness of blame *in conditions like ours.* I've remained insistent that my project is compatible with realism about responsibility for precisely this reason.

Even so, there is something compelling (if controversial) about what a view like Reis-Dennis' is driving at: at the intersection of our real-world practices and our ideal theory we must pay attention to the material conditions and social structures that mediate our blaming. When we ask whether our practices are just, or whether they can be improved, we shouldn't attempt to answer merely based on idealized conceptions of blame and responsibility. We need to look and see how things really work in the messy world we inhabit. The reason that power-imbalances matter so much is because they condition what our blame *does.* If we care about instrumental and forward looking justifications of our responsibility practices, then we must also care about what our attributions of responsibility amount to, practically speaking. It is not enough to show that blame would be justified by pointing out that an actor has the relevant agential capacities – this is a necessary but non-sufficient answer to the question of whether blaming is appropriate.

One extreme, then, is to pay no attention to the real world. The other extreme is to jettison (as Reis-Dennis does to some extent) any theoretical constraints and focus only on outcomes. This goes too far in the other direction: in giving up on our theoretical architecture we give up on the game itself. That is, we still need to know when we *are* blaming, that we are engaging in practices of responsibility attribution, that we are holding one another accountable and so on. If instrumental value is our only guide, we lose sight of these basic questions of justification and fairness in such a way that our practices themselves cease to be rational, sensible, or justifiable.

We want, in other words, to know whether our blame is *useful*, while still retaining it as a distinct and defensible concept. In order to pursue this middle ground, one plausible candidate for an overarching theoretical perspective is the kind of ameliorative ideology critique argued for by Haslanger (2012). When we examine blame from a critical lens we open up the space for new conceptual choices and revisions. But these revisions are not ad-hoc or unguided. As Haslanger points out, "the social theorist's task is to situate a practice within a broader causal and moral context that those engaged in the practice ordinarily aren't aware of" (20). It is this kind of perspective I've taken up in this dissertation. And, as I've said, I am drawn to the perspective of authors like Holroyd (2018), McGreer (2013), and Vargas (2018b) who argue (in importantly distinct ways) that our responsibility practices can enhance, scaffold, or partially constitute our agency.

If all of this is on the right track, we can have a theory of the appropriateness of blame that holds on to the theoretical architecture of responsibility *and* allows for significant guard-rails and overrides due to forward looking concerns about agency cultivation. It is precisely in this hybrid spirit that I defended instrumentalism and introduced Powerful Restraint in Chapter Three. And, indeed, in which the high stakes nature of blame was surveyed, and in which the Standpoint Epistemic and Social Signalling justifications were put forward in this Chapter.

In conclusion, let me revisit one final objection to my project which has lingered through the second half of this dissertation. I've painted our reactive lives as substantially captured by non-cognitive and non-deliberative factors. Affect and emotion, I've claimed, often distort our epistemic processes such that their accuracy is doubtful and blame has, overall, a justification problem. Yet, my solution appears to ask agents to do a better job of thinking through and reacting to potential instances of blameworthiness. Isn't this exactly the wrong kind of solution? If the problem with blame is that epistemic disruption occurs at points in the ping-pong model where we do not have direct, cognitive control, then what good is it to ask the powerful to blame less or the powerless to blame more? These just aren't the kind of things we can decide to do. One of P.F. Strawson's original points, in calling our attention to the reactive attitudes, after all, was that we are party to them whether we believe in freedom of the will (or whatever other metaphysical conditions) or not.

In one sense this objection is easily answered by recalling that I've discussed blame as having both an assessment and an expression phase.[17] It's true that we can do little to control (directly) what happens in the assessment phase. Even here, of course, we can adopt strategies that modify our environment and behavior so that we are less prone to assess agents as blameworthy (or more prone to do so). These kind of indirect methods have been much discussed in the literature on implicit bias, for instance, in how we might mitigate bias in hiring decisions.[18]

Yet, the lack of control at the assessment phase does not mean we have no control over the expression phase. It is sometimes the case that we are overcome with emotion (or, perhaps, overcome with reason – "It just *has* to be true that it's a good idea to blame!"), and cannot help but express our blame. But very often we need to decide whether to proceed with some active expression of blame – whether that is

[17] Malle, Guglielmo, and Monroe (2014) identify these two aspects as "cognitive blame" - a private judgment - and "social blame" - a public act.

[18] For representative discussions of strategies, pitfalls, and so on, see: Baumeister, Ainsworth, and Vohs (2015), Holroyd (2012), Moss-Racusin et al. (2012), and Spaulding (2018), chapter six.

self-blame, interpersonal blame, or registering some kind of moral protest. Thus, a surface level answer is available, and it is one I have already been pursuing: that agents pause and consider the likelihood of their being in error, as well as the likely effects of their expression of blame, before they move from assessment to expression.

Things are, of course, rarely only surface deep. Hieronymi (2019) has recently objected, for instance, that thinking of blame as a voluntary reaction obscures it's point and makes us defensive. As she puts it, "by focusing on merited consequences and overlooking non-voluntariness, we risk misunderstanding the significance of moral criticism and of certain reactions to moral failure" (2). Her concern is the messy middle ground between passivity and activity when it comes to the reactive attitudes. Reactive Attitudes seem to involve both non-voluntary mental states and voluntary actions, such as expressed blame. However, as is a familiar point in the literature on responsibility, especially for those who favor self-disclosure views, there is an important sense in which our non-voluntary mental states are still "up to us." That is, my reactions to the world around me are expressive and disclosing of my values - my "take on things." This is what quality of will theorists mean when they say that an action can express something about what I stand for and how I perceive you. And, Hieronymi is noting, this seems equally the case for our *reactions.* As she puts it: "Although these reactions are non-voluntary, they are neither involuntary nor out of one's control – they are... up to the person reacting, something that person can revise and can be criticized for. Better, I think, to say they are manifestations of the way in which people matter to other people" (31-31).

These non-voluntary aspects of blame matter because they pick out blame as a unique moral emotion. Hieronymi's idea is that if I blame you, I do not merely criticize your actions. If reactive attitudes were merely critiques, they could be purely forward looking, and have nothing to do with notions of desert. Blame would merely suggest, for instance, that an agent ought to focus on changing or bettering themselves so that they don't do blameworthy things in the future. I have focused much of my attention on the forward-looking aspects of blame, but, of course, I have also described the ways in which blame is backward looking and bound up with notions of desert. It is not merely a critique precisely because, in focusing on the wrong done to a victim, it suggests the need for recompense, atonement, or, at the very least, conversation.

Hieronymi concludes that she is, "tempted to say that being the target of resentment or of indignation is more like suffering from a hangover than like being sent to your room: it is, in a sense, a natural consequence of your disrespect or disregard of others" (30). But here we must depart from her account. It is precisely because blame can be mistaken - that it can be an in-apt reaction - that it is not simply a natural, causal consequence of our actions. And, in particular, it is the presence of two gaps (which I've discussed at length in this dissertation) that matter when we think about voluntariness. First, the interpretive gap - the space in which we make sense of what an agent did and judge whether it was blameworthy, and second, the gap between assessment and expression - the space we have to decided whether to move from an assessment of blameworthiness to an expression of blame.

These explorations of social signalling, the ecology of blame, and blame's voluntariness all add up to the following: what's needed "on the ground" are new interpretive norms in the practices of responsibility. What's needed at the level of theory is a greater recognition that practices of responsibility *are* epistemic practices largely governed by norms of interpretation. In this dissertation, I've begun that work, but there is much more to do. Norm change is difficult, and so too is self-change – but both are possible. In the case of responsibility and blame, change can be aided by fruitful work that interrogates what good interpretive norms would look like in our responsibility practices, given what we know about the epistemology of responsibility. Linking the last several decades of work in the metaphysics of responsibility with work on narrative, interpretation, and the epistemic and social dimensions of responsibility is a path ripe for the taking.

Works Cited

Abizadeh, Aresh (2021). “The Grammar of Social Power”. In: *Political Studies*, 1–17. ISSN: 00016993.

Agule, Craig K. (2016). “Resisting Tracing’s Siren Song”. In: *Journal of Ethics & Social Philosophy* 10.1.

Alcoff, Linda Martín (2007). “Epistemologies of Ignorance: Three Types”. In: *Race and Epistemologies of Ignorance*. Ed. by Shannon Sullivan Nancy Tuana. 2007.

Alicke, Mark and Constantine Sedikides (2009). In: *European Review of Social Psychology* 20.1, 1–48.

Alicke, Mark D (2000). “Culpable Control and the Psychology of Blame”. In: *Psychological Bulletin* 126.October, 556–574. `https://doi.org/10.1037//0033-2909.126.4.556`.

Alicke, Mark D. et al. (2008). “Culpable control and counterfactual reasoning in the psychology of blame”. In: *Personality and Social Psychology Bulletin* 34.10, 1371–1381. ISSN: 01461672. `https://doi.org/10.1177/0146167208321594`.

Alicke, Mark D. et al. (2015). “Causal Conceptions in Social Explanation and Moral Evaluation: A Historical Tour”. In: *Perspectives on Psychological Science* 10.6, 790–812. ISSN: 17456924. `https://doi.org/10.1177/1745691615601888`.

Allen, Amy (2016). “Feminist Perspectives on Power”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2016. Metaphysics Research Lab, Stanford University, 2016.

Barden, Jamie et al. (2004). “Contextual moderation of racial bias: The impact of social roles on controlled and automatically activated attitudes”. In: *Journal of Personality and Social Psychology* 87.1, 5–22. ISSN: 00223514. https://doi.org/10.1037/0022-3514.87.1.5.

Bargh, John A. and Tanya L. Chartrand (1999). “The unbearable automaticity of being.” In: *American Psychologist* 54.7, 462–479.

Basu, Rima (2021). “The Specter of Normative Conflict: Does Fairness Require Inaccuracy?” In: *Applied Epistemology.* 2021. Chap. 10.

Basu, Rima and Mark Schroeder (2018). “Doxastic wronging”. In: *Pragmatic Encroachment in Epistemology.* Ed. by Brian Kim and Matthew McGrath. New York: Routledge, 2018, 181–205. ISBN: 9781351685252. https://doi.org/10.4324/9781315168197-11.

Baumeister, Roy F., Sarah E. Ainsworth, and Kathleen D. Vohs (2015). “Are Groups More or Less than the Sum of their Members? The Moderating Role of Individual Identification”. In: *Behavioral and Brain Sciences* 2016. ISSN: 14691825. https://doi.org/10.1017/S0140525X15000618.

Bayles, Michael D. (1982). “Character, Purpose, and Criminal Responsibility”. In: *Law and Philosophy* 1, 5–20.

Bicchieri, Cristina (2017). *Norms in the Wild.* Oxford: Oxford University Press, 2017. ISBN: 9780190622046.

Bjornsson, Gunnar (2017). “Explaining (Away) the Epistemic Condition on Moral Responsibility”. In: *Responsibility: The Epistemic Condition.* Ed. by Philip Robichaud and Jan Willem Wieland. Oxford: Oxford University Press, 2017. Chap. 11.

Bloom, Paul (2016). *Against Empathy: The Case for Rational Compassion.* Ecco Press, 2016. ISBN: 9788578110796. https://doi.org/10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

Bradford, Gwen (2017). “Hard to Know”. In: *Responsibility: The Epistemic COndition.* Ed. by Philip Robichaud and Jan Willem Wieland. Oxford University Press, 2017. Chap. 10.

Brewer, Marilynn B. (1977). “An information-processing approach to attribution of responsibility”. In: *Journal of Experimental Social Psychology* 13.1, 58–69. ISSN: 10960465. https://doi.org/10.1016/0022-1031(77)90013-0.

Brink, David O and Dana Kay Nelkin (2013). “Fairness and the Architecture of Responsibility”. In: *Oxford Studies in Agency and Responsibility*, 284–314.

— (2022). “The Nature and Significance of Blame”. In: *The Oxford Handbook of Moral Psychology.* Ed. by John Doris and Manuel Vargas. New York: Oxford University Press, 2022, 1–17.

Buunk, Bram P. and Thomas Mussweiler (2001). “New directions in social comparison research”. In: *European Journal of Social Psychology* 31.5, 467–475. ISSN: 00462772. https://doi.org/10.1002/ejsp.77.

Calhoun, Cheshire (1989). “Responsibility and Reproach”. In: *Ethics* 99.2, 389–406.

Ciurria, Michelle (2020a). *An Intersectional Feminist Theory of Moral Responsibility.* New York: Routledge, 2020. ISBN: 9780367343972. https://doi.org/10.4324/9780429327117.

— (2020b). “The Mysterious Case of the Missing Perpetrators”. In: *Feminist Philosophy Quarterly* 6.2. ISSN: 2371-2570. https://doi.org/10.5206/fpq/2020.2.7322.

Clarke, Randolph (2014). *Omissions: Agency, Metaphysics, and Responsibility.* Oxford: Oxford University Press, 2014.

Cohen, Stewart (2018). “Pragmatic Encroachment and Having Reasons”. In: *Pragmatic Encroachment in Epistemology.* Ed. by Brian Kim and Matthew McGrath. New York: Routledge, 2018, 101–106. ISBN: 9781351685252. https://doi.org/10.4324/9781315168197-11.

Collins, Rebecca L. (2000). “Among the Better Ones: Upward Assimilation in Social Comparison”. In: *Handbook of Social Comparison*. Ed. by Jerry Suls and Ladd Wheeler. New York: Kluwer Academic Publishers, 2000.

Cudd, Ann E. (2006). “Analyzing Oppression”. In: *Analyzing Oppression*, 1–304. `https://doi.org/10.1093/0195187431.001.0001`.

Cuddy, Amy J.C., Susan T. Fiske, and Peter Glick (2007). “The BIAS Map: Behaviors From Intergroup Affect and Stereotypes”. In: *Journal of Personality and Social Psychology* 92.4, 631–648. ISSN: 00223514. `https://doi.org/10.1037/0022-3514.92.4.631`.

Deery, Oisín and Eddy Nahmias (2017). “Defeating Manipulation Arguments: Interventionist causation and compatibilist sourcehood”. In: *Philosophical Studies* 174.5, 1255–1276. ISSN: 15730883. `https://doi.org/10.1007/s11098-016-0754-8`.

Dennett, Daniel (1987). *The Intentional Stance*. Cambridge, MA: MIT Press, 1987.

Devine, D.J. and D.E. Caughlin (2014). “Do they matter? A meta-analytic investigation of individual characteristics and guilt judgments”. In: *Psychology and Public Policy Law* 20.109.

Doris, John (2002). *Lack of Character: Personality and moral behavior*. Cambridge: Cambridge University Press, 2002.

Dotson, Kristie (2014). “Conceptualizing Epistemic Oppression”. In: *Social Epistemology* 28.2, 115–138. `https://doi.org/10.1080/02691728.2013.782585`.

Ellison, L. and V. Munro (2008). “Reacting to rape: exploring mock jurors’ assessments of complainant credibility”. In: *British Journal of Criminal Law* 49, 202–219.

Fantel, Jeremy and Matthew McGrath (2002). “Evidence, Pragmatics, and Justification”. In: *Philosophical Review* 111, 67–94.

Feigenson, Neal (2016). "Jurors' Emotions and Judgments of Legal Responsibility and Blame: What Does the Experimental Research Tell Us?" In: *Emotion Review* 8.1, 26–31. ISSN: 17540739. https://doi.org/10.1177/1754073915601223.

Fischer, John Martin (2006). *My Way : Essays on Moral Responsibility: Essays on Moral Responsibility.* Oxford: Oxford University Press, 2006, 272. ISBN: 0195346289. http://books.google.com/books?id=muxlMFSMYZ8C%7B%5C&%7Dpgis=1.

— (2012). *Deep Control: Essays on Free Will and Value.* 2012, viii–244. ISBN: 9780199742981. http://global.oup.com/academic/product/deep-control-9780199742981.

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

Fischer, John Martin and Neal Tognazzini (2009). "The Truth About Tracing". In: *Nous* 43.3, 531–556.

Fishbein, Martin and Icek Ajzen (1973). "Attribution of responsibility: A theoretical note". In: *Journal of Experimental Social Psychology* 9.2, 148–153. ISSN: 10960465. https://doi.org/10.1016/0022-1031(73)90006-1.

Fiske, Susan T., Amy J.C. Cuddy, and Peter Glick (2007). "Universal dimensions of social cognition: warmth and competence". In: *Trends in Cognitive Sciences* 11.2, 77–83. ISSN: 13646613. https://doi.org/10.1016/j.tics.2006.11.005.

Fiske, Susan T. et al. (2002). "A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition". In: *Journal of Personality and Social Psychology* 82.6, 878–902. ISSN: 00223514. https://doi.org/10.1037/0022-3514.82.6.878.

Frankfurt, Harry (1969). "Alternate Possibilities and Moral Responsibility". In: *The Journal of Philosoph* 66.23, 829–839.

Fricker, Miranda (2007a). *Epistemic Injustice: Power & the Ethics of Knowing.* Oxford: Oxford University Press, 2007.

— (2007b). *Epistemic Injustice: Power & the Ethics of Knowing.* Oxford: Oxford University Press, 2007.

Gailey, Jeannine A. and R. Frank Falk (2008). “Attribution of responsibility as a multidimensional concept”. In: *Sociological Spectrum* 28.6, 659–680. ISSN: 15210707. https://doi.org/10.1080/02732170802342958.

Gallagher, Shaun (2006). “The narrative alternative to theory of mind”. In: *Radical Enactivism: Intentionality, Phenomenology and Narrative: Focus on the Philosophy of Daniel D. Hutto*. Ed. by Richard R. Menary. Philadelphia: John Benjamins Publishing Company, 2006, 223–230.

Gawronski, Bertram (2009). “The multiple inference model of social perception: Two conceptual problems and some thoughts on how to resolve them”. In: *Psychological Inquiry* 20.1, 24–29. ISSN: 1047840X. https://doi.org/10.1080/10478400902744261.

Gerber, J. P., Ladd Wheeler, and Jerry Suls (2018). “A social comparison theory meta-analysis 60+ years on”. In: *Psychological Bulletin* 144.2, 177–197. ISSN: 00332909. https://doi.org/10.1037/bul0000127.

Gerstenberg, Tobias and David A. Lagnado (2012). “When contributions make a difference: Explaining order effects in responsibility attribution”. In: *Psychonomic Bulletin and Review* 19.4, 729–736. ISSN: 10699384. https://doi.org/10.3758/s13423-012-0256-4.

— (2014). “Attributing Responsibility: Actual and Counterfactual Worlds”. In: *Oxford Studies in Experimental Philosophy: Volume 1*, 91–130. ISSN: 0873626X. https://doi.org/10.1093/acprof:oso/9780198718765.003.0005. http://www.oxfordscholarship.com/view/10.1093/acprof:oso/9780198718765.001.0001/acprof-9780198718765-chapter-5.

Gerstenberg, Tobias et al. (2018). “Lucky or clever? From expectations to responsibility judgments”. In: *Cognition* 177.December 2017, 122–141. ISSN: 18737838. https://doi.org/10.1016/j.cognition.2018.03.019. https://doi.org/10.1016/j.cognition.2018.03.019.

Ginet, Carl (2000). “The Epistemic Requirements for Moral Responsibility”. In: *Philosophical Perspectives* 14, 267–277.

Goldie, Peter (2009). “Narrative thinking, emotion, and planning”. In: *Journal of Aesthetics and Art Criticism* 67.1, 97–106. ISSN: 00218529.

— (2012). *The Mess Inside: Narrative, Emotion, & the Mind.* Oxford: Oxford University Press, 2012.

Griffith, Meghan (2019). “Narrative Capacity and Moral Responsibility”. In: October, 1–31.

Guerrero, Alexander (2017). “Intellectual Difficulty and Moral Responsibility”. In: *Responsibility: The Epistemic Condition.* Ed. by Philip Robichaud and Jan Willem Wieland. Oxford: Oxford University Press, 2017. Chap. 11.

Guglielmo, Steve and Bertram F. Malle (2017). “Information-Acquisition Processes in Moral Judgments of Blame”. In: *Personality and Social Psychology Bulletin* 43.7, 957–971. ISSN: 15527433. `https://doi.org/10.1177/0146167217702375`.

Haji, Ishtiyaque (1997). “An Epistemic Dimension of Blameworthiness”. In: *Philosophy and Phenomenological Research* 57.3, 523–544.

Harman, Elizabeth (2011). “Does Moral Ignorance Exculpate?” In: *Ratio* 24.4, 443–468.

— (2015). “The Irrelevance of Moral Uncertainty”. In: *Oxford Studies in Metaethics 10*, 53–79. `https://doi.org/10.1093/acprof:oso/9780198738695.003.0003`.

Hart, H. L. A. and A. M. Honoré (1959). *Causation in the law.* Oxford: Oxford University Press, 1959.

Haslanger, Sally (2012). *Resisting Reality.* 2012. ISBN: 9780199892624. `https://doi.org/10.1093/acprof:oso/9780199892631.001.0001`.

Hechler, S. and T. Kessler (2018). “Talking to Others: The importance of responsibility attributions by observers”. In: *Behavioral and Brain Sciences* 41.E46.

Heider, F. (1958). *The psychology of interpersonal relations.* New York: Wiley, 1958.

Hieronymi, Pamela (2001). “Articulating an Uncompromising Forgiveness”. In: *Philosophy and Phenomenological Research* 62.3, 529–555. https://doi.org/10.2307/2653535. http://www.jstor.org/stable/2653535%5Cnhttp://www.jstor.org/stable/pdfplus/2653535.pdf.

— (2019). “I’ll Bet You Think This Blame is About You”. In: *Oxford Studies in Agency and Responsibility Volume 5*, 60–87. https://doi.org/10.1093/oso/9780198830238.003.0004.

Holroyd, Jules (2012). “Responsibility for Implicit Bias”. In: *Journal of Social Philosophy* 43.3, 274–306.

— (2018). “Two Ways of Socializing Moral Responsibility”. In: *Social Dimensions of Moral Responsibility*, 137–162.

Hutto, Daniel (2006). “Narrative practice and understanding reasons”. In: *Radical Enactivism: Intentionality, Phenomenology and Narrative: Focus on the Philosophy of Daniel D. Hutto*. Ed. by Richard R. Menary. Philadelphia: John Benjamins Publishing Company, 2006, 231–248.

— (2012). *Folk psychological narratives: The sociocultural basis of understanding reasons*. Cambridge, MA: MIT Press, 2012.

Hutto, Daniel D (2016). “Narrative self-shaping : a modest proposal”. In: February 2014, 21–41. https://doi.org/10.1007/s11097-014-9352-4.

Jefferson, Anneli (2020). In: *Topoi* 39.1, 219–227.

Kahneman, D. and A. Tversky (1982). “The simulation heuristic”. In: *Judgment under uncertainty: Heuristics and biases*. Ed. by D. Khaneman, P. Slovic, and A. Tversky. New York: Cambridge University Press, 1982, 201–208.

Kelley, H.H. (1973). “The processes of causal attribution”. In: *American Psychologist* 28, 107–128.

Kim, Brian and Matthew McGrath, eds. (2018). *Pragmatic encroachment in epistemology*. 2018, 1–216. ISBN: 9781351685252. https://doi.org/10.1017/9781108992985.005.

Knobe, J. and B. Fraser (2008). “Causal judgments and moral judgment: Two experiments”. In: *Moral psychology, volume 2: The cognitive science of morality.* Ed. by Walter Sinnott-Armstrong. Cambridge, MA: MIT Press, 2008, 441–447.

Knobe, Joshua (2010). “Person as scientist, person as moralist”. In: *Behavioral and Brain Sciences* 33, 315–365.

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

Krueger, Joachim (2000). “The Projective Perception of the Social World”. In: *Handbook of Social Comparison.* Ed. by Jerry Suls and Ladd Wheeler. New York: Kluwer Academic Publishers, 2000.

Lagnado, David A. and Shelley Channon (2008). “Judgments of cause and blame: The effects of intentionality and foreseeability”. In: *Cognition* 108.3, 754–770. ISSN: 00100277. https://doi.org/10.1016/j.cognition.2008.06.009.

Lamarque, Peter (2004). “On not expecting too much from narrative”. In: *Mind and Language* 19.4, 393–408. ISSN: 02681064.

Levy, Neil (2007). “Doxastic Responsibility”. In: *Synthese* 155.1, 127–155.

Lewis, Christopher (2016). “Inequality, incentives, criminality, and blame”. In: *Legal Theory* 22.2, 153–180. ISSN: 14698048. https://doi.org/10.1017/S1352325217000052.

Malle, Bertram F., Steve Guglielmo, and Andrew E. Monroe (2014). “A Theory of Blame”. In: *Psychological Inquiry* 25.2, 147–186. ISSN: 1047840X. https://doi.org/10.1080/1047840X.2014.877340.

Manne, Kate (2018). *Down Girl: The Logic of Misogyny.* Oxford: Oxford University Press, 2018. ISBN: 978-0-19-060498-1.

Mazella, R and A Feingold (1994). “The effects of physical attractiveness, race, socioeconomic status, and gender of defendants and victims of judgments of mock jurors: a meta analysis”. In: *Journal of Applied Social Psychology* 24, 1315–1338.

McGeer, Victoria (2012). “Co-reactive attitudes and the making of moral community”. In: *Emotions, Imagination, and Moral Reasoning* 1974, 299–326. https://doi.org/10.4324/9780203803134.

— (2019). “Scaffolding agency : A proleptic account of the reactive attitudes”. In: *European Journal of Philosophy* March 2018, 301–323. https://doi.org/10.1111/ejop.12408.

McGreer, Victoria (2013). “Civilizing Blame”. In: *Blame: Its Nature and Norms*. Ed. by D.J. Coates and N.A. Tognazzini. New York: Oxford University Press, 2013, 162–188.

McIntosh, Peggy (1988). “White Privilege and Male Privilege: A Personal Account of Coming to See Correspondences through Work in Women’s Studies”. 1988.

McKenna, Michael (2012). *Conversation and Responsibility*. New York: Oxford University Press, 2012.

— (2018). “Power, Social Inequities, and the Conversational Theory of Moral Responsibility”. In: *Social Dimensions of Moral Responsibility*. Ed. by Katrina Hutchison, Catriona Mackenzie, and Marina Oshana. Oxford: Oxford University Press, 2018. Chap. 1, 38–58.

Mele, Alfred (2010). “Moral Responsibility for Actions: Epistemic and Freedom Conditions”. In: *Philosophical Explorations* 13.2, 101–111.

Mele, Alfred R. (2019). *Manipulated Agents*. Oxford: Oxford University Press, 2019. isbn: 9780190927998.

Menge, Torsten (2020). “Fictional expectations and the ontology of power”. In: *Philosophers Imprint* 20.29, 1–22. issn: 1533628X.

Mill, John Stuart (1909). *On Liberty*. New York: P.F. Collier / Son, 1909. isbn: 9781775410652.

Mills, Charles W. (1997). *The Racial Contract*. 1997.

— (2017). *Black Rights/ White Wrongs*. Oxford: Oxford University Press, 2017.

Mitchell, T.L. et al. (2005). “Racial bias in mock juror decision-making: a meta-analytic review of defendant treatment”. In: *Law and Human Behavior* 29, 621–637.

Morgan, Mary S. and M. Norton Wise (2017). “Narrative science and narrative knowing. Introduction to special issue on narrative science”. In: *Studies in History and Philosophy of Science Part A* 62, 1–5. ISSN: 18792510.

Moss-Racusin, C. a. et al. (2012). “Science faculty’s subtle gender biases favor male students”. In: *Proceedings of the National Academy of Sciences* 109.41, 16474–16479. ISSN: 0027-8424. https://doi.org/10.1073/pnas.1211286109.

Nadler, Janice (2012). “Blaming as a social process: The influence of character and moral emotion on blame”. In: *Law and Contemporary Problems* 75.2, 1–31. ISSN: 00239186. https://www.law.northwestern.edu/faculty/fulltime/nadler/Nadler_LCP_2012.pdf.

Nadler, Janice and Mary Hunter Mcdonnell (2012). “Moral character, motive, and the psychology of blame”. In: *Cornell Law Review* 97.2, 255–304. ISSN: 00108847.

Nelkin, Dana Kay (2014). “Difficulty and Degrees of Moral Praiseworthiness and Blameworthiness”. In: *Nous* 50, 356–378.

Nolfi, Kate (2018). “Another Kind of Pragmatic Encroachment”. In: *Pragmatic Encroachment in Epistemology*. Ed. by Brian Kim and Matthew McGrath. New York: Routledge, 2018, 35–55. ISBN: 9781351685252. https://doi.org/10.4324/9781315168197-11.

Oshana, Marina (2018). “Ascriptions of Responsibility Given Commonplace Relations of Power”. In: *Social Dimensions of Moral Responsibility*. Ed. by Katrina Hutchison, Catriona Mackenzie, and Marina Oshana. Oxford: Oxford University Press, 2018. Chap. 3, 81–110.

Pereboom, Derk (2001). *Living Without Free Will*. Cambridge: Cambridge University Press, 2001. https://doi.org/10.5840/faithphil200219330.

Pereboom, Derk (2014). *Free Will, Agency, and Meaning in Life*. Oxford: Oxford University Press, 2014. ISBN: 9788578110796. https://doi.org/10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

Pickard, Hanna (2011). "Responsibility Without Blame: Empathy and the Effective Treatment of Personality Disorder". In: *PPP* 18.3.

— (2013). "Responsibility Without Blame: Philosophical Reflections on Clinical Practice". In: *Cure and Care* 2010, 1134–1154.

Pizarro, David A. and David Tannenbaum (2012). "Bringing Character Back: How the Motivation to Evaluate Character Influences Judgments of Moral Blame". In: *The Social Psychology or Morality: Exploring the causes of good and evil.* Ed. by Mario Mikulincer and Philip R. Shaver. Washington, DC: American Psychological Association, 2012. Chap. 5, 91–108. ISBN: 9788578110796. https://doi.org/10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

Rahimi, Sonia, Nathan C. Hall, and Timothy A. Pychyl (2016). "Attributions of responsibility and blame for procrastination behavior". In: *Frontiers in Psychology* 7.AUG, 1–7. ISSN: 16641078. https://doi.org/10.3389/fpsyg.2016.01179.

Rai, Tage Shakti and Alan Page Fiske (2011). "Moral Psychology Is Relationship Regulation: Moral Motives for Unity, Hierarchy, Equality, and Proportionality". In: *Psychological Review* 118.1, 57–75. ISSN: 0033295X. https://doi.org/10.1037/a0021867.

Railton, Peter (2014). "The Affective Dog and Its Rational Tale : Intuition and Attunement *". In: *Ethics* 124.4, 813–859. ISSN: 00141704. https://doi.org/10.1086/675876.

Reis-Dennis, Samuel (2021). "Rank Offence: The Ecological Theory of Resentment". In: *Mind* 00, 1–19. ISSN: 0026-4423. https://doi.org/10.1093/mind/fzab006.

Roth, Paul A. (2017). "Essentially narrative explanations". In: *Studies in History and Philosophy of Science Part A* 62, 42–50. ISSN: 18792510.

Scaife, Robin et al. (2020). “To Blame? the Effects of Moralized Feedback on Implicit Racial Bias”. In: *Collabra: Psychology* 6.1, 1–12. ISSN: 24747394. `https://doi.org/10.1525/collabra.251`.

Scanlon, T. M. (2008). *Moral Dimmensions*. Cambridge, MA: Belknap Press of Harvard University Press, 2008. ISBN: 9788578110796. `https://doi.org/10.1017/CBO9781107415324.004`. arXiv: `arXiv:1011.1669v3`.

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

Schechtman, Marya (2011). “The Narrative Self”. In: *The Oxford Handbook of the Self* February. ISSN: 00218308. `https://doi.org/10.1093/oxfordhb/9780199548019.003.0018`.

Shaver, Kelly G. (1985). *The Attribution of Blame: Causality, Responsibility, and Blameworthiness*. New York: Springer, 1985. ISBN: 9781461295617. `https://doi.org/10.1007/978-1-4612-5094-4`.

Shaw, Elizabeth, Gregg Caruso, and Derk Pereboom, eds. (2019). *Free Will Skepticism in Law and Society: Challenging Retributive Justice*. Cambridge: Cambridge University Press, 2019. `https://doi.org/10.1017/9781108655583`.

Sher, George (2009). *Who Knew? Responsibility without Awareness*. New York: Oxford University Press, 2009.

Shoemaker, David and Manuel Vargas (2019). “Moral torch fishing : A signaling theory of blame”. In: *Nous* July, 1–22. `https://doi.org/10.1111/nous.12316`.

Smiley, Marion (2016). “Volitional excuses, self-narration, and blame”. In: *Phenomenology and Cognitive Science* 15, 85–101. `https://doi.org/10.1007/s11097-014-9367-x`.

Smith, Adam (1976). *The Theory of Moral Sentiments*. Ed. by A.L. MacFie and D.D. Raphael. The Glasgo. Oxford: Oxford University Press, 1976. ISBN: 9781626239777.

Smith, Richard (2000). “Assimilative and Contrastive Emotional Reactions to Upward and Downward Social Comparisons”. In: *Handbook of Social Comparison*.

Ed. by Jerry Suls and Ladd Wheeler. New York: Kluwer Academic Publishers, 2000.

Sommers, S.R. and P.C. Ellsworth (2000). "Race in the courtroom: perceptions of guilt and dispositional attributions". In: *Pers. Social Psychology Bulletin* 26, 1367–1379.

Spaulding, Shannon (2018). *How We Understand Others*. Routledge, 2018. ISBN: 9781138221581. https://doi.org/10.4324/9781315396064.

Stanley, Jason (2005). *Knowledge and Practical Interests*. Oxford: Oxford University Press, 2005.

Strawson, Peter (1989). "Freedom and Resentment". In: *Free Will*. Ed. by Gary Watson. Oxford: Oxford University Press, 1989. Chap. 4, 72–93. ISBN: 9788578110796. https://doi.org/10.1017/CBO9781107415324.004. arXiv: arXiv:1011.1669v3.

— (2003). "Freedom and Resentment". In: *Free Will*. Ed. by Gary Watson. Oxford: Oxford University Press, 2003. Chap. 4, 72–93.

Suedfeld, Peter et al. (1985). "Ascription of Responsibility as a Personality Variable". In: *Journal of Applied Social Psychology* 15.3, 285–311. ISSN: 15591816. https://doi.org/10.1111/j.1559-1816.1985.tb00902.x.

Suls, Jerry, René Martin, and Ladd Wheeler (2002). "Social comparison: Why, with whom, and with what effect?" In: *Current Directions in Psychological Science* 11.5, 159–163. ISSN: 09637214. https://doi.org/10.1111/1467-8721.00191.

Suls, Jerry and Ladd Wheeler, eds. (2000). *Handbook of Social Comparison: Theory and Research*. New York: Springer Science+Business Media, LLC, 2000. ISBN: 9781787284395.

Táíwò, Olúfmi (2020). "Being-in-the-room Privilege: Elite Capture and Epistemic Deference". In: *The Philosopher* 108.4.

Talbert, Matthew (2022). “Moral Responsibility”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Fall 2022. Metaphysics Research Lab, Stanford University, 2022.

Tilton, Emily C. R. (2023). “"That’s Above My Paygrade": Woke Excuses for Ignorance”. In: *Philosophers’ Imprint*.

Tognazzini, Neal A. (2015). “The Strains of Involvement”. In: *The Nature of Moral Responsibility*. Ed. by Randolph Clarke, Michael McKenna, and Angela Smith. Oxford: Oxford University Press, 2015, 19–45.

Vargas, Manuel (2005). “The Trouble With Tracing”. In: *Midwest Studies in Philosophy* 29.

— (2013). *Building better beings: A theory of moral responsibility*. Oxford University Press, 2013.

— (2018a). “Revisionism”. In: *A Companion to Free Will*. Ed. by Joseph Campbell. 2018, 1–23.

— (2018b). “The Social Constitution of Agency and Responsibility”. In: *Social Dimensions of Moral Responsibility*. Ed. by Katrina Hutchison, Catriona Mackenzie, and Marina Oshana. Oxford: Oxford University Press, 2018. Chap. 4, 110–136. ISBN: 9780190609610.

Velleman, J David (2009). *How We Get Along*. Cambridge: Cambridge University Press, 2009.

Vitale, Alex S. (2018). *The End of Policing*. Verso, 2018.

Waller, Bruce N. (2014). *The stubborn system of moral responsibility*. 2014, 1–294. ISBN: 9780262327404. `https://doi.org/10.1080/10848770.2018.1430928`.

Wang, Tinghao (2021). “The Communication Argument and the Pluralist Challenge”. In: *Canadian Journal of Philosophy* 51.5, 384–399. `https://doi.org/10.1017/can.2021.30`.

— (2022). “How Blame Functions: Essays on Blame, Responsiblity, and Moral Emotions”. PhD thesis. University of California San Diego, 2022.

Watson, Gary (1996). “Two faces of responsibility”. In: *Philosophical Topics* 24.2, 227–248.

Weiner, Bernard (2006). *Social Motivation, Justice, and the Moral Emotions: An Attributional Approach*. Mahwah, New Jersey: Lawrenec Erlbaum Associates, 2006. ISBN: 0805855262.

Westra, Evan (2018a). “Character and theory of mind: an integrative approach”. In: *Philosophical Studies* 175.5, 1217–1241. ISSN: 15730883. https://doi.org/10.1007/s11098-017-0908-3.

— (2018b). “Character and theory of mind: an integrative approach”. In: *Philosophical Studies* 175.5, 1217–1241. ISSN: 15730883. https://doi.org/10.1007/s11098-017-0908-3.

— (2019). “Stereotypes, theory of mind, and the action–prediction hierarchy”. In: *Synthese* 196.7, 2821–2846. ISSN: 15730964. https://doi.org/10.1007/s11229-017-1575-9.

Wiggleton-Little, Jada (2023). “Let Me Have Your Attention! Taking Pain Utterances Seriously”. PhD thesis. University of California San Diego, 2023.

Willemsen, Pascale, Albert Newen, and Kai Kaspar (2018). “A new look at the attribution of moral responsibility: The underestimated relevance of social roles”. In: *Philosophical Psychology* 31.4, 595–608. ISSN: 1465394X. https://doi.org/10.1080/09515089.2018.1429592. http://doi.org/10.1080/09515089.2018.1429592.

Wolf, Susan (2011). “Blame, Italian Style”. In: *Reason and Recognition: Essays on the Philosophy of T.M. Scanlon*. Ed. by R. Jay Wallace, Rahul Kumar, and Samuel Freeman. New York: Oxford University Press, 2011, 332–347.

Wylie, Alison (2003). “Why Standpoint Matters”. In: *Science and Other Cultures: Issues in Philosophies of Science and Technology*. Ed. by Robert Figueroa and Sandra G. Harding. Routledge, 2003, 26–48.

Yancy, George (2012). *Look, A White!: Philosophical Essays on Whiteness*. Temple University Press, 2012.

Young, Iris Marion (1990). *Justice and the Politics of Difference*. Princeton: Princeton University Press, 1990. ISBN: 0691078327. https://doi.org/10.1177/1522637916656379.

Zhao, Jun and Christabel L. Rogalin (2017). "Heinous Crime or Unfortunate Incident: Does Gender Matter?" In: *Social Psychology Quarterly* 80.4, 330–341. ISSN: 01902725. https://doi.org/10.1177/0190272517728923.