UNIVERSITY OF CALIFORNIA SAN DIEGO

Recklessness and Responsibility: A Theory of the Epistemic Dimension

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

in

Philosophy

by

William Albuquerque

Committee in charge:

Professor David Brink, Co-Chair
Professor Dana Nelkin, Co-Chair
Professor Christine Harris
Professor Manuel Vargas
Professor Caren Walker
Professor Monique Wonderly

2024

The Dissertation of William Albuquerque is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2024

DEDICATION

I dedicate my dissertation to my mother Melanie, father Gilberto, sister Emma, and partner Tara.

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to acknowledge the support and guidance of the many people who contributed to the work in this dissertation. First and foremost are my two co-chairs: David Brink and Dana Nelkin. When I was first reading through the literature on the epistemic dimension and trying to form some initial ideas, I would meet with both David and Dana to discuss various papers and views that I had come across. They generously gave me time to bounce ideas off them and provided commentary that guided my early thinking. When I then started to put my ideas to writing, their feedback was invaluable. Each time I asked them to look at a draft, I received extremely detailed comments. Often, they would set aside time to discuss these comments in person or over Zoom. These exchanges had a substantial impact on my theory of the epistemic dimension. While writing, I often thought about how fortunate I was to be working with two of the best philosophers working in moral responsibility.

Another person that significantly shaped the dissertation is Manuel Vargas. Manuel was instrumental in helping me see the big picture regarding certain key issues in moral responsibility. Whenever we would meet, I would try to take copious notes about whichever dialectic or framework he was describing because it was so useful in reframing my thoughts. As with Dana and David, I also have deep respect for Manuel's work; and as his own approach to moral responsibility differs from David and Dana's in important ways, I feel that I gained considerably from engaging with his perspective.

I would also like to acknowledge my other committee members: Monique Wonderly, Caren Walker, and Christine Harris. Monique came to UCSD part way through my time in the program, but I immediately connected with her thoughts about the emotions and their relevance for moral responsibility. I always found our talks very rewarding and appreciated her feedback

regarding my work. Because I was previously focused on work regarding moral psychology, I took a seminar with both Caren and Christine in the psychology department. Although my dissertation didn't end up centering on moral psychology, those two seminars were some of my favorites during my coursework. As Caren has a philosophy background, she was able to discuss the intersection of philosophy and developmental psychology at a very high level. Moreover, she asked valuable questions at my prospectus defense. I took Christine's seminar remotely during the early days of COVID-19, and it was a highlight of my week. She allowed me to ask many questions about the emotions literature and test some thoughts I had developed from thinking about moral emotions. If I go back to working on moral psychology and the philosophy of emotions, I will certainly draw on what I learned from her.

VITA

2011 Bachelor of Arts in Philosophy and Biology, Bowdoin College

2016 Master of Arts in Philosophy, University of Wisconsin-Milwaukee

2024 Doctor of Philosophy in Philosophy, University of California San Diego

ABSTRACT OF THE DISSERTATION

Recklessness and Responsibility: A Theory of the Epistemic Dimension

by

William Albuquerque

Doctor of Philosophy in Philosophy

University of California San Diego, 2024

Professor David Brink, Co-Chair
Professor Dana Nelkin, Co-Chair

Most people acknowledge that epistemic considerations matter for moral responsibility. For example, it matters for responsibility whether a pharmacist knowingly or unknowingly gave a customer the wrong medication. It might be that even though the pharmacist made an unwitting error, she's still responsible for her wrongdoing – but we would want to know more about the nature and etiology of her epistemic state. Despite this near universal recognition that these considerations matter, however, the epistemic dimension of moral responsibility has traditionally taken a backseat to issues regarding free will. The purpose of this dissertation is to better understand how this epistemic dimension influences responsibility and construct a core theory.

In chapter one, I introduce the epistemic dimension and address some important issues that aren't central to my project before moving on to the main project at hand. In chapter two, I evaluate the prospects for *attributionism*, an approach to moral responsibility that grounds responsibility in an agent's *evaluative orientation*. In chapter three, I then consider a rival approach, *capacitarianism*, which emphasizes the *capacities* of an agent. In chapter four, I expand on the awareness of risk condition that forms the foundation of my account of the epistemic dimension. Finally, in chapter five, I address remaining issues regarding my account of the epistemic dimension.

The centerpiece of the dissertation is my theory of the awareness of risk condition. I argue that in order to be blameworthy for some outcome, the agent must have: (1) an *occurrent* belief of the general risk of her conduct, and (2) a *disposition to believe* the specific risk of her conduct. As far as the occurrent belief, the agent must entertain a belief about the riskiness of her actions at the time of wrongdoing, but this doesn't mean she must *explicitly* entertain that belief. As for the disposition to believe, I articulate a notion whereby such dispositions are grounded in other beliefs and perceptions and require some mediating process to form the relevant belief.

THE EPISTEMIC DIMENSION

Introduction

Consider the following scenario. Jack is a pharmacist filling a prescription for Michelle. As it turns out, the wrong medicine was put in the wrong capsules, and those capsules were placed in the corresponding container. When Jack fills Michelle's prescription, then, he unwittingly gives her the wrong medication. As a result, Michelle suffers a minor seizure.

Now, obviously Michelle was wronged by being given the incorrect medication. Yet it's far from obvious that Jack is morally responsible for this wrongdoing, despite being the most proximate cause. Note, though, that Jack plausibly acts freely in giving Michelle the wrong medication. No one coerced Jack into filling the subscription, for instance, and he didn't act under some sort of irresistible compulsion. If Jack is excused for harming Michelle, then, it isn't because his free agency was undermined or interfered with, it's because of the nature of his ignorance. In this way, there clearly seems to be an *epistemic* dimension to moral responsibility.

Naturally, the observation that epistemic considerations can influence culpability isn't novel to either philosophy or legal theory. Yet, particularly in work on moral responsibility, these epistemic considerations have traditionally played a secondary role to other issues, such as the nature and significance of free will. More recently, though, a growing number of theorists have come to recognize the central importance of the epistemic dimension. The purpose of this dissertation is to better understand how this epistemic dimension influences moral responsibility. My basic approach is to first explore how prominent theories of moral responsibility – not usually developed with epistemic considerations at the forefront – might capture the epistemic dimension. Based on the inadequacies of these theories, I then attempt to construct an independent epistemic requirement that emphasizes the importance of *awareness of risk*.

In chapter two, I evaluate the prospects for *attributionism*, an approach to moral responsibility that grounds responsibility in an agent's *evaluative orientation*.[1] Ultimately, although attributionism has significant explanatory power, I argue that it appears extensionally inadequate in capturing intuitions (or considered judgments) that track the influence of certain epistemic factors and can gain extensional adequacy only by drawing on resources that undermine its foundational commitments. In chapter three, I then consider a rival approach, *capacitarianism*, which emphasizes the *capacities* of an agent.[2] Although capacitarianism technically straddles multiple theories of moral responsibility, it's strongly related to the *reasons-responsiveness* tradition. I contend that although a reasons-responsiveness approach better captures and explains the influence of epistemic considerations, it must be supplemented by additional constraints. Specifically, I argue that awareness of risk is a necessary condition on blameworthiness because such awareness is required for an agent to possess the *fair opportunity to avoid wrongdoing* that grounds moral responsibility.[3] Because such a fair opportunity best explains why capacities are relevant to responsibility, I understand the awareness of risk condition as a natural extension of a reasons-responsiveness theory.

In chapter four, I expand on the awareness of risk condition that forms the foundation of my account of the epistemic dimension. In constructing this account, I rely on pertinent work in the philosophy of law regarding the concept of *recklessness*, which can be found as a class of elemental *mens rea* in the criminal law.[4] Broadly speaking, criminal recklessness involves

---

[1] Attributionist accounts include Adams (1985), Arpaly (2002), Scanlon (1998), Sher (2006), and Smith (2005). Neil Levy (2005) originally coined the term to refer to these types of views.
[2] Capacitarian views include Clarke (2017), Murray and Vargas (2020), Rudy-Hiller (2017), and Sher (2009).
[3] The basic idea that blame (or punishment) requires that the agent has a fair opportunity to avoid the relevant wrongdoing can be found in a number of philosophical and legal works, including Brink (2021), Brink and Nelkin (2013), Hart (1968a), and Moore (1997).
[4] Elemental *mens rea* (or guilty mind) refers to the mental elements of an offense. It is contrasted with *actus reus* (or guilty act), representing the objective or material elements of an offense (conduct, result, and attendant circumstances). The Model Penal Code (1985) lists four culpable classes of mental state: purpose, knowledge,

awareness of an unjustifiable risk, and so discussions in legal theory regarding recklessness help inform my requirement on moral responsibility. Ultimately, my view is a mixed account that recognizes different requirements at different cognitive levels. According to this theory, to be blameworthy for some outcome the agent must have a certain kind of belief regarding the *general* risk of her conduct, and a *disposition* to believe the specific risk of her conduct. Importantly, this constraint is less subjective than it might initially appear. First, the relevant notion of risk at play is determined by the perspective of a *reasonable person* with the agent's available evidence, not the agent's own assessment of risk. Secondly, the agent needn't believe that the risk she's imposing is unjustified in order to be blameworthy, as long as the risk is *actually* unjustified.

In chapter five, I address remaining issues regarding my account of the epistemic dimension. Some of these issues involve specific kinds of cases, whereas others involve broader questions regarding moral and legal responsibility. One specific kind of case that is relevant to the epistemic dimension is *pure epistemic recklessness*. In these rare cases, agents are aware that their actions risk *ignorance* but unaware that this ignorance consequently risks harm. Applying my theory of the awareness of risk condition, I argue that although we might criticize such agents for their epistemic practices, they aren't blameworthy for their ignorant wrongdoing.

Finally, it's important to note that on any plausible theory of moral responsibility, epistemic considerations strongly interact with other dimensions of responsibility. Thus, any account that primarily focuses on the influence of epistemic considerations will necessary be incomplete. I accept this limitation of my dissertation. My goal is to build enough of a theory of the epistemic dimension to provide a framework for answering lingering questions, including

---

recklessness, and negligence. An agent acts recklessly when she "consciously disregards a substantial and unjustifiable risk that the material element exists or will result from his conduct" (§ 2.02(2)(c)).

those involving the interaction of various components within the best theory of responsibility. Indeed, I believe that no simple theory can capture the influence of epistemic considerations without significant revisionism. I aim to avoid such revisionism wherever possible and best capture our intuitions, despite how disordered they may sometimes appear.

Before beginning the work of chapter two – investigating how prominent theories of moral responsibility can capture the epistemic dimension – there are a few topics worth discussing. Again, my goal in this dissertation isn't to answer every question one could have about the epistemic dimension; even if my theory *could* provide an answer, I won't weigh in on every debate. Still, there are some important issues that I want to address before moving on to the main project at hand. This introductory work is important for setting the stage for the rest of the discussion, but it also serves to demonstrate that I'm aware of certain issues and provides space to explain why they're not central to the rest of the dissertation. In the next section, I temporarily suspend my focus on the epistemic dimension and discuss the basic nature and motivations of the main theories of moral responsibility that I will later critique. This context is important for evaluating these theories as they relate to the epistemic dimension.

1. Theories of Moral Responsibility

As previously mentioned, epistemic considerations have historically played a secondary role in theorizing about moral responsibility. The nature and significance of free will has instead been the dominant focus of this work. This is not to say that those focusing on free will were necessarily unaware of the influence of epistemic considerations, only that these considerations were often set aside.[5] But is it even possible to neatly set aside the epistemic dimension of

[5] John Fischer and Mark Ravizza's (1998) theory is a notable example of work that explicitly distinguishes between epistemic and free will considerations and claims to exclusively focus on free will.

responsibility? Fernando Rudy-Hiller (2022) explains that "philosophers usually acknowledge two individually necessary and jointly sufficient conditions for a person to be morally responsible for an action…a control condition (also called freedom condition) and an epistemic condition (also called knowledge, cognitive, or mental condition)." Yet, even if it's in principle possible to separate these strands of responsibility, I believe that the most plausible theory understands them as at least strongly interacting factors. For this reason, I eschew the notion of a separate epistemic condition in favor of the concept of the epistemic *dimension*. Still, there's no denying that many theories of moral responsibility were developed with free will considerations at the forefront. In this section, I will briefly explain the "freedom condition" and how it motivates different theories of responsibility.

Perhaps unsurprisingly, there's no universally applied definition of the freedom (or control) condition, but a prominent conception understands it simply as "whatever ability is required of a morally responsible agent to have sufficient control over her culpable conduct that she would deserve blame for it" (McKenna, 2022, p. 28).[6] It's important to note that the notion of freedom here is explicitly linked to moral responsibility, whereas in other contexts the target notion is conceptually prior to the issue of responsibility. Still, even if these two concepts of freedom are technically distinct, there's no denying that they're at least closely related. Because of this, we can understand much of the theorizing about the nature and significance of free will to be relevant to the freedom condition, even if the actual focus of some of this theorizing isn't specifically moral responsibility.

[6] At least, this is a prominent conception of the *internal* dimension of the freedom condition. Although most accounts of the freedom condition focus on features of the agent herself, moral responsibility also requires certain *external* conditions, such the agent being free from coercion. These external conditions allow the agent to exercise the necessary capacities to act freely. In this proceeding discussion, I will set aside the external dimension of the freedom condition, given that the details are much less disputed.

With the previous caveat in place, I can now discuss the basic structure of the main theories of moral responsibility, and how they understand the freedom condition. Although there is significant variation in how individual accounts understand the freedom condition, I will group theories into two broad categories: *reasons-responsiveness* and *attributionist* views. Furthermore, I will focus on *compatibilist* theories that assert that their given freedom condition is compatible with determinism.[7] Of course, there are many other ways to categorize theories of moral responsibility, depending on the features one wants to highlight and contrast. For current purposes, though, my focus is on freedom (or control), and whether moral responsibility requires some form of reasons-responsiveness.

1.1 Broadly speaking, reasons-responsiveness theories "explain exercises of free agency in terms of responsiveness or sensitivity to reasons" (McKenna, 2022, p. 27), especially *moral* reasons. Crucially, this central notion of responsiveness is a *modal* feature of agents,[8] meaning that it's concerned with an agent's *ability* to recognize and respond to reasons, not whether the agent actually exercises this ability. For example, Scott plausibly satisfies the freedom condition on a reasons-responsiveness theory insofar as he possessed the ability to recognize and respond to the fact that his comments were wrongfully hurtful, even if he failed to exercise this ability in the moment.[9] Key to evaluating the required possession of these abilities are relevant *counterfactuals* in which the pertinent reasons are present.[10] Consider, for instance, relevantly similar situations where there are reasons for Scott to avoid making these hurtful comments.

[7] By (causal) determinism, I mean the thesis that every event is necessitated by the past and laws of nature.
[8] For ease of exposition, I will discuss reasons-responsiveness as a property of agents. Crucially, though, Fischer and Ravizza's (1998) prominent account understands reasons-responsiveness as a property of sub-agential *mechanisms*. Although the distinction between agent-based and mechanism-based views is significant in certain contexts, construing reasons-responsiveness as a property of agents shouldn't matter for the purposes of this section.
[9] Assuming that nothing unduly interfered with Scott exercising this ability.
[10] The range of counterfactuals (usually understood in terms of *possible worlds*) depends on the range of reasons, which in turn depends on whether the assessed reasons-responsive ability is more *general* or *specific*.

Does Scott recognize these reasons and respond accordingly in these scenarios?[11] If so, then reasons-responsiveness theories maintain that he had sufficient freedom (or control) in his *actual* situation.[12]

The reasons-responsiveness conception of the freedom condition has a number of attractive features. First, reasons-responsiveness seems to be an essential property of *persons*, or at least responsible persons. Not only do animals plausibly lack the requisite reasons-responsiveness, but so do many small children that we wouldn't hold responsible. In this way, reasons-responsiveness theories effectively categorize and explain the kind of free agents that are of interest when theorizing about moral responsibility.[13]

Secondly, reasons-responsiveness is both extensionally and explanatorily powerful, meaning that it can both capture and explain intuitive judgments regarding moral responsibility. Consider some paradigmatic types of non-responsible agents: those who are insane, immature, or suffer from irresistible desires. All these agents intuitively fail the freedom condition on a reasons-responsiveness view, and the fact that they're insensitive to certain reasons seems like a credible explanation for why they lack sufficient freedom. The same is true for otherwise responsible agents whose freedom is undermined by external factors, such as coercive forces. Such coercion plausibly excuses these agents because it interferes sufficiently with the exercise of their reasons-responsive abilities.[14]

[11] More accurately, does he recognize/respond in a sufficient *proportion* of these scenarios (or possible worlds)? On a more scalar account, his *degree* of freedom would be relative to this proportion.
[12] As I will discuss shortly, some reasons-responsiveness theories require the ability to do otherwise. On this view, freedom includes the ability to *access* the alternative scenarios that determine reasons-responsiveness.
[13] Brink and Nelkin (2013) and McKenna (2022) both note this feature of reasons-responsiveness.
[14] There's disagreement among theorists if external factors like coercion undermine reasons-responsiveness itself or just the exercise of this ability. Regardless, though, all reasons-responsiveness theories accept that these factors can undermine freedom (or responsibility) by somehow interfering with reasons-responsiveness.

Finally, reasons-responsiveness is an attractive compatibilist conception of free will. A reasons-responsiveness account of the freedom condition seems *prima facie* compatible with determinism in that the type of abilities it requires aren't contra-causal – meaning that their possession and exercise don't require that determinism is false.[15] However, it's worth noting here that reasons-responsive compatibilism has actually been developed in two main ways, corresponding to two prominent models of freedom. As McKenna (2022) explains the dichotomy:

> Some philosophers explain freedom in terms of alternatives to what an agent does do. These philosophers, *leeway theorists*, focus on the leeway an agent has to do something other than what she does, and her freedom consists, at least in part, in her ability to access this leeway and do something other than what she does. Others, *source theorists*, focus on the actual source of an agent's action and on the etiology pertaining to what she does do. On such an approach, freedom is not explained in terms of what other things an agent might have done or was able to do but on what she did do and what agential ability was manifested in her doing that thing. (pp. 31-2)

Two of the most influential reasons-responsiveness theorists, John Fischer and Mark Ravizza, are both source theorists. Source theories have alleged advantages in responding to certain important arguments and cases in the literature, including those related to compatibilism.[16] However, leeway theorists have developed sophisticated resources to support their own conception of freedom. Either way, the reasons-responsiveness approach represents a powerful compatibilist strategy.

1.2 'Attributionism' designates an even more diverse collection of views, with an obscurer freedom condition. The primary reason for these characteristics is that attributionism is perhaps best defined in terms of the conditions that it *rejects* regarding moral responsibility. Most

[15] It's worth noting that even if reasons-responsiveness is a compatibilist ability, reasons-responsiveness theories are still subject to incompatibilist arguments targeting the *source* of this ability (see, e.g., Pereboom, 2001; and Mele, 2019).

[16] In particular, source theories have a supposed advantage responding to the *Consequence Argument for Incompatibilism* (see, e.g., Ginet, 1966; van Inwagen, 1975).

significantly, attributionist views generally reject conceptions of the freedom condition that understand it in terms of voluntary or reasons-responsive control. As attributionist Matthew Talbert (2022) explains:

> It's often assumed that we are blameworthy only for what is in our control, either in an immediate or a mediated fashion… However, attributionism holds that we are open to blame on account of those things that reflect our objectionable evaluative judgments, and that not all such things are under our immediate control or are associated with prior exercises of control. (p. 59)

On my broader understanding of attributionism, the theory focuses on how actions (or attitudes) express an agent's *evaluative orientation*, whether this orientation is grounded in judgements, desires, or other states. The key is that actions can express an agent's evaluative orientation without most substantive forms of control.

In terms of a positive freedom condition, attributionists resemble source theorists in that they both deny the necessity of leeway freedom and stress the significance of the relation between one's actions and some important agential feature. This similarity is no coincidence, as the two species of views share a common ancestor in the work of Harry Frankfurt (1969, 1971), who influentially argued for his own kind of proto-source account of the freedom condition.[17] Unlike reasons-responsive source theories, however, the relevant relation between action and agential feature isn't specified in merely causal terms for attributionists. Instead, a responsible act *expresses* a certain evaluative orientation, where this notion of expression represents a more

[17] Attributionism arguably draws more directly from the work of Watson (1975, 1996), whose theory builds on Frankfurt's in important ways. Crucially, though, Watson doesn't consider his view to be a theory of moral responsibility as I understand it here (i.e., responsibility as *accountability*), but instead a theory of an antecedent form of moral appraisal (responsibility as *attributability*). Attributionists, however, take themselves to be providing a comprehensive theory of moral responsibility. I discuss these different forms of moral appraisal in the next chapter.

nuanced connection between act and evaluative orientation.[18] Crucially, though, the abilities required to express a certain evaluative orientation are fairly minimal.

Because attributionism requires such a minimal conception of freedom, it's arguably even more plausible that moral responsibility is compatible with determinism on this picture than on a reasons-responsiveness account. After all, not only does attributionism deny the necessity of contra-causal abilities, but it also denies the necessity of most substantive forms of control. Unfortunately, attributionists haven't engaged as much with issues of free will and determinism as reasons-responsiveness theorists, but intuitively it seems an action can express a certain evaluative orientation regardless of whether that action was casually determined. If this is right, then attributionism also represents an attractive compatibilist view. Furthermore, possession of an evaluative orientation seems to be an important property of persons, and so attributionism also identifies a conception of free agency that is theoretically significant.

Furthermore, attributionism is also both extensionally and explanatorily powerful. Consider again paradigmatic types of non-responsible agents: those who are insane, immature, or suffer from irresistible desires. Although these agents aren't plausibly reasons-responsive, they also appear to either lack a (coherent) evaluative orientation or fail to express it in action. Indeed, the agent who succumbs to irresistible desires that she judges to be wrong seemingly acts *contrary* to her evaluative orientation. An appealing explanation for why we don't hold these types of agents responsible is that our responsibility practices and judgments appear principally concerned with agents' evaluative orientation toward us. As P.F. Strawson (1962/1993) influentially argued, a central feature of ordinary interpersonal relationships is "the very great

[18] In Smith's (2005) terms, expression involves a "rational relation" between an act/attitude and an evaluative orientation, which is to say that "that [act/attitude] is, or should be, sensitive to her evaluative judgments and that she therefore can properly be asked to defend or justify it" (p. 267).

importance that we attach to the attitudes and intentions towards us of other human beings …[and] whether the actions of other people – and particularly some people – reflect attitudes towards us of goodwill, affection, or esteem on the one hand and contempt, indifference, or malevolence on the other" (p. 48). If this is right, then it makes sense that moral responsibility might be grounded in how an action expresses a certain evaluative orientation.

Lastly, attributionists claim advantages over other theories in capturing certain responsibility judgments. For instance, Angela Smith notably argues that her attributionist view can vindicate the intuition that agents are responsible for certain *involuntary* states, such as attitudes, desires, and failures to notice or remember certain things. If a parent forgets to pick up her child from school, for example, she may be responsible for her omission even though her forgetting was totally involuntary. Smith's (2005) reasoning is that there's "a rational connection between many of the thoughts and desires that occur to us and the evaluative judgments and commitments we accept" (p. 247). In this way, involuntary features of an agent can nevertheless have expressive evaluative significance that grounds responsibility. Insofar as there's broad agreement that agents can be responsible for these involuntary features and the resultant acts, then, attributionism can claim an extensional advantage over views that fail to support this shared judgment.

Perhaps more controversially, some attributionists argue that *psychopaths* are morally responsible. Now, the precise definition of psychopathy is disputed, but within these discussions psychopaths are generally categorized as agents who are unable to recognize and be appropriately motivated by *moral* reasons. Despite this incapacity, attributionists like Talbert (2008) contend that "there are good reasons to respond to the hurtful, intentional actions of psychopaths in the same way that we respond to the hurtful, intentional actions of more

psychologically normal wrongdoers" (p. 519). Specifically, insofar as psychopaths are able to make judgments about reasons *at all*, their actions can still express an objective evaluative orientation that makes them blameworthy. This conclusion supposedly vindicates our natural reactions toward the wrongdoing of psychopaths. If this is right, then attributionism has an apparent extensional advantage over reasons-responsiveness theories, given that psychopaths lack the relevant reasons-responsive abilities.[19]

2. The Revisionist Argument

Before investigating how these two prominent theories of moral responsibility can capture the epistemic dimension, there are two other topics worth discussing. First, I will discuss a rather influential argument that has shaped much of the literature on the epistemic dimension in recent years. Looking back at the example of Jack the pharmacist, I suggested that he might be excused for his actions based on his ignorance. If this is possible, then ignorance at least sometimes excuses. According to the conclusion of the argument at hand, however, ignorance *always* excuses. This is a revisionist conclusion insofar as its acceptance would require us to revise many of our intuitive judgments about responsibility. In fact, if the requisite ignorance is pervasive enough, the argument entails that agents are almost never morally responsible. So, what is this revisionist argument?

Although there are slightly different versions of the argument, I will focus on Gideon Rosen's (2003) prominent rendering.[20] Rosen begins with the observation that an action done from non-culpable factual ignorance is itself non-culpable. For instance, Jack isn't blameworthy

[19] See, e.g., Nelkin (2015a) for an argument that psychopaths *aren't* morally responsible.
[20] It's worth noting that Rosen (2004) himself has an epistemic version of the argument with the conclusion that "it would be unreasonable to repose much confidence in any particular positive judgment of responsibility" (p. 308). The much more influential version of the argument doesn't have this epistemic character, focusing instead on the metaphysical issue of whether agents are actually responsible for ignorant wrongdoing.

for wronging Michelle on the assumption that he wasn't blameworthy for his ignorance that he was filling the incorrect medication for her prescription. From this observation, Rosen infers that all blameworthiness for factually ignorant wrongdoing is *derivative*, meaning that it must stem from the culpability of the ignorance. Although not uncontroversial,[21] it's a seemingly plausible inference to draw from considering cases like Jack's.

Of course, this derivative structure naturally prompts the question, what does it mean for factual ignorance to be culpable? Rosen's answer is that factual ignorance is culpable only when it is the upshot of *epistemic irresponsibility*.[22] Now, the precise contours of epistemic irresponsibility require further investigation, but the basic notion is fairly clear: such irresponsibility involves the mismanagement of one's epistemic situation. For example, suppose that Jack had read a news article in the morning reporting widespread errors with the medications contained in certain labelled pharmaceuticals. Given this information, he would be epistemically irresponsible if he didn't somehow verify the prescriptions that he was filling for the day. As Rosen (2003) explains, "we are under an array of standing obligations to inform ourselves about matters relevant to the moral permissibility of our conduct: to look around, to reflect, to seek advice, and so on" (p. 63). Insofar as we fail to discharge such duties, we are epistemically irresponsible.[23]

[21] Attributionism, for example, rejects this claim about derivative responsibility; as long as the agent's ignorant actions express an inadequate evaluative orientation, she's blameworthy – regardless of whether this ignorance is culpable or not.

[22] The reasoning here is that because an agent's beliefs aren't voluntary, she can only be culpable for the *management* of her beliefs. Rudy-Hiller (2017) and other capacitarian accounts deny this claim for a variety of reasons pertaining to capacities.

[23] Holly Smith (1983) explains the same phenomenon in terms of "benighting acts." For example, if Arthur comes to falsely believe that Rio de Janeiro is the current capital of Brazil by reading an outdated textbook, then the benighting act that caused his ignorance was reading the book. Often, though, what Smith calls benighting acts are more like *omissions*, such as failing to gather certain information. Regardless, on this picture an agent is culpable for her ignorance if she's culpable for the benighting act that caused this ignorance.

At this point, however, a regress looms. After all, if we accept Rosen's original premise that all blameworthiness for factually ignorant wrongdoing is derivative, then epistemically irresponsible agents are blameworthy for their wrongful action only if: (1) it wasn't performed under ignorance, or (2) it was performed under culpable factual ignorance. In the latter case, we would need to determine whether the agent was epistemically irresponsible regarding *this* instance of ignorance, and the regress continues. Ultimately, then the regress can only end in blameworthiness if some level of wrongdoing wasn't performed under ignorance. In Jack's case, this condition might seem easy to fulfill. Given that Jack was aware of the widespread errors with medications, it's intuitive that he wouldn't be acting under ignorance if he omitted verifying his prescriptions. But this is where the argument takes a more controversial turn. According to Rosen, wrongdoing without ignorance requires full *akrasia* – that is, occurrent and conscious awareness that one's actions are all-thing-considered wrong.[24]

Obviously, akrasia is a demanding standard for awareness. It entails that Jack could be excused for omitting to verify his prescriptions, despite his awareness of the widespread errors with medications, if he merely believed that such an omission was *pro tanto* wrong;[25] for instance, if he falsely believed that it wasn't worth his time to verify that he correctly filled the prescriptions. In order to determine Jack's ultimate culpability in these circumstances, we would need to assess whether his false belief about the moral status of omitting to verify the

[24] Occurrent awareness means that the agent believes that her actions are all-thing-considered wrong *at that moment*. Full akrasia is also sometimes referred to as "clear-eyed" akrasia. Other, weaker forms of akrasia are occasionally discussed in other contexts. Note that the agent must believe that her actions are all-thing-considered wrong *as such*, not just that they have the features that make them all-thing-considered wrong. Thus, for instance, it would not be enough for Jack to merely believe that omitting to verify his prescriptions could lead to harm, he must also believe that causing such harm is (all-things-considered) wrong. In the literature, the distinction between these two kinds of beliefs regarding the moral significance of one's actions is sometimes explicated in terms of *de re* vs *de dicto* awareness, where *de dicto* awareness designates the further belief that one's actions are wrong.

[25] By '*pro tanto* wrong' I mean that there are moral considerations that count against it, but they can be outweighed by competing considerations. An act is all-things-considered-wrong if competing considerations don't outweigh the moral considerations against it.

prescriptions is *itself* an instance of culpable ignorance, and the regress continues. The conclusion of this argument is that all blameworthy wrongdoing must ultimately be akratic in the relevant sense. It's worth noting that even though the argument starts with a claim about factual ignorance, it turns out that *moral* ignorance also excuses, as culpability requires occurrent awareness that one's actions are all-thing-considered wrong.

Rosen (2003) admits that accepting the conclusion of this argument would entail major revisions to our responsibility practices:

> What follows if I'm right? People normally do what they believe they have most reason to do; and people normally have most reason to do the right thing. It follows that when people act badly, it is almost always because they have a mistaken belief of this sort. So *if these beliefs are typically blameless*, our excuse applies in an enormous range of cases. (pp. 82-3)

In other words, his argument has the revisionary implication that agents are almost never morally responsible for their wrongdoing, as it's rare that agents act fully akratically.

Several philosophers accept some version of this argument and its revisionary conclusion.[26] However, the vast majority rejects at least one of its assumptions. Indeed, since the revisionist argument was put forth, much of the work on the epistemic dimension has focused on refuting it by various means. One rather obvious response is to simply deny the claim that blameworthiness ultimately requires akrasia, especially insofar as that involves occurrent awareness that the action is all-things-considered wrong.[27] It seems plausible that the standard for both the *kind* and *content* of awareness necessary for blameworthiness is weaker. This is a rather conservative response, as it only refutes one element of the argument, but others reject

[26] See, e.g., Levy (2011) and Zimmerman (1997).
[27] See, e.g., Haji (1997), Peels (2011), and Timpe (2011).

much more. Attributionists, for instance, tend to reject *every* main claim because of commitments that seemingly flow from their antecedent theory of moral responsibility.[28]

Given that much of the contemporary work on the epistemic dimension has focused on refuting the revisionist argument, it's unsurprising that recent overviews and introductions of the epistemic dimension have structured their discussion around this argument.[29] Nevertheless, the revisionist argument won't be a focus of this dissertation. My reasons for this departure are as follows. First, as suggested before, I believe that there are plausible solutions to the regress that focus on the awareness condition. Indeed, my final account of the epistemic dimension represents one such solution (although it's not presented this way). Of course, proponents of the revisionist argument have responses to even this more conservative approach,[30] but I don't think these responses are convincing. Ultimately, then, I just don't believe that the revisionist argument should worry most theorists as much as other threats to moral responsibility.

Even if the revisionist argument isn't so threatening, though, it might be worth engaging with insofar as it provides useful framing. We might compare approaches to the epistemic dimension in terms of which assumptions of the argument they reject, for example. Still, although I acknowledge that there is some value here, this framing often obscures more than it clarifies. Consider, for instance, the premise that all blameworthiness for factually ignorant wrongdoing is derivative. It's not obvious what a capacitarian account should say about culpability here. If the agent possesses the relevant capacities for awareness, then she's blameworthy for her ignorant wrongdoing on such an account; but does the possession of such

[28] See, e.g., Arpaly (2002), Smith (2005), and Talbert (2017a).
[29] See, e.g., Rudy-Hiller (2022) and Wieland (2017a). Although I depart from their focus on the revisionist argument in the rest of the dissertation, my introduction draws heavily from their framing. In general, their work characterizing the dialectic regarding the epistemic dimension has been invaluable to my own project.
[30] See, e.g., Levy (2016).

capacities also imply that her ignorance is culpable? Capacitarians disagree on this point, partially because this notion of culpability is ambiguous within the revisionist argument.[31] It would be more instructive, then, if we could just discuss capacitarianism outside of its relation to this argument. The same is true for other views on the epistemic dimension grounded in independent theories of moral responsibility.

Finally, there are many interesting questions regarding the epistemic dimension outside of the framework of the revisionist argument. For instance, what is the relation between epistemic (or intellectual) difficulty and moral responsibility? Alexander Guerrero (2017) and others plausibly argue that the difficulty involved in achieving the requisite awareness for responsibility should mitigate blameworthiness under certain conditions. Fully explicating this relation seems like a worthwhile project that isn't directly connected to the revisionist argument. By narrowly structuring the discussion around this argument, I worry that worthwhile issues like this will largely go unnoticed. I would like my own investigation of the epistemic dimension to be free to explore the full range of possible questions. That being said, in chapter five I will return to the revisionist argument in order to briefly explain how my own theory of the epistemic dimension would address the challenge.

3. The Data Set

Another important issue that I want to address is the boundaries of my examination of the epistemic dimension. In investigating the influence of epistemic considerations, I will often draw on cases of ignorant wrongdoing to generate and support intuitions, and it's worth delineating this data set. As it turns out, there's no uncontroversial account of the mental states and contents

[31] For instance, Fernando Rudy-Hiller (2017) accepts the claim that an action done from non-culpable factual ignorance is itself non-culpable, whereas fellow capacitarian Randolph Clarke (2017) denies it but argues that "substandard" awareness can still ground blameworthiness.

that might be relevant to moral responsibility, and so it's best to be perspicuous about what one considers to be the *explanandum* at hand. Furthermore, as each state has a distinctive character, it's important to distinguish them when necessary.

I hold that there are five epistemic (or doxastic) states potentially relevant to the epistemic dimension. The first two should be uncontroversial: *false belief* and *unconsidered true belief*. Indeed, when we discuss ignorance of some fact in ordinary language, we usually have one of these two states in mind. For example, we would say that Alex is ignorant of the fact that Irvin Kershner directed Star Wars: Episode V if he either falsely believed that George Lucas directed the film, or if he had never even considered who directed it. The latter is sometimes called *deep ignorance* because the agent never even entertained the relevant proposition. Regardless of their differences, though, it's clear that both can influence moral responsibility.

The third epistemic state relevant to the epistemic dimension has recently played a significant role in disputes between attributionism and capacitarianism: *forgotten beliefs* (specifically, forgotten true beliefs). In general, capacitarians argue that an agent can be blameworthy for wrongdoing resulting from her forgetting some morally relevant information, despite her actions expressing the proper evaluative orientation. Unlike the previous two states, forgetting essentially involves a *temporal* dimension. When an agent forgets some fact, she no longer occurrently believes it. This description leaves room for different versions of the phenomenon. For example, Susan might temporarily forget the name of a classmate from elementary school but recall it later in the day. Conversely, she might permanently forget the address of her elementary school, failing to recall it no matter how hard she tries. The temporary version is usually the focus of the aforementioned disputes between capacitarianism and attributionism, as agents in cases of temporary forgetfulness appear to possess relevant capacities

for awareness. Nevertheless, both types of forgetting appear to be relevant to assessing moral responsibility in certain cases.

The fourth state, *uncertainty*, is much less discussed in relation to the epistemic dimension.[32] Although the details are contentious,[33] it seems intuitive that there's a category that's importantly distinct from both genuine belief and disbelief. The paradigm case would be 0.5 degrees of belief toward a proposition, but perhaps uncertainty covers some *range* of values that doesn't cross the threshold into genuine belief or disbelief.[34] Insofar as uncertainty is worth separating as a discrete category, it presents potentially unique issues regarding the epistemic dimension. For instance, how should we assess the actions of agents who are *morally* conflicted; that is, they're uncertain whether their actions are permissible?[35] Are they obligated to exercise special caution under such circumstances? Depending on how one understands uncertainty, I will either have very little or much to say about this state.[36]

Finally, *suspension of belief* represents a fifth state potentially relevant to the epistemic dimension. As with uncertainty, the nature of this state is controversial,[37] but most epistemologists agree that there is a neutral attitude complementing belief and disbelief. As Matthew McGrath (2021) explains it, this third option "is not the mere absence of a doxastic attitude…To be neutral whether p, instead, is to be in a positive state on the question whether p,

[32] One notable exception is Guerrero (2007), which focuses on uncertainty.
[33] One source of controversy is whether uncertainty is best understood in terms of belief or credence. Although I choose to cast things in terms of degree of belief below, I don't mean to take a stand on debates about the relation between beliefs and credences.
[34] My use of 'uncertainty' in this context differs from a more technical sense sometimes used in decision theory to denote the absence of any genuine doxastic state toward a proposition.
[35] Often when we're uncertain about the permissibility of our actions, it seems plausible that we're obligated to seek out information that could improve our epistemic situation. Sometimes, however, we don't have the necessary time (or access) and must act under uncertainty.
[36] Specifically, if uncertainty is understood broadly as simply any belief about the probability of some proposition, then chapter four's discussion of awareness of risk is essentially a chapter about uncertainty.
[37] Among the sources of controversy: (1) is suspension of belief a genuine *doxastic* state; (2) is suspension of belief distinct from uncertainty; (3) are there multiple versions of suspension of belief?

and in that sense to be an attitude" (p. 464). Regardless of how this attitude is properly construed, though, I won't discuss it in the proceeding chapters. As Rik Peels (2014) argues, "it seems that suspending ignorance, in opposition to disbelieving and deep ignorance, gives rise to further obligations, namely an obligation to investigate or find something out if the stakes are sufficiently high" (pp. 492-3). Because of this fundamental difference, it's dubious that suspension of belief has direct implications for moral responsibility, and so I'll follow most others in setting it aside while exploring the epistemic dimension.

Having discussed the mental states that might be involved in ignorant wrongdoing, I turn now to the contents of these states. Again, there are certain elements that clearly have implications for moral responsibility. First, awareness of one's actions and their consequences. If Doug unwittingly parks in a handicap spot because the sign blew away in a storm, then he's plausibly excused for parking where he shouldn't, given that he wasn't aware he was doing it. Second, awareness of the moral significance of one's actions.[38] If Doug falsely believes that it's permissible to park in a handicap spot when you're in a rush, then this moral ignorance might at least mitigate his blameworthiness, especially if this belief was inculcated in him by certain social circumstances. Some will flatly deny that this kind of moral ignorance ever mitigates blameworthiness, but most at least agree that one's awareness of the moral significance of one's actions is sometimes relevant to responsibility.

Other kinds of contents of awareness are much less discussed, perhaps because they're less obviously relevant. I group these contents in the broader category of awareness of one's abilities and alternatives. For example, imagine that Doug is taking a friend to the emergency

[38] As discussed in fn. 24, there's a distinction between two kinds of beliefs regarding the moral significance of one's actions: beliefs about the wrong-making *features* of certain actions (*de re* awareness) versus beliefs about the wrongness *as such* of certain actions (*de dicto* awareness). There's some disagreement about whether lack of *de re* awareness counts as moral ignorance, but clearly both kinds of beliefs are relevant to moral responsibility.

room, and he parks in a handicap spot because he falsely believes that there are no other open spots. Some, most notably Neil Levy (2011), argue that agents like Doug can't be blamed unless they believed that (reasonable and permissible) alternatives were open to them. Now, suppose that Doug knows that there are other open spots – perhaps there was a car counter at the entrance of the parking structure – but the structure is so complex that he doesn't know *how* to get to these spots in a practical amount of time. Peels (2014) and others argue that Doug is plausibly excused if he ends up parking in a handicap spot because he lacks the requisite "how to" knowledge.[39]

As these kinds of contents aren't prominent in the literature, it's difficult to say how controversial they are as possible constraints on moral responsibility. For my part, it seems that their relevance to responsibility largely depends on the details of the case. Obviously, these kinds of ignorance aren't excusing if the ignorance itself is culpable. For instance, Doug's false belief that there are no other open spots isn't excusing if he made no conscious effort to find one. Assuming that the ignorance isn't culpable, though, ignorance of one's abilities and alternatives plausibly excuses (or mitigates) blameworthiness in certain situations. But even if such cases are both interesting and relevant, I won't focus on them in the proceeding chapters. In my estimation, these kinds of contents aren't central to the epistemic dimension, and so they shouldn't be the emphasis in constructing a theory.

Conclusion

In the following chapters, I will evaluate attempts by prominent theories of moral responsibility to capture the epistemic dimension. The deficiencies of these efforts point toward a better theory. The failure of attributionism to capture intuitions regarding certain cases of

[39] See, e.g., Rudy-Hiller (2019) for an argument that voluntarily exercising the cognitive control necessary for moral responsibility requires that "the agent must know how to avoid the risk in question, that is, what to do in order to achieve the desired cognitive state" (pp. 724-5).

ignorant wrongdoing suggests that blameworthiness requires certain capacities and opportunities characteristic of a kind of control that attributionism constitutively rejects. Yet, I argue that these capacities are insufficient for blameworthiness because fair opportunity also requires awareness of risk. Since fair opportunity best explains why capacities are relevant to responsibility in the first place, I understand the awareness of risk condition as a natural extension of a reasons-responsiveness theory based on fair opportunity. Ultimately, then, the best theory of the epistemic dimension combines reasons-responsive capacities and awareness of risk.

In evaluating various accounts of the epistemic dimension, I will focus on assessing cases in which agents are ignorant of either their actions, the consequences of those actions, or the moral significance of those actions. The pertinent types of ignorance include false belief, unconsidered true belief, forgotten true belief, and perhaps uncertainty. Such cases represent the core of the epistemic dimension of moral responsibility. Although intuitions may vary with regards to certain cases, my goal is to develop a theory of the epistemic dimension that limits revisionism overall.

ATTRIBUTIONIST THEORIES OF MORAL RESPONSIBILITY

Introduction

A reasonable starting point for developing an account that can explain our intuitions regarding the epistemic dimension is to look at extant theories of moral responsibility and see whether they already have the resources to do this explanatory work. If this were successful, then the task at hand would just be carrying out this explanatory work. One reason for skepticism, though, is that these theories were primarily developed to address issues relating to freedom (or control). Thus, if our intuitions suggest a more direct relation between epistemic considerations and moral responsibility – i.e., one that does not necessarily flow through issues related to freedom – then it's conceivable that existing theories might not have the resources to account for these intuitions. In this case, these theories can perhaps be *supplemented* to include the influence of epistemic considerations. Still, the relevant supplementation should be consistent with the core of the existing theory, so that an independent account of the epistemic dimension isn't simply being grafted on.

One theory that appears well equipped to account for epistemic considerations is attributionism.[40] Although attributionism is perhaps best defined negatively, the view broadly holds that "assessments of moral responsibility are, and ought to be, centrally concerned with the morally significant features of an agent's orientation toward others that are attributable to her…" (Talbert, 2022, p. 54).[41] Among attributionist theories, a common way to explain an agent's evaluative orientation is in terms of the quality of her will, and so many attributionist accounts

[40] Examples of attributionist views include Adams (1985), Arpaly (2002), Scanlon (1998), Sher (2006), Smith (2005), and Talbert (2012). Neil Levy (2005) originally coined the term to refer to these kinds of views.
[41] Overall, there is significant variation among attributionist account – so much so that Manuel Vargas (2020) contends that, "it is unclear whether so-called attributionist accounts are unified in any deep way" (p. 412). I just want to pick out a loose grouping of views that explain responsibility in terms of expression of certain attitudes, and which deny that certain kinds of control are necessary for responsibility. Less importantly, attributionist theories also tend to reject *historical* conditions on moral responsibility.

are also so-called *quality of will* views. It's important to note, though, that not all attributionist theories are quality of will views, and not all quality of will views are attributionist. Michael McKenna (2012), for example, holds a quality of will view that is reasons-responsive rather than attributionist. Furthermore, even among attributionist quality of will views there is significant disagreement about *which* attitudes comprise an agent's will. A prominent family of views that traces back to T.M. Scanlon's work explains quality of will in terms of evaluative judgments.[42] However, other attributionist views are more conative, describing quality of will in terms of states like desires and cares.[43]

Although my discussion primarily focuses on quality of will accounts, as they are the largest family of attributionist views, I propose an expansive understanding of attributionism that also includes at least some *deep self* (or real self) theories, such as a recent version by Chandra Sripada (2016).[44] Hopefully, including such theories will not invite more confusion regarding theoretical boundaries; but given that deep self views like Sripada's also focus on the attributions (or expressions) of morally significant attitudes, and reject control conditions, I believe it makes sense to treat them as attributionist. Eventually, it will be necessary for me to differentiate some of these attributionist views, but for now I just want to portray a broad tradition that seems to have resources for capturing certain intuitions regarding epistemic considerations.

Specifically, attributionism appears well equipped to capture intuitions in cases where the ignorance is suitably related to morally objectionable attitudes. An especially stark subclass of these cases are instances of *motivated (or affected) ignorance*, wherein an agent purposefully

[42] See, e.g., Hieronymi (2008), Smith (2005), Talbert (2012).
[43] See, e.g., Arpaly (2002) and Björnsson (2017).
[44] Indeed, Sripada himself considers his view to have a similar approach as the kinds of quality of will views previously mentioned. The broader category that includes both quality of will and deep self views is sometimes designated as *self-expression* theories. However, I believe that both types of theories can usefully be considered attributionist.

avoids potentially morally relevant information.[45] For example, suppose that a believer of the conspiracy theory "QAnon" chooses to avoid information that challenges her false belief that the 2020 U.S. presidential election was rigged against Donald Trump. Specifically, she avoids the information because she believes that the media is controlled by a cabal of Satanic, cannibalistic pedophiles. If she then goes on to storm the Capitol based on her belief that the election was rigged, she's not plausibly excused for acting on it, even if she wouldn't have stormed the Capitol had she corrected her belief. According to attributionism, this is because her ignorance is motivated by the same morally objectionable attitudes that would ordinarily make an agent blameworthy for consequent wrongdoing; and thus, her ignorant wrongdoing is also an (indirect) expression of these attitudes. Insofar as this explanation makes sense, then, attributionism can support and explain the common intuition that motivated ignorance isn't excusing.

Yet, despite this explanatory power, attributionism has difficulty capturing and explaining intuitions in other sorts of cases involving epistemic considerations. In the following chapter, I present two such cases that best demonstrate this limitation. The first case (in section 1) shows that an agent who expresses a bad (or indifferent) will can still be excused for wrongdoing when social context makes it suitably *difficult* for him to ascertain the moral truth. The second case (in section 2) shows that an agent who expresses a good will can still be morally responsible for wrongdoing when he's *reckless* in the management of his epistemic situation. Taken together, these two cases indicate that expressing a certain quality of will is neither necessary nor sufficient for moral responsibility. In section 3, I evaluate possible responses on behalf of attributionism. Ultimately, I argue that attributionism can't plausibly capture these

[45] Moody-Adams (1994) represents the *locus classicus* for discussions of motivated (or affected) ignorance.

types of cases without appealing to considerations that undermine essential features of attributionism.

Before presenting the two cases, however, I must make an important clarifying remark about the proceeding discussion. Because I'll assess the moral responsibility of agents in certain situations, it's crucial that I'm perspicuous about how I understand the concept of moral responsibility. To this end, I'll assume a conception whereby an agent's responsibility for her actions is based on reasonable demands arising from moral obligations.[46] On this account, when an agent violates a reasonable demand that she had a fair opportunity to comply with, she is *deserving* of blame and the associated reactive attitudes (e.g., resentment, guilt, indignation).[47] This conception is commonly referred to as responsibility as *accountability,* following Gary Watson's (1996) distinction between accountability and attributability. The latter notion of *attributability* is more controversial, but it generally concerns "whether some action can be attributed to an agent in the way that is required in order for it to be a basis for moral appraisal" (Scanlon, 1998, p. 248).[48]

Now, it might seem illegitimate for me to assume a conception of moral responsibility as accountability going forward, given that attributionism tends to focus on responsibility as attributability. However, although attributionism gains its name from focusing on attributability,

[46] It's worth noting that reasonable demands are necessary but not sufficient for accountability. As I discuss in the next section, it's possible for an agent to be excused from accountability despite being the subject of reasonable demands. Nevertheless, such demands are characteristic of accountability in contrast to attributability.

[47] This conception of moral responsibility in terms of the deservingness of praise and blame is often referred to as the *basic desert* sense. See Pereboom (2014) for a canonical discussion of basic desert responsibility. There is debate about whether basic desert responsibility and accountability are equivalent. Although I believe they are, for current purposes I simply stipulate that accountability requires desert of blame.

[48] Following Shoemaker (2011), some acknowledge a third conception – namely, responsibility as *answerability*. Although I agree that this distinction makes sense, answerability is really a version of attributability. Thus, I believe the most important contrast is between accountability and attributability. It is worth nothing, though, that I will focus on attributionist views that are concerned with answerability, as they seem to have sophisticated resources for handling the challenges that I raise.

most attributionist views hold that satisfying the conditions for responsibility as attributability are necessary *and* sufficient for accountability. Because of this, it seems fair to evaluate the following cases in terms of accountability, while acknowledging that there's another conception that might be important for other sorts of evaluations. In fact, in discussing possible responses on behalf of attributionism in section 3, I assess the strategy of using the distinction between attributability and accountability to argue for the separation of moral responsibility and blameworthiness. This strategy would potentially allow attributionists to avoid some counterintuitive implications of their theory by arguing that certain excused agents aren't blameworthy despite fulfilling the conditions for attributability. In this context, it's obviously necessary to carefully separate and consider each conception, but unless otherwise noted I assume that evaluations of moral responsibility involve accountability.[49]

1. Case One: Difficulty for Dennis

With the preceding clarifying remarks out of the way, I now present the first of two main cases that create problems for attributionism. I initially focus on Scanlonian quality of will views that ground moral responsibility in evaluative judgments, as these are the most prominent attributionist accounts. This first case demonstrates that an agent can be excused for wrongdoing, despite expressing a bad (or indifferent) will, based on epistemic considerations.[50]

> **Difficulty for Dennis**: Dennis is a young man who grew up in an insular community of ethnically homogenous people. One of the shared beliefs in this community is that

[49] One issue that I will not discuss is the *nature* of blame. Although most moral responsibility theorists claim to be focusing on responsibility as accountability, there's seemingly less agreement on the target concept of blame. Because of this variance, I don't want to presume any conception in my analysis of the epistemic dimension. That being said, my general sympathies lie with Brink and Nelkin's (2019) *core and syndrome* approach. According to this view, there is a core that is necessary and sufficient for blame, but also normal manifestations of blame that constitute a non-accidental syndrome.

[50] An action expresses a bad will insofar as it expresses an evaluative orientation that is vicious in some way, whereas an *indifferent* will is simply insensitive to the pertinent moral considerations. Indifference is normally sufficient for responsibility on a quality of will view, but unless a distinction is necessary I will often use 'bad will' as inclusive of an indifferent will.

> members of another ethnicity are untrustworthy. Although Dennis has not had contact with this other ethnic group, he grew up hearing false stories about them. In particular, they were often invoked to explain his community's current and historical problems. These beliefs and stories were not challenged within Dennis's community, and so Dennis accepted them. One day, Dennis finally ventures outside of his community and encounters a member of this other group in need of help. Dennis is the only person available to help, but he declines to do so because he judges this person to be untrustworthy.

I assume that Dennis acts wrongly by violating a moral obligation to provide help when no significant sacrifice is required and no one else is available to provide aid. Such an obligation seems consistent with a fairly weak demand for beneficence. Yet, Dennis also acts from ignorance (i.e., false belief) in omitting to help, assuming that he would have helped had he not judged the person untrustworthy.[51] Does this ignorance excuse his wrongdoing? I argue that intuitions suggest that Dennis is (at least partially) excused under these conditions. Still, he clearly seems to express a bad will in refusing to help, as his omission reflects his morally objectionable judgment that members of a certain ethnic group are untrustworthy. Therefore, attributionism seems committed to the counterintuitive result that Dennis is blameworthy for refusing to help.[52] In the remainder of this section I'll support this conclusion.

1.1 First, why think that Dennis is excused for refusing to help? After all, as demonstrated in the discussion of motivated ignorance, not all ignorance excuses wrongdoing. In Difficulty for Dennis, though, the *source* of his ignorance is much different than in motivated ignorance cases. Instead of being self-imposed, Dennis's false belief that members of another ethnic group are untrustworthy is the product of his community and upbringing. Now, one might argue that many of our evaluative judgments are informed by these same factors, but Dennis's situation is clearly

[51] An agent acts *from* ignorance when she would have acted differently, if not for her ignorance. This kind of ignorance contrasts with acting *in* ignorance, which does not involve this counterfactual. It is generally held that only acting *from* ignorance is potentially excusing.

[52] I assume here that being morally responsible (in the accountability sense) for wrongdoing is sufficient for blameworthiness. Later, I examine this assumption.

exceptional in certain respects. Specifically, his community is remarkably homogenous and insular, meaning his potential sources of information within this social context are very limited. Because of this, it seems unreasonable to expect Dennis to form a contrary belief regarding the trustworthiness of the relevant ethnic group and thus avoid subsequent wrongdoing. Given that moral responsibility requires reasonable expectations, then, Dennis is plausibly excused for refusing to help.[53]

If this line of reasoning is intuitive, it's worth considering why it seems unreasonable to expect Dennis to form a contrary belief. In this case, I believe the key consideration is the *difficulty* involved in recognizing the falsity of the claim that the relevant ethnic group is untrustworthy. After all, not only is this prejudice universally shared within the community, but it's also integrated into their cultural and historical understanding. Moreover, Dennis was never afforded the opportunity to meet people from this ethnic group and perhaps correct his own prejudice through first-hand experience. This distinguishes Dennis from more typical cases of enculturated prejudice, where there's usually at least some chance to gain countervailing evidence about the relevant group. Of course, these interactions are likely to be colored by biases, but we're still less willing to excuse related wrongdoing from agents who have been afforded such opportunities. In this way, the relevant difficulty generated by one's social context isn't necessarily excusing, but Dennis's particular situation seems sufficiently difficult to at least mitigate responsibility.

In response, one might acknowledge that difficulty can sometimes excuse wrongdoing while denying that Difficulty for Dennis involves the requisite difficulty. A strong version of this

[53] There are different ways to account for this excuse. At this point, I intend to remain neutral between these various explanations, and so I discuss things in terms of the shared condition of reasonable expectations for action. This concept of reasonable expectations will be important in section 3.

position asserts that *epistemic* difficulty never excuses.[54] Yet, this strong claim is implausible. After all, morally relevant reasoning can be extremely complicated.[55] Imagine, for instance, that some head of state must decide whether to get involved in a foreign conflict; and suppose that she has gathered as much information as possible prior to the time of action. Still, even if she has sufficient evidence to make the right decision, she's plausibly excused for making the wrong decision if the required reasoning is sufficiently complex. Under these conditions, it seems unreasonable to expect her to come to the correct conclusion. Thus, some threshold of epistemic difficulty plausibly excuses, either based on available evidence or the complexity of the required reasoning.[56]

A weaker version of the above position might instead hold that only *factual* ignorance can excuse based on difficulty. The basic argument behind this claim would be that although realizing non-moral truths can sometimes be unreasonably difficult (e.g., deciphering the geopolitical ramifications of state intervention), realizing moral truths never meets this threshold. In this case, *moral* ignorance like Dennis's would never excuse. Such a position is perhaps captured by Elizabeth Harman's (2011) claim that "ordinary people who know the non-moral facts of what they are doing, when they do wrong things, often do have *sufficient evidence* that their actions are wrong" (pp. 461-2). According to this view, awareness of the non-moral facts that make an action wrong undercuts excuse because it's never suitably difficult to reason from these facts to the conclusion that the act is wrong.[57]

---

[54] Epistemic difficulty is sometimes referred to as *intellectual difficulty*. See, e.g., Guerrero (2017).

[55] I would like to remain neutral regarding the right account of difficulty. However, difficulty is plausibly indexed to particular agents or kinds of agents. In this way, required reasoning that might be sufficiently difficult to excuse one agent might not excuse another with greater cognitive capacities.

[56] Sher (2017) makes this claim for factual mistakes (p. 105).

[57] Although I characterize Difficulty for Dennis as a case of moral ignorance, it's perhaps more accurate to describe it as a *mixed* case. After all, Dennis also has false non-moral beliefs about the past actions of this ethnic group that help explain why they're untrustworthy. Indeed, most cases of moral ignorance are plausibly mixed cases like this. Still, I think that Harman and others would claim that Dennis has enough true beliefs to reason to the conclusion that

Yet even this weaker claim is implausible. Consider cases where an agent is mistaken about the correct *weight* of relevant moral considerations.[58] Suppose, for example, that Andre faces a dilemma: he is scheduled to volunteer at the local soup kitchen on Friday night, but this happens to be the same night as his best friend's 30th birthday party. In this case, Andre has competing moral obligations. His obligation to his friend derives from their close relationship, and his obligation to volunteer comes from a promised commitment. Given that Andre cannot fulfill both obligations, he must decide what he ought to do *all-things-considered*. Now, suppose that Andre makes the wrong decision because he incorrectly weighs these moral considerations. It's still plausible that we would excuse Andre's wrongdoing because of the difficulty in correctly balancing these considerations.[59] Indeed, if theorizing about morality teaches us anything, it's that determining the right action can sometimes be extremely challenging. Thus, a strict asymmetry between factual and moral ignorance based on difficulty is artificial and under-motivated.[60]

Finally, consider an even weaker version of the current position, which accepts that epistemic difficulty of any type can excuse but simply denies that Difficulty for Dennis meets the necessary conditions. This version might argue that although agents like Andre have difficult

---

he ought to provide aid. Moreover, his wrongdoing is most proximately caused by his false moral belief, not any false non-moral beliefs about his actions. Wieland (2017b) might characterize Difficulty for Dennis as an "impure" case of moral ignorance, insofar as Dennis's moral ignorance is (at least partially) based on non-moral ignorance.

[58] See King (2020) and Sher (2017) for more examples like this. It also seems possible that moral ignorance can be excusing in cases where an agent has a false belief about the relative weights of moral and *non*-moral considerations (e.g., prudential considerations). However, the case of competing moral considerations appears strongest for excuse.

[59] One might have a view that in cases of competing moral obligations an agent is only blameworthy if she acts on an obligation that is significantly weaker than the other(s), *regardless* of any epistemic (difficulty) considerations. In this way, blameworthiness requires a certain level of suboptimal action. On this account, Andre might not be blameworthy merely because his competing obligations are similarly weighty. Nevertheless, one could imagine another case where Andre's obligations have greater divergence in weight, and yet it's still sufficiently difficult to reason toward the right conclusion to provide an excuse. Ultimately, it's intuitive that certain failures to correctly weight moral considerations can be undermined by difficulty and not by other factors.

[60] Such an asymmetry also conflicts with the criminal law's *M'Naghten rule* and much of the discussion regarding whether psychopathy is excusing.

routes to recognizing the moral truth, Dennis doesn't face the same difficulty. After all, there's no indication that Dennis has reduced capacities for moral reasoning, or that he must balance competing moral considerations. Because of this, coming to the correct belief shouldn't be too challenging for him, given the available evidence. In fact, reasoning along these same lines, George Sher (2017) argues that even someone who is indoctrinated from childhood to hold false moral beliefs is usually blameworthy for related wrongdoing because "the basic facts about persons that lead by the familiar non-deductive routes to the requirements of commonsense morality—the fact that each person has an interiority, holds various things dear, is vulnerable in many ways, and so on—are as available to him as they are to us" (p. 114).

Although this weaker version is the most plausible, though, there are strong reasons for thinking that Dennis's situation is sufficient for excuse (or at least mitigation). Regarding Sher's above claim, it's important to note he's mostly thinking about social doctrines that advocate rather *extreme* wrongdoing. Indeed, ever since Rosen (2003) introduced his rather influential case involving an ancient slaveholder,[61] discussions about the effect of social context on the epistemic dimension of moral responsibility have tended to focus on cases like this. This feature of the literature is unfortunate insofar as focusing on these extreme cases has led to a biased conception of enculturated ignorance that tends to disfavor excuse.

Conversely, Difficulty for Dennis doesn't involve severe wrongdoing like slavery. Dennis doesn't directly harm a member of the allegedly untrustworthy ethnic group – he simply fails to provide aid – and we needn't assume that his culture permits more significant

[61] Rosen (2003): "*Ancient Slavery*. In the ancient Near East in the Biblical period the legitimacy of chattel slavery was simply taken for granted. No one denied that it was bad to be a slave, just as it is bad to be sick or deformed. The evidence suggests, however, that until quite late in antiquity it never occurred to anyone to object to slavery on grounds of moral or religious principle. So consider an ordinary Hittite lord. He buys and sells human beings, forces labour without compensation, and separates families to suit his purposes. Needless to say, what he does is wrong. The landlord is not entitled to do these things. But of course he thinks he is" (p. 64).

wrongdoing.[62] In Dennis's case, then, it's much less plausible that the basic facts about persons that Sher cites would easily lead to the recognition that his enculturated belief is false. After all, the claim that a certain group is untrustworthy – and so it's imprudent to try to help them – conflicts much less with the requirements of commonsense morality, especially when it doesn't license inflicting serious harm. Even assuming that Dennis's insular community didn't deprive him of sufficient evidence to recognize the moral truth, then, it's still plausible that this recognition would be sufficiently difficult to (at least partially) excuse. At the very least, claims like Sher's about the relative ease of coming to the moral truth are much less plausible in Difficulty for Dennis, where the potential wrongdoing isn't so significant.

Going further, I argue that Dennis is intuitively excused based on difficulty even *if* he could have rather easily corrected his false belief had he genuinely questioned it. This is because, as Alex Guerrero (2017) points out, not all difficulty derives from one's lack of skills, or the amount of effort required, there is also what he terms "difficulty in trying." As Guerrero (2017) explains the phenomenon, "it might be difficult for a reasoner, R, to try to come to have a justified true belief about M, because doing so would go against R's fundamental beliefs and/or values, be difficult for R to steadily resolve to do, or simply not be something that R would think to do" (p. 204). This appears to be precisely the kind of difficulty involved in Dennis's situation. Not only would coming to the correct belief involve going against other fundamental social beliefs and/or values, but he would also have to first think to reconsider his false belief, and nothing in his social context recommends applying such scrutiny. If anything, the certainty of his prejudice belief is constantly reinforced. Therefore, it seems unreasonable to expect Dennis to

[62] Of course, failing to provide aid can be a serious (and salient) wrong in cases where such a failure results in significant harm. For current purposes, I assume that Difficulty for Dennis is not such an extreme case.

come to the moral truth and avoid wrongdoing. Dennis simply isn't afforded an adequate opportunity to reconsider his false belief from within his social context.

1.2 Of course, the fact that Dennis is intuitively excused for his wrongdoing only creates problems for attributionism if the theory entails that Dennis is blameworthy. But why should we think that attributionism is committed to this conclusion? I argued earlier that Dennis expresses a bad will in refusing to help because his omission reflects his objectionable judgment that members of a certain ethnic group are untrustworthy; given a Scanlonian quality of will view that grounds responsibility in evaluative judgment, then, Dennis is intuitively blameworthy for refusing to help. Still, it is worth justifying this judgment and evaluating whether there are versions of attributionism that might have the resources to excuse Dennis.[63]

Broadly speaking, attributionist accounts have two main components: (1) a privileged set of *attitudes*, and (2) a certain *expression relation* between these attitudes and actions (or other attitudes) that explains responsibility. As mentioned before, evaluative judgments are the relevant attitudes on Scanlonian quality of will views, but the notion of an evaluative judgment requires more clarification. Attributionist Angela Smith (2005) contends that such judgments "are not necessarily consciously held propositional beliefs, but rather tendencies to regard certain things as having evaluative significance" (p. 251). Furthermore, she claims that what makes these judgments evaluative is that they are about reasons, specifically *moral* reasons. On this view, then, an agent is blameworthy insofar as her action bears the appropriate expression relation to an objectionable judgment about moral reasons. According to Smith (2008), an action

[63] If it's intuitive that Dennis is only *partially* excused for his wrongdoing, then attributionism need only explain this mitigation of blameworthiness. However, although quality of will plausibly comes in degrees, epistemic difficulty isn't a mitigating factor for the same reasons that I will argue that it isn't an excusing factor. Perhaps Dennis expresses a better quality of will than other prejudiced agents, such as those who wrong members of another group purely out of disregard for their interests, but this difference isn't explained in terms of their respective epistemic states.

expresses an agent's evaluative judgment if "she can, in principle, be called to defend it with reasons and to acknowledge fault if an adequate defense cannot be provided" (p. 370). Such a notion of expression rules out as blameworthy implanted attitudes, and certain unwitting omissions, but it includes many other non-voluntary attitudes and actions.

Given that Smith has perhaps the most prominent and developed Scanlonian account, I use her view as the exemplar – and on this view Dennis appears rather straightforwardly blameworthy. First, Dennis's omission clearly expresses an evaluative judgment. If asked to justify his actions, he would cite his belief that members of the relevant ethnic group are untrustworthy and so it would be imprudent to try to help.[64] Secondly, this evaluative is clearly objectionable. Again, Dennis's enculturated prejudice might not allow more extreme forms of wrongdoing, but it does allow a wrongful omission in the situation. Therefore, Dennis is blameworthy because he expresses a bad will in refusing to help.

In fact, quality of will attributionists often openly embrace the implication that agents in cases like Difficulty for Dennis are responsible for their wrongdoing. Matthew Talbert (2017a), for example, draws this conclusion in a similar case featuring "a young man who, as a result of an extremely insular upbringing and heavy indoctrination, has acquired a virulently homophobic outlook" (p. 54). In response to William Fitzpatrick's (2008) original argument that this young

[64] In correspondence, Rosalind Chaplin offers the following possible challenge to the claim that Dennis expresses a bad will: "Suppose the attributionist has the intuition that Dennis is excused because he's acting from the (admittedly false) belief that he'll be putting himself in danger if he helps the person—and as they want to say, this means he's not acting from ill-will after all." In other words, why not read Difficulty for Dennis as a case of mistaken prudence rather than expressing a bad will? In response, I first concede that perhaps Dennis expresses a *better* will than other kinds of prejudiced agents, such as those who wrong members of another group purely out of disregard for their interests. Still, I believe that attributionist views of the sort I'm targeting would maintain that actions grounded in the belief that another group is untrustworthy are inherently expressive of a bad will. If this isn't right, however, the case can be easily changed to generate the desired result. For example, Chaplin herself suggests a version where Dennis refuses to help simply because he believes untrustworthy (or otherwise flawed) people don't deserve help. In this alternate version, Dennis is still intuitively excused, despite expressing a bad will.

man, Daniel, is excused for subsequent wrongdoing because it's "not his fault that he came to be this way" (p. 35), Talbert replies as follows:

> I agree that if Daniel is not at fault for being the way he is, then he is not *to blame* for his faults in the causal sense of that expression. But on my view, desert of blame—that is, being *blameworthy*, being an appropriate target of blaming responses—doesn't depend on whether an agent is causally responsible for his faults...So, if facts about how a wrongdoer came to be the way he is do not call into question the moral status of the judgments that inform his behavior, or the attributability of these judgments to him, then they do not call into question the aptness of blaming responses. (pp. 54-5)[65]

In this way, Talbert argues that etiological facts about how an agent's social context caused her evaluative orientation are irrelevant to her blameworthiness, except insofar as these facts undermine the attribution of a relevant evaluative judgment.[66] Given that these facts rarely undermine such attributions, then, enculturated prejudices will rarely excuse. Thus, even if Daniel represents a slightly more extreme version of Dennis, it seems clear that Talbert would also blame Dennis – as would most Scanlonian views. Importantly, I choose to focus on Talbert here because of the starkness of his statement about the irrelevance of social context to blameworthiness, not because this commitment is exceptional.

What about (non-Scanlonian) quality of will views that ground moral responsibility in attitudes that are *conative* rather than cognitive? Might such a view have the resources to excuse Dennis for his wrongdoing? One representative conative view is Chandra Sripada's (2016) *Self-Expression* account, which he characterizes as a deep self theory of moral responsibility.[67] As with all attributionist theories, the Self-Expression account has two main components: (1) a

[65] I've omitted Talbert's (2017a) explanation of the function of blame, for simplicity; according to this account, "blaming responses like resentment are largely means of marking and protesting a wrongdoer's objectionable evaluative judgments" (p. 55).
[66] Indeed, it's common for attributionist theories to deny such *historical* conditions on moral responsibility.
[67] According to Sripada (2016), "all deep self theories share the view that, of the totality of attitudes in a person's psychology, there is a distinguished subset of them that are fundamental to her practical identity… once this subset is specified, all deep self views agree that a person is morally responsible for an action only if it expresses her deep self" (pp. 1204-5).

privileged set of attitudes, and (2) a certain expression relation between these attitudes and actions (or other attitudes) that explains responsibility. For Sripada, the privileged attitudes are *cares*, which are distinguished from other attitudes by their *functional role*; specifically, "they exhibit a syndrome of dispositional effects that includes motivational, commitmental, evaluative, and affective elements…" (p. 1209) As for the pertinent expression relation, the Self-Expression account explains expression in terms of *motivational support*, whereby "an action A expresses a motive M if and only if during the operation of the action-direction psychological mechanisms that are involved in the etiology of A, M exerts motivational influences (of sufficient strength) in favor of A-ing" (p. 1216). Taken together, these two components entail that "an action expresses one's self if and only if the motive expressed in the action is one of one's cares" (p. 1216).

Perhaps the Self-Expression view has certain advantages over cognitive attributionist theories,[68] but it doesn't appear to have the resources to excuse Dennis for his wrongdoing. Abstracting away from some of the finer details of Sripada's view, Dennis's omission in Difficulty for Dennis appears to be primarily motivated by a suite of dispositions that favor not helping a member of the relevant ethnic group. Perhaps this motive can be expressed in terms of a particular care, but it seems more accurate that the motive derives from a *lack* of care – namely, for the rights and interests of this ethnic group. Still, a lack of care can also ground responsibility. As Sripada acknowledges, "we sometimes say that an action is expressive not of something a person *does* care about, but rather what he *fails* to care about, his attitudes of disregard or indifference" (p. 1220). This kind of expressive absence is often characterized as *insufficient concern* by other accounts and seemingly applies to Dennis's situation.[69] After all,

[68] For example, Sripada (2016) argues that the Self-Expression account can better answer the "Which Judgments?" problem; that is, it can better identify the attitudes that genuinely reflect an agent's evaluative orientation.
[69] See, e.g., Arpaly and Schroeder (2013), who argue for a conative view grounded in *de re* concern for what is morally important.

we can assume that if Dennis adequately cared for the rights and interests of the relevant ethnic group, he would have helped. Therefore, Dennis's omission is properly expressive, and so he's blameworthy on this conative account as well.

Of course, one could construct a similar case in which Dennis's omission isn't expressive of his cares. Specifically, there's a version where Dennis acts *contrary* to his cares in refusing to provide aid. In this case, he properly cares for the rights and interests of all people, but his enculturated belief about the untrustworthiness of the relevant ethnic group motivates him to omit helping.[70] Therefore, Dennis's omission isn't motivated by a lack of care – either the wrongdoing isn't appropriately expressive, or it expresses a different care that isn't morally objectionable.[71] Although such a case might be possible, however, this isn't how I construe Difficulty for Dennis. Dennis's enculturated beliefs influence him so that he lacks a care for the rights and interests of the relevant ethnic group. Indeed, this seems like a realistic outcome of his upbringing. Given the original description of the case, then, Dennis clearly appears to be blameworthy on a conative account like Sripada's.

Ultimately, then, it seems that both cognitive and conative versions of attributionism must hold Dennis responsible for his wrongful omission, even though he's intuitively excused due to difficulty. Yet this result shouldn't be surprising. After all, as Talbert's comments demonstrate, attributionism is largely *ahistorical*; that is, what matters for responsibility is whether an action bears the right relation to the attitudes that constitute one's evaluative orientation, *not* how an agent came to have these attitudes. But epistemic difficulty is a morally

[70] Sripada (2016) states that "the care-based view is consistent with a person's caring for X and, due to the operation of a defeater, failing to make the relevant evaluative judgments regarding X, even in idealized epistemic circumstances" (p. 1210, fn. 12). Thus, Sripada admits that caring and judgment can come apart in this way.
[71] For example, perhaps Dennis's omission expresses a care for the norms and practices of his community. Such a care doesn't seem morally objectionable in the absence of awareness that these norms and practices conflict with for the rights and interests of all people.

relevant factor that often involves the etiology of one's attitudes.[72] In Difficulty for Dennis, Dennis acts wrongly because of objectionable attitudes, but the fact that his community made it extremely difficult for him to correct these attitudes seems morally relevant. Insofar as this factor intuitively influences responsibility, then, attributionism doesn't appear to have the resources to capture or explain it.

2. Case Two: Reckless Ralph

Cases involving epistemic difficulty, like Difficulty for Dennis, demonstrate that expressing a bad (or indifferent) will is insufficient for blameworthiness. This conclusion likely won't surprise the numerous theorists who hold that responsibility requires a kind of control that attributionism rejects. Indeed, following Susan Wolf's (1990) influential argument, many have argued that attributability is insufficient for accountability.[73] Still, it's more controversial among these theorists whether attributability is *necessary* for accountability – that is, whether blameworthiness requires that an action (or attitude) express a bad will. I argue that the following case of ignorant wrongdoing demonstrates that expressing a bad will is also unnecessary for blameworthiness.

> **Reckless Ralph**: Ralph works in the kitchen of a popular fast-causal restaurant. During a lunch rush one day, he gets an order from a customer with the instruction to omit pork from the entrée. Ralph correctly infers that this customer keeps kosher, and he's determined to respect her dietary laws. Because of this determination, Ralph considers whether he ought to write down her instruction on her order ticket, given that he will fulfill several orders before hers and could potentially forget. But because of the rush, he decides to trust his memory and not write down the instruction, as he's very good at remembering orders. Unfortunately, though, Ralph forgets and ends up adding pork to the entrée. This error ends up upsetting the customer, who realizes the mistake after inadvertently eating some pork.

[72] To be clear, the central problem for attributionism is that it can't capture certain intuitive cases of excuse due to epistemic difficulty. However, part of the *explanation* of this failure is the ahistoricism of the theory.
[73] It's worth noting that Wolf (1990) herself never used 'accountability' in explicating different forms of moral responsibility. Moreover, she has since changed her views regarding these different forms in ways that don't clearly map onto this distinction.

First, I assume that it's morally wrong to serve someone food against their wishes, especially if these wishes are tied to religious convictions. Thus, it should be uncontroversial that Ralph acts wrongly in the case. On the other hand, Ralph commits this wrong unwittingly – at the moment he adds pork to the entrée he's unaware that he's going against the customer's wishes.[74] Because of this, Ralph seems even more ignorant than Dennis; whereas Dennis is ignorant that his omission is morally wrong, Ralph is even unaware of his action under an appropriate description. If asked to describe his conduct, Ralph wouldn't say that he was going against the customer's wishes in the way that Dennis would admit to refusing to provide aid. While Dennis is only morally ignorant, then, Ralph is also *factually* ignorant.

Nevertheless, although Ralph might be more ignorant than Dennis at the time of wrongdoing, he's still intuitively blameworthy for his error.[75] Presumably, this is because he should have written down the customer's instruction when he had the chance. By choosing to rely on his memory, he seems *epistemically reckless* in the situation. I will say more about epistemic recklessness shortly, but note that Ralph doesn't appear to express a bad will in serving the food. After all, he judges that the customer's wishes are important and is determined to respect them. Even though Ralph voluntarily chooses to omit writing down the instructions, then, he doesn't appear to express any objectionable attitudes that would ground blameworthiness on

[74] At least, Ralph is *consciously* unaware of his actions under a certain description. It's controversial what kind of awareness is required for blameworthiness, although most hold that conscious awareness is too strong and some sort of weaker awareness should suffice. Still, I needn't take a stand on this issue at this moment, as I will argue that his earlier (conscious) recklessness is the source of his blameworthiness. Therefore, even those who might deny that Ralph satisfies the necessary awareness condition at the moment of wrongdoing will likely accept that he satisfies it at the moment he makes his reckless decision to trust his memory. In fact, as long as most have the intuition that Ralph is blameworthy for *some* reason that doesn't relate to the quality of his will, the case is problematic for attributionism.

[75] Perhaps there are those who don't initially share this intuition. Although I use cases to generate certain intuitions, I also try to justify these intuitions by appealing to relevant moral principles and considerations. In the proceeding section, I attempt to justify the intuition that Ralph is blameworthy by appealing to the notion of culpable recklessness in the management of one's epistemic state.

an attributionist account. His reasons for deciding to rely on his memory – the lunch rush and his normally strong memory – are morally unobjectionable. Therefore, Reckless Ralph is another counterexample to attributionism that relies on epistemic considerations – in this case, epistemic recklessness rather than difficulty. As with section 1, I will use the subsequent section to support this conclusion.

2.1 First, what justifies the intuition that Ralph is blameworthy? If we just focus on the moment when he puts pork in the entrée, it might seem like he's excused. After all, at that moment Ralph is unaware that he's failing to respect the wishes of the customer, and thus unaware of the facts that make his action wrong.[76] Yet, as we zoom out from the actual wrongdoing, we find grounds for holding him responsible for his actions. Specifically, there's the prior moment when Ralph makes a conscious decision to trust his memory and forgo writing down the customer's instruction. It's at this moment that he appears culpably reckless, and we can plausibly anchor his responsibility for the subsequent wrong to this moment. Had Ralph chosen to write down the customer's instruction, he would have avoided wrongdoing.

Hopefully it's fairly intuitive that Ralph acts recklessly in choosing to trust his memory, but it's worth briefly discussing the concept of recklessness. In the criminal law, recklessness is commonly one of four grades of mental states that constitute elemental *mens rea*, establishing (narrow) culpability for an offense[77] – the others being purpose, knowledge, and negligence. Although there are different formulations, the basic conception of recklessness holds that an agent is reckless about X if, and only if, (1) she is *aware* that there is a risk of X, and (2) running

[76] I'll have much more to say about awareness in chapter four, but assume that Ralph has completely forgotten the customer's instruction. In other words, he's not even *implicitly* aware that he's failing to respect the wishes of the customer.

[77] See Brink (2019) for an examination of different culpability concepts within the criminal law.

the risk of which she is aware is *unjustified*.[78] Although (inclusive) culpability for negligence is controversial within legal and moral theory, culpability for recklessness is not nearly so.[79] However, specific conduct will fall into different categories depending on how one understands awareness and the standard of unjustifiability.

In Reckless Ralph, I simply stipulate that Ralph is consciously aware that omitting to write down the customer's instruction risks forgetting them and thus causing harm. Given that conscious awareness is a particularly strong form of awareness, then, he satisfies any plausible construal of the awareness condition for recklessness. What about the justifiability of Ralph's decision to run the risk of causing harm?[80] Again, different theories apply different standards, but I maintain that Ralph's decision is clearly unjustifiable. One way to see this is by conceptualizing the standard of justifiability as a balancing test between the *risk* of harm, as a product of the probability and degree of harm, and the *reasons* for running the risk.[81] According to this standard, running certain risks are justifiable when the reasons for running them are weightier than the risks themselves. For example, it might be justifiable for someone to speed to the hospital, risking harm to other drivers, if there's a legitimate medical emergency. Yet, in Reckless Ralph his reason for running the risk – to keep up with the lunch rush – doesn't plausibly outweigh the risk of harm, even if the probability of harm is low because of his good

[78] I adapt this formulation from Edwards (2018). According to a prominent definition from the Model Penal Code, "an agent acts recklessly if she consciously disregards a substantial and unjustifiable risk that the material element exists or will result from her conduct" (§2.02). In my formulation, I abstract away from are the consciousness and substantiality criteria in the Model Penal Code definition, as I don't believe that these are fundamental to the concept of recklessness. Moreover, they introduce issues that would obscure the current discussion.
[79] By 'inclusive culpability' I mean "the combination of wrongdoing and responsibility or broad culpability that functions as the retributivist desert basis for punishment" (Brink, 2019, p. 347).
[80] I'm simplifying here by focusing on just one risk related to one candidate action. A more realistic assessment of Ralph's moral responsibility would include every conceivable candidate action and the risks associated with each one.
[81] I draw on Alexander and Ferzan (2009) here.

memory.[82] The calculus might eventually flip as the probability of forgetting approaches zero, but Ralph has an ordinary good memory. Because of this, Ralph is reckless in deciding to trust his memory and forgo writing down the customer's instruction, and ultimately blameworthy for putting pork in the entrée.[83]

More specifically, Ralph is *epistemically* reckless because his omission risks ignorance most proximally. Contrast Ralph with a chef who serves food that she reasonably believes might be spoiled, to avoid losing money. In this case, the chef's decision directly risks *harm*, rather than risking ignorance that could reasonably lead to harm – so it's not an instance of epistemic recklessness. Importantly, most instances of risking ignorance aren't blameworthy, even if there are moral stakes to remaining ignorant. I'm not plausibly blameworthy for refusing to learn how to defuse a bomb right now, for example, even if remaining ignorant risks unnecessary harm in the unlikely event that the information becomes pertinent; were that event to actualize, I wouldn't be blameworthy for the resultant harm because I could have informed myself now. Fortunately, I don't need to take a stand on when exactly risking ignorance is blameworthy to vindicate the upshot of Reckless Ralph. After all, not only is Ralph aware that he could forget the information, and that this would likely lead to harm, but he also clearly has insufficient reason for taking the risk. Because of this, Reckless Ralph isn't a dubious case – Ralph clearly acts recklessly and blameworthy in choosing to trust his memory.

[82] Even if the calculus is definite, might Ralph be excused due to the difficulty of getting it right, like the case of Andre before? I maintain that the cases are dissimilar for two main reasons. First, Andre's calculus involves two *moral* considerations that are approximately similar in weight, whereas Ralph's involves balancing one moral consideration (harm) against a non-moral one (keeping up with the lunch rush). Secondly, even though the risk of harm might be low, getting it right in Ralph's situation isn't nearly as difficult as Andre's – the calculus is rather clear. Thank you to Dana Nelkin for suggesting this possibility.

[83] Ralph is plausibly blameworthy for *two* things: (1) putting pork in the entrée, and (2) the consequential harm from this action. However, to simplify, I mostly talk in terms of the wrongful action that results in harm.

2.2 It's less clear that Ralph doesn't express a bad (or indifferent) will in Reckless Ralph. Especially if an indifferent will is synonymous with insufficient moral concern, it might seem that recklessness is a *paradigmatic* form of such indifference. The chef who recklessly serves food that she reasonably believes might be spoiled, for instance, clearly seems to express an indifferent will. Yet, although many cases of recklessness are also undoubtedly cases of an indifferent will, Reckless Ralph is meant to be an informative exception. I will further explain why it's an exception in the following section.

Focusing first on the moment of wrongdoing – when Ralph puts pork into the entrée – his action doesn't express a bad (or indifferent) will. Rather than acting from an objectionable evaluative judgment that the customer's wishes are unimportant, he fails to act on his virtuous judgment that these wishes *are* important. In this way, his action fails to express his good will, rather than expressing a bad will. Contrast Reckless Ralph, then, with a case where Ralph remembers the instruction but judges that the customer's wishes don't give him sufficient reason to act accordingly. In this alternate version, Ralph is morally ignorant that he ought to exclude the pork, but his evaluative judgment provides grounds for attributionist blame. In the original version, though, Ralph isn't even aware that he's going against the customer's wishes. If asked to justify his action in the moment, Ralph would presumably respond that he's simply preparing the dish as it normally comes; and if asked to justify his action *after* the harm occurred, he would cite his failure to remember. In either case, the action doesn't express a bad will.

However, although Ralph isn't *originally* blameworthy on an attributionist account, perhaps blameworthiness can be traced back to some earlier, related expression of bad will?[84]

[84] The term *tracing* is often used to refer to a specific strategy for capturing moral responsibility. My usage here is not meant to invoke this sense. Instead, I merely intend to suggest that attributionists can also anchor responsibility for some wrongdoing to a (properly related) prior action.

One clear candidate is the moment Ralph forgets the instruction. Indeed, Smith and other attributionists often argue that non-voluntary actions and attitudes, like forgetting, can express a bad will.[85] If Clarke forgets to fulfill his wife's request to pick up milk on his way home because he ultimately judges her interests unimportant, for example, then his unwitting omission is blameworthy.[86] In this case, his forgetting reflects an objectionable evaluative judgment toward his wife's interests, and so he's blameworthy for failing to pick up the milk as a consequence of this forgetting. Still, although forgetting is *sometimes* a target of blame for attributionism, Ralph's instance isn't a plausible anchor point. Again, Ralph judges that the customer's wishes are important and is determined to respect them. In this way, his attitudes are crucially different than Clarke's.

Now, one could argue that if Ralph *really* respected the customer's wishes, he would have remembered; but this is implausible. This claim posits a necessary relation between certain actions and attitudes that intuitively have a much weaker connection. Indeed, both Smith and Talbert admit that although forgetting *can* reflect a bad will, there's no necessary connection between forgetting and expressing a certain evaluative orientation. Because of this, forgetting morally relevant information isn't always blameworthy. In reference to Clarke forgetting the milk, for instance, Talbert (2017a) grants that "if a condemnable lack of concern for his wife's interests played no role in explaining Clarke's omission, then he isn't a proper target for blaming responses" (p. 57). Reckless Ralph is meant to be analogous to this version of the Clarke case. Indifference toward the customer's wishes and interests played no role in explaining Ralph's wrongdoing because he wasn't indifferent. The forgetting that explains Ralph's actions was caused by situational factors like the stress of the lunch rush, not any sort of lack of concern or

[85] See, e.g., Smith (2005, 2008, 2017).
[86] I draw this example from Clarke (2014).

indifference.[87] Therefore, attributionism can't plausibly blame Ralph's wrongdoing on the basis of his forgetting.

A more promising strategy for attributionism is grounding blame in Ralph's reckless decision to trust his memory, given that this is the actual source of his blameworthiness on my understanding of the case. Again, though, this strategy seems to require a necessary relation between certain actions and attitudes that intuitively have a much weaker connection. As mentioned at the outset, recklessness might *often* express a bad (or indifferent) will, but Reckless Ralph is an exception. Just as attributionists admit that there's no necessary connection between forgetting and expressing a bad will, then, they should concede that there's no necessary connection between recklessness and expressing a bad will. After all, if recklessness is ultimately just a matter of consciously taking unjustifiable risks,[88] then it's doubtful that every instance involves a bad will. The key point here is that there's nothing inherent to the notion of conscious unjustified risk-taking that entails a bad will, even if most actual risk-taking of this sort involves a bad will; and as long as there's no such necessary connection, attributionism is vulnerable to counterexamples involving intuitively blameworthy recklessness without bad will.

To further illustrate this point, suppose that Ralph correctly assessed the probability that he would forget the customer's instruction, but incorrectly balanced this risk against the value of keeping up with the lunch rush. Under these circumstances, his reckless decision to trust his memory seems blameworthy but needn't express a bad will. Instead, it could be that although Ralph takes the customer's interests to provide strong reasons for compliance, the low

[87] Amaya and Doris (2015) make a similar point regarding cases like Reckless Ralph in their discussion of *performance mistakes*, arguing that "many of the mistakes people make are not a reflection of deep seated attitudes in them but are rather due to small lapses of concentration and memory" (p. 264).
[88] To be clear, the current conception of recklessness only requires (*de dicto*) awareness of a risk that is *in fact* unjustifiable; it does not require (*de re*) awareness *that* this risk is unjustifiable. This is consistent with the common criminal law principle that ignorance of the law is no defense.

probability of forgetting lead him to mistakenly believe that he was justified in taking the risk.[89] It's worth pointing out that this kind of calculated risk is often intuitively justified. In another situation, for example, Ralph might come to realize that there's some nonzero risk that the restaurant's water heater could explode, causing significant harm to everyone in the vicinity. If the risk is infinitesimal, he might decide to forgo a safety inspection at that exact moment in favor of keeping up with the lunch rush. This decision is surely justified, even if it involves outweighing legitimate moral reasons with considerations that wouldn't normally counterbalance them. More importantly, it's implausible that Ralph expresses a bad will by not performing an immediate inspection, given that he's taking a justified risk. Similarly, it's implausible that he expresses a bad will in Reckless Ralph just because he miscalculates. Such miscalculation simply isn't plausibly expressive under these circumstances. As with forgetting, then, there's no necessary connection between recklessness and an objectionable evaluative judgment.[90]

Finally, it's worth noting that conative attributionist views have no significant advantage in capturing cases like Reckless Ralph. Using Sripada's framework, Ralph's recklessness isn't motivationally supported by any cares, nor any expressive lack of cares. If anything, Ralph's reckless decision to trust his memory actually *belies* his determination to respect the customer's wishes, which plausibly derives from genuinely caring for the legitimate interests of others.

---

[89] Alternatively, one could imagine that Ralph slightly underestimates the probability that he would forget the customer's instruction. In this case, Ralph's miscalculation would involve underestimating the risk, rather than overestimating the reason for running the risk (i.e., the value of keeping up with the lunch rush). I leave it to the reader to choose which version he/she finds more plausible as an instance of blameworthy recklessness without bad will. For my part, I believe that both versions represent counterexamples to attributionism.

[90] In fact, the connection appears even weaker in the case of *epistemic* recklessness, as the relation between the action and consequences is even more indirect. Thus, Ralph's recklessness doesn't risk wrongdoing itself, but only forgetting information that consequently risks wrongdoing. Indeed, we can imagine a scenario where Ralph forgets the instruction but nevertheless omits the pork anyway. Because of this, Ralph can't be certain that risking ignorance necessarily risks wrongdoing. All cases of epistemic recklessness have this same feature, namely, that there are really two risks involved: (1) the risk of ignorance, and (2) the risk that this ignorance will lead to wrongdoing. This indirect relation makes it less likely that the epistemic recklessness can be traced to an objectionable evaluative attitude.

When making this decision, he simply miscalculates.[91] Insofar as his recklessness expresses any motive, then, it's simply a desire to keep up with the lunch rush. Such a motive doesn't plausibly constitute a care on Sripada's account, and so Ralph's wrongdoing doesn't express his deep self.[92] Indeed, this result is intuitive; if the notion of a deep self is meant to capture anything practically significant, then shallow motives like this desire can't be expressive of it. Ultimately, then, conative attributionist views have similar deficiencies in capturing cases like Reckless Ralph; namely, they can't explain why recklessness is blameworthy in certain intuitive cases.[93]

3. Attributionist Response

The previous two cases demonstrate that expressing a certain quality of will (or deep self) is neither necessary nor sufficient for blameworthiness. Both cases rely on epistemic considerations to support their respective intuitions. In Difficulty for Dennis, it's the epistemic difficulty of ascertaining the moral truth that excuses Dennis, despite him expressing a bad will. In Reckless Ralph, it's the reckless mismanagement of his epistemic situation that makes Ralph blameworthy, despite him not expressing a bad (or indifferent) will. If these intuitions are convincing, then the only non-revisionary move for attributionism is to try to capture these cases

[91] One possible explanation of this error is that the saliency of the lunch rush causes Ralph to overestimate the value of keeping up with it. Indeed, it's common for people to have these sorts of immediacy biases when balancing competing considerations.

[92] As Sripada (2016) explains, "when we consider the vast array of motivational attitudes within a person's psychology, since caring requires being a source of intrinsic motivation, it must be fairly rare. Most every other motivational state rises and falls in the service of other more basic motives. Cares are distinctive in lying exclusively at the foundations of this hierarchy of motives" (p. 1209). Thus, it's implausible that Ralph's desire to keep up with the lunch rush is such a foundational motive, or that it necessarily derives from one.

[93] Although I consider Sripada's account to be the paradigm conative attributionist view, a more thorough investigation would discuss the intricacies of different versions. Perhaps the most notable omission from the above discussion is Arpaly and Schroeder's (2013) view, which explicates the moral worth of an action in terms of the agent's intrinsic desire (*de* re) that morality be followed. Although this kind of account might have the best chance of capturing intuitions in Reckless Ralph, it's also the most peripherally attributionist. Indeed, Vargas (2020) classifies it as a reasons-responsiveness view, like Nelkin (2011).

by modifying or supplementing the core theory.[94] In this section, I explore and evaluate possible strategies for attributionism to capture these cases. Specifically, I focus on strategies for capturing intuitions in Difficulty for Dennis, as there's more indication how attributionists might respond to this alleged counterexample.

3.1 One way that attributionism might try to explain why Dennis is excused is by maintaining that expressing a certain quality of will is merely *necessary* for moral responsibility. In fact, Sripada's (2016) explicit formulation of his deep self theory understands self-expression as necessary but not sufficient for responsibility, and he mentions that a possible further condition could be an *epistemic* requirement. Because of this, it might appear that his account already has the resources to capture Difficulty for Dennis by appealing to an epistemic condition that Dennis fails to satisfy because of the relevant difficulty of ascertaining the moral truth. But although Sripada hedges by entertaining such a condition, it's important to note that his considered view is that "these epistemic conditions are already fully *built into* a deep self theory's requirement of self-expression…there isn't a separate freestanding epistemic requirement that one must meet over and above the requirement of self-expression" (p. 1205, fn. 2). In this way, epistemic considerations influence moral responsibility only by affecting the expression of an agent's deep self, and so they're just another factor that can *indirectly* influence responsibility within an attributionist framework.

It's ultimately unsurprising that Sripada's considered view attempts to incorporate epistemic considerations into the theory's central condition of self-expression. Although nothing prevents attributionism from adding further conditions for moral responsibility, there are

[94] The relevant *revisionary* move would be to argue that these intuitions are mistaken, and we should instead accept the attributionist conclusion that Dennis is responsible, and Ralph is not. But this move would require not just going against intuitions in these particular cases, but all cases that are relevantly similar. I assume, then, that the attributionist would prefer to attempt to capture these intuitions.

significant theoretical costs for such additions. First, simply adding something like an epistemic condition would make the resultant view rather disjointed and ad hoc. Certainly, many theories of moral responsibility assert multiple conditions, but these conditions are usually unified by some broader concept. Reasons-responsiveness theories, for instance, often unite various cognitive, volitional, and situational conditions by appealing to underlying considerations of control or fairness.[95] Similarly, attributionism would need to explain why an epistemic condition should supplement the central notion of an agent's evaluative orientation. Perhaps the attributionist could argue that epistemic considerations are also *directly* relevant to the broader notion of attributability,[96] but this argument doesn't seem promising. Insofar as epistemic considerations seem germane to attributability, it's because of their relevance to quality of will, not because they directly relate to a certain kind of moral appraisal. Yet this just reinforces that such considerations aren't best captured as an independent condition within attributionism.

Secondly, and more importantly, incorporating further conditions for moral responsibility risks undermining foundational commitments of attributionism. One such commitment is the rejection of the requirement that agents possess certain substantive forms of control over their actions to be morally responsible. Smith (2005), for example, asserts that "what makes us responsible for our [actions and] attitudes is not that we have…voluntary control over them but that they are the kinds of states that reflect and are in principle sensitive to our rational judgments" (p. 271). If attributionism admits that epistemic considerations can bear directly on

[95] See, e.g., Brink (2021) and Brink and Nelkin (2013), who unify conditions for moral responsibility under the concept of the *fair opportunity to avoid wrongdoing*. Some reasons-responsiveness theorists prefer to set aside the epistemic condition altogether and simply focus on the control condition. For these theorists, who seemingly admit that they offer a partial theory of moral responsibility, there doesn't seem to be a theoretical cost to later incorporating epistemic conditions, as long as they offer some unifying explanation.

[96] By *directly* relevant I mean that an agent's epistemic situation is somehow central to her evaluative orientation. Difficulty for Dennis and Reckless Ralph suggest, however, that epistemic considerations are separable from evaluative orientation. Therefore, these considerations are unlikely to be essential to the notion of attributability.

moral responsibility, though, this move risks reintroducing these kinds of control as necessary for responsibility. After all, the best explanation for why Dennis's ignorance is excusing plausibly appeals to a certain lack of control that he had in forming and revising his enculturated beliefs. More specifically, the exculpatory difficulty Dennis experienced in ascertaining the moral truth resulted from his community impeding his ability to exercise certain cognitive capacities that are central to reasons-responsive control.[97] Many other cases involving epistemic considerations similarly undermine responsibility by plausibly compromising forms of control that attributionists fundamentally reject as necessary. If this is right, then an independent epistemic condition will generate an overall account that is not recognizably attributionist.[98]

3.2 A more promising strategy, hinted at in the introduction, involves asserting a distinction between moral responsibility and blameworthiness. On this approach, the attributionist would maintain that Dennis is morally responsible in the attributability sense but argue that he's not *blameworthy*.[99] In fact, it's common for attributionists to stress this distinction in other contexts. Smith (2015), for instance, emphasizes that she takes "the most basic question of moral responsibility to be *prior to* questions of moral praiseworthiness or blameworthiness" (p. 128, fn. 8). Still, if Dennis's wrongdoing expresses a bad will, then the attributionist needs to explain why he's excused from blame. After all, regardless of the theoretical validity of the distinction, expressing a bad will is normally sufficient for both responsibility *and* blameworthiness.

---

[97] Cf. Nelkin (2016, p. 371).

[98] Consider the following passage from Talbert (2022): "Angela Smith notes that what is distinctive about attributionism is *not* its interest in specifying grounds for attributing attitudes, actions, and omissions to agents—many other approaches share this feature (Smith, unpublished). Instead, and as I meant the added emphases above to bring out, the distinctive feature of the view is that the relevant attributions are taken to be *all* that is required for responsibility" (p. 54). This shared view between Talbert and Smith further corroborates the claim that an account that requires the kinds of control discussed above is not recognizably attributionist.

[99] Note that this strategy will do nothing to capture the intuition that Dennis isn't morally responsible. I argue that Dennis isn't blameworthy *because* he isn't morally responsible, according to a conception of responsibility as accountability. However, I believe that it's pretheoretically intuitive that Dennis isn't morally responsible.

Of course, the most natural explanation of Dennis's excuse appeals to the difficulty in ascertaining the moral truth; but the attributionist owes an explanation for *why* this difficulty precludes blameworthiness when the agent nevertheless expresses a bad will. This explanation can't appeal to considerations that undermine foundational commitments of attributionism either, otherwise the ultimate account won't be recognizably attributionist. For her part, Smith appeals to the notion of *reasonable expectations* to distinguish blameworthiness from responsibility. This strategy is particularly evident in Smith's (2008) response to an argument from Neil Levy (2005) against attributionism. Levy argues that attributionists must maintain that agents are blameworthy for actions that cause harm, even if they had no way of knowing that they were causing harm, because these actions are still attributable to them. In response, Smith accuses Levy of conflating responsibility and blameworthiness, arguing that "the attributionist, like the volitionist, can say that an agent is blameworthy for an action or attitude only if that action or attitude violates a moral norm we expect reasonable persons to accept" (p. 390, n. 30). Moreover, she explicitly cites *epistemic* considerations as relevant to these expectations, and so it seems plausible that she might make a similar argument in response to Difficulty for Dennis.

Reasonable expectations actually play a fundamental role in many *non*-attributionist theories of moral responsibility,[100] and so it might seem that there is some independent plausibility to Smith's appeal to these expectations. Indeed, my own analysis of Difficulty for Dennis cited reasonable expectations as part of the explanation of his excuse. To assess this move, however, it's important to be clear about the nature of these expectations. Following Rudy-Hiller (2020), I contend that the relevant normativity of these expectations derives from both "a legitimate moral demand the agent is subject to and her having a fair opportunity to

[100] See, e.g., Brink (2021), Brink and Nelkin (2013), FitzPatrick (2017), Nelkin (2016), Rosen (2003), Wallace (1994), and Watson (1996).

comply with it" (p. 2948). Notice that this conception of reasonable expectations isn't merely *predictive*. In other words, it's pertinently unreasonable to expect certain behavior from an agent in a given situation just because one would predict this behavior, perhaps based on statistical information.[101] Instead, reasonable expectations are based on the propriety of the constituent demands and opportunities. To my knowledge, attributionists haven't offered a competing account of reasonable expectations, and so I work within this framework going forward.

Assuming this account, how might the attributionist deny Dennis's blameworthiness based on reasonable expectations? It seems that the attributionist must at least refute either the legitimacy of the relevant demands or the existence of a fair opportunity. Yet, the former approach seems implausible. After all, if demands are based on obligations, then we can intuitively demand that Dennis provide aid, given that this moral obligation plausibly still applies to him. This intuition reflects the important distinction between (a) being excused for failing to meet a demand and (b) not being legitimately subject to a demand in the first place. Difficulty for Dennis represents a case of (a), not (b). The difficulty of ascertaining the moral truth doesn't undermine Dennis's obligation to provide aid, it simply excuses him for failing to meet the associated demand. Therefore, appealing to this component of reasonable expectations won't help the attributionist explain why Dennis isn't blameworthy.

Instead, then, the only move currently available to the attributionist is to deny Dennis's blameworthiness based on a lack of fair opportunity to comply with the relevant demand. This route is actually more intuitive, as Dennis's social context seemingly deprived him of this opportunity by making it sufficiently difficult to ascertain the moral truth. But the attributionist

[101] Smith (2017) also explicitly denies a conception of reasonable expectations that is merely predictive: "What is reasonable, in other words, is not someone's prediction about how someone is likely to behave, but rather a norm specifying how someone ought (in some sense) to behave" (p. 46).

owes an account of fair opportunity that doesn't appeal to considerations that undermine foundational commitments of attributionism.[102] One natural candidate would be something along the lines of Smith's earlier response to Levy, which appeals to epistemic considerations that are also central to Difficulty for Dennis. Unfortunately, Smith doesn't elaborate much on this response, but the details of such an account can be found elsewhere in her work. Although Smith (2017) never explicitly references fair opportunities, in discussing unwitting omissions she argues that "an agent can only be said to 'deviate from' or to 'violate' a relevant practical norm if she either knew about of should have known about the existence of that practical norm in the circumstances" (pp. 53-4). The way Smith phrases this "knowledge condition" seemingly invokes legitimate demands,[103] but we can simply convert it to a criterion for fair opportunity. The crucial claim is that blameworthiness requires that the agent was aware or *should* have been aware of the relevant practical norm.

Regardless of the independent merits of such a condition,[104] though, the problem is that it's unclear how attributionism could ground it. The most intuitive explanations for why blameworthiness requires that the agent was aware or should have been aware of the relevant practical norm would appeal to considerations that undermine foundational commitments of attributionism. For example, one might naturally appeal to the way that Dennis's community impeded his ability to exercise certain cognitive capacities to explain why it's false that he

[102] For example, Brink (2021) and Brink and Nelkin (2013) advocate a reasons-responsiveness view explicitly grounded in a conception of blameworthiness that entails a fair opportunity to avoid wrongdoing, where such an opportunity requires a kind of control that is denied by attributionism. Attributionists can't appeal to this kind of control and remain recognizably attributionist.

[103] That is, Smith (2017) seems to assert that an agent is only subject to a legitimate demand if she knew about of should have known about the existence of that practical norm in the circumstances. Yet, as we saw before, this claim is implausible. Instead, it's more intuitive that the agent is subject to the demand (whether she knew about, or should have known about it, or not), but that she's excused for failing to meet it.

[104] Indeed, many non-attributionists argue for some version of this condition. See, e.g., Clarke (2014), Murray and Vargas (2020), and Rudy-Hiller (2017).

should have known that he ought to provide aid in the situation. Yet it's precisely these kinds of abilities for control that attributionism rejects as necessary for moral responsibility, and so it would be incongruous for attributionism to rely on them to explain blameworthiness in these cases.[105]

3.3 Finally, Gunnar Björnsson (2017) provides a related attributionist strategy for capturing cases like Difficulty for Dennis.[106] Instead of relying on the distinction between moral responsibility and blameworthiness, however, Björnsson seemingly builds reasonable expectations directly into his quality of will account.[107] On this conative theory, blameworthiness "requires not only that the quality of will be bad. It also requires that the quality fall below what can be *properly demanded*, which in turn depends on the difficulty of caring…" (p. 154, my emphasis). Setting aside the conative component of this account, I take it that these proper demands are essentially equivalent to the legitimate moral demands that partially constitute the current interpretation of reasonable expectations. If this is right, then reasonable expectations are integral to Björnsson's theory of moral responsibility in that they determine what level of bad will is sufficient for blameworthiness.

Unfortunately, Björnsson doesn't give a full account of the nature of the demands that factor into his account. As evident in the above quote, though, difficulty is a key consideration in

[105] Of course, there's nothing inherently contradictory about having different conditions for moral responsibility and blameworthiness. However, attributionists generally reject these kinds of abilities (or opportunities) as necessary for *blameworthiness* as well, and so it seems inconsistent to rely on them to differentiate moral responsibility from blameworthiness in Difficulty for Dennis. Furthermore, it's hard to see how the influence of Dennis's community could only affect his blameworthiness, and not his moral responsibility. Although there might be considerations that only bear on one's blameworthiness, an agent's ability to exercise his cognitive capacities doesn't seem like one of them.

[106] Although Björnsson (2017) advocates a quality of will theory of moral responsibility (and blameworthiness), it's not clear that he endorses the foundational commitments of attributionism. In this way, his account might be more similar to non-attributionist quality of will views, like McKenna's (2012, 2013). For the purposes of this section, though, I'm only concerned with his account insofar as it presents a possible strategy for attributionism to capture cases like Difficulty for Dennis. My argument is that this strategy faces similar issues as the previous strategy.

[107] An added benefit of building reasonable expectations into one's theory or moral responsibility is that it potentially allows one to capture that intuition that Dennis is neither blameworthy *nor* morally responsible.

what can be properly demanded, which is obviously relevant to Difficulty for Dennis. Björnsson (2017) explains the nature of this difficulty as follows:

> The difficulty of achieving a certain language proficiency, for example, depends on how much of the language one can access and what type of feedback is available. Similarly, how easily I can sustain a concern for various morally relevant aspects of my actions might depend on the extent to which such a concern is supported by circumstances: it matters whether others share and voice their support for such concerns, and it might matter what sorts of opportunities I have to act on such concerns and to make them part of my identity (cf. Vargas 2013: Chs. 7-8). (pp. 154-5)

It's important to note that this passage occurs within a discussion about difficult *epistemic* circumstances.[108] Given this context, Björnsson seemingly asserts that certain epistemic circumstances – such as an agent's opportunities – can make it more difficult to achieve and sustain an adequate quality of will, which in turn influences reasonable expectations regarding quality of will. If this is right, then Björnsson might plausibly excuse Dennis by appealing to the lack of opportunity that his circumstances provided him to cultivate a sufficiently good quality of will. It's precisely this lack of opportunity that explains why it would be unreasonable to expect him to express a better quality of will in Difficulty for Dennis.

Despite the intuitive plausibility of this explanation, however, appealing to a lack of opportunity to cultivate a certain quality of will seems to violate foundational commitments of attributionism. As mentioned before, attributionism is largely ahistorical – what matters for responsibility is whether an action bears the appropriate relation to the privileged attitudes that constitute one's will, and how an agent came to have these attitudes is largely irrelevant. Indeed, it's rather telling that Björnsson cites Manuel Vargas (2013) in the passage above, when mentioning opportunities, given that Vargas explicitly rejects an attributionist gloss on the

[108] Specifically, Björnsson is responding to a family of cases from Rosen (2003) that involve difficult epistemic circumstances that allegedly undermine or mitigate blameworthiness. Björnsson contends that his quality of will account can account for the intuition that blame is significantly undermined in these cases.

significance of these features. Ultimately, it's hard to reconcile the relevance of such opportunities with an attributionist theory of moral responsibility. While Björnsson might have identified resources for his unique quality of will theory to capture cases like Difficulty for Dennis, then, they are not resources available to attributionism.[109]

3.4 Given these problems incorporating reasonable expectations into an attributionist theory, perhaps it's unsurprising that many attributionists simply reject any such requirement on blameworthiness. Talbert (2017a), for example, explicitly states that "the reasonable expectation that a person avoid wrongdoing is neither necessary nor sufficient for blameworthiness" (p. 48). Instead of trying to capture the relevant intuitions, then, Talbert simply denies that it's unfair to blame agents like Dennis, for whom it wasn't reasonable to expect them to act otherwise. Although this strategy avoids undermining what is distinctive of attributionism, though, I contend that it's unacceptably revisionary. As argued before, Dennis's social context made it sufficiently difficult to ascertain the moral truth in such a way that he's intuitively excused (at least, partially). Moreover, we can imagine many similar situations in which difficulty plausibly excuses wrongdoing based on reasonable expectations. Therefore, attributionism is better off trying to capture these intuitions than denying them. But as I argue, the prospects of this endeavor aren't promising.

Conclusion

In this chapter I argued that attributionism has difficulty capturing intuitions in certain cases involving epistemic considerations. Specifically, I presented two cases demonstrating that

[109] Björnsson (2017) and McKenna (2012, 2013) both advocate non-attributionist quality of will views, and both seemingly appeal to resources from the reasons-responsiveness tradition. In this way, their views represent *hybrid* approaches. I save evaluation of these hybrid theories for the next chapter when I discuss reasons-responsiveness views in much more detail.

expressing a certain quality of will is neither necessary nor sufficient for blameworthiness. Difficulty for Dennis showed that an agent can still be excused, despite expressing a bad (or indifferent) will, if his epistemic situation made it sufficiently difficult to realize the moral truth. Reckless Ralph showed that an agent can still be blameworthy, despite expressing a good will, if he was reckless in the management of his epistemic situation. Because of this, attributionism (both the cognitive and conative variety) appears extensionally inadequate in capturing intuitions that track the influence of these epistemic factors.

In section 3, I explored possible responses on behalf of attributionism and found them wanting. In particular, I focused on strategies for capturing intuitions in Difficulty for Dennis, as there's more indication how attributionists might respond to this case.[110] In evaluating these strategies, a dilemma seemed to emerge: attributionists can gain extensional adequacy only if they draw on resources that undermine foundational commitments of attributionism. Specifically, capturing intuitions in these cases appears to require certain capacities and opportunities characteristic of a kind of control that attributionism constitutively rejects as necessary for moral responsibility. If this is right, then attributionism must offer a revisionary account of these cases. I argued that this is too high of a price to pay, but ultimately attributionism should be judged against its rivals. If these rival approaches end up being even more revisionary, then we have reason to reconsider attributionism. In the next chapter, I discuss and evaluate a major alternative to attributionism. Rather than eschewing capacities, these views put them at the center of responsibility.

[110] Indeed, there's much more focus in the literature on the *insufficiency* of attributability for blameworthiness. Because of this, there's more evidence of how an attributionist might respond to a case like Difficulty for Dennis. It's less clear how an attributionist might try to capture Reckless Ralph. One possible strategy might be to expand the mental states that constitute an agent's quality of will, so that acting with awareness that one unjustifiably risks ignorance might express a bad will. A problem with this strategy, though, is that expanding the states the constitute an agent's quality of will risks holding her counter-intuitively responsible for other actions that appear excused.

CAPACITARIANISM

Introduction

Examining the inadequacies of attributionism more closely can help suggest better approaches for capturing epistemic considerations. In Difficulty for Dennis, evaluative orientation failed to explain intuitions because the difficulty of realizing the moral truth excused Dennis's wrongdoing despite him expressing a bad will. It is worth considering, then, why exactly difficulty excuses in this case, and whether there are theories of moral responsibility that better support this explanation.[111] One rather intuitive answer is that Dennis's social context sufficiently impeded his ability to come to the correct moral belief and avoid wrongdoing. In other words, his circumstances compromised his *capacity* to recognize the reasons that bear on his action in the situation. Because of this interference, it's false that he could and should have known better, and so it's unfair to hold him responsible for his wrongdoing.[112]

Indeed, this kind of capacities-based explanation is well represented in the literature on epistemic considerations for moral responsibility. Although individual views vary, so-called capacitarian accounts are unified in holding that "a central element in the explanation of blameworthiness for ignorant wrongdoing is the fact that the agent should and could have known better than she did" (Rudy-Hiller, 2022).[113] Such views trace their lineage to an analogous approach to explaining criminal liability for *negligence*;[114] specifically, H.L.A. Hart's (1968b)

---

[111] Guerrero (2017, pp. 216-7) disagrees that this kind of difficulty ("difficulty in trying") excuses in cases like Difficulty for Dennis. Because of this, he claims that *agential revelation views* offer a more plausible diagnosis of these cases.

[112] One could deny that Dennis is sufficiently incapacitated and still agree with the normative claim that *were* he incapacitated he would be excused because it would be unfair to hold him responsible. The plausibility of the current explanatory strategy doesn't rely on one's intuitions about Dennis's specific capacities.

[113] Capacitarian views include Clarke (2017), Murray and Vargas (2020), Rudy-Hiller (2017), and Sher (2009).

[114] There are various definitions of negligence within legal theory and criminal law. According to the Model Penal Code (1985) version, "a person acts negligently with respect to a material element of an offense when he should be aware of a substantial and unjustifiable risk that the material element exists or will result from his conduct" (§ 202(2)(d)). For current purposes, the key features of negligence are (1) lack of an awareness condition and (2) the

influential account grounding culpability for negligence in unexercised capacities to avert to the risk of harm. Note that, unlike attributionism, capacitarianism is essentially modal, meaning that blameworthiness is based on what agents *could* have done in certain situations. This modality is key to capturing cases where an agent never actually recognizes the pertinent risks and reasons, which I discuss in detail later in the chapter.

Strictly speaking, 'capacitarianism' usually refers to an approach to explaining blameworthiness for ignorant wrongdoing, not an overall theory of moral responsibility. Nevertheless, the focus on capacities for recognizing morally relevant features is most closely associated with the reasons-responsiveness tradition. Because of this, most capacitarian accounts derive from broader reasons-responsiveness theories, even though not all reasons-responsiveness views are capacitarian, and vice versa.[115] In the following chapter, I focus on reasons-responsiveness theories that take the capacitarian approach to capturing epistemic considerations. Recall that reasons-responsiveness theories explain moral responsibility in terms of agents' capacities for responsiveness to reasons, especially moral reasons.

In section 1, I explain how a reasons-responsiveness theory might use the capacitarian approach to explain intuitions regarding certain key cases of ignorant wrongdoing. In particular, I focus on cases of unwitting omissions where the wrongdoers appear blameworthy. According to capacitarianism, these agents are blameworthy because they possess unexercised capacities to become aware that they are omitting and thus risking harm. In section 2, I then explore two types of justifications for why these unexercised capacities should render agents blameworthy, a backward-looking and forward-looking justification. The backward-looking justification appeals

claim that the agent *should* have been aware. Negligence is generally contrasted with *recklessness*, where the agent is aware of the relevant risk, and *strict liability*, where it's not the case that the agent should have been aware.

[115] For example, Sher (2009) is an attributionist who argues for capacitarianism, and Haji (1997) is a reasons-responsiveness theorist who doesn't espouse capacitarianism.

to considerations of fairness and reasonable expectations, while the forward-looking one appeals to the value of a certain kind of self-control. In section 3, I present two main problems for capacitarianism: the first relates to *capacity* claims, and the second relates to the *justification* of blame. Finally, in section 4, I present and evaluate what I take to be the strongest evidence in favor of capacitarianism; namely, the tendency for people to blame agents in cases of ignorant wrongdoing. In response, I argue that there are reasons to put less weight on the significance of this tendency. Ultimately, I conclude that although capacitarianism has significant advantages over attributionism, it's an unsatisfactory account because it rejects a necessary condition for blameworthiness: *awareness of risk*.

1. Capacities

Before turning to how capacities might ground blameworthiness, it's worth briefly considering how capacities can *excuse* blame. Although I previously suggested that a capacitarian account could excuse Dennis, it will be instructive to examine a more straightforward case. For example, suppose that Abby spoons arsenic into Martha's tea, thinking it's sugar, causing Martha's death by poisoning.[116] Moreover, suppose that Abby had no reason to suspect that the substance in the sugar bowl was anything other than sugar; perhaps her mortal enemy surreptitiously switched the sugar out for arsenic. Is Abby blameworthy for Martha's death? Surely not. According to attributionism, Abby is excused because her accidental poisoning doesn't express a bad will. But even if this is true about her evaluative orientation, the more fundamental explanation of Abby's excuse is that she couldn't have known it was arsenic – or perhaps more accurately, it's false that she *should* have known. After all, although Abby hypothetically could have known it was arsenic had she chemically tested the substance, it's

[116] This is an adaptation of a frequently cited example from Rosen (2004).

unreasonable to expect her to do so. Therefore, she didn't have sufficient capacity to know better, and so her ignorant wrongdoing is excused.[117]

Yet, although capacitarianism offers a better explanation of excuse for many cases of blameless ignorant wrongdoing, it's most significant advantage over attributionism is in capturing certain cases of intuitive blameworthiness. Specifically, capacitarianism appears to have resources to support intuitions regarding cases of *fully* unwitting wrongdoing, where it's implausible that the agent expresses a bad will. In these cases, not only is the agent unaware that what she does is wrong, but she's also unaware of the considerations that make her actions wrong. In other words, she's both morally *and* factually ignorant. This situation contrasts with a case like Difficulty for Dennis, where Dennis is aware that he's omitting to provide aid to someone in need.

What's an example of the kind of fully unwitting wrongdoing that capacitarianism is seemingly well equipped to capture? One of the most cited cases comes from George Sher (2009):

> **Hot Dog:** Alessandra, a soccer mom, has gone to pick up her children at their elementary school. As usual, Alessandra is accompanied by the family's border collie, Bathsheba, who rides in the back of the van. Although it is very hot, the pick-up has never taken long, so Alessandra leaves Sheba in the van while she goes to gather her children. This time, however, Alessandra is greeted by a tangled tale of misbehavior, ill-considered punishment, and administrative bungling which requires several hours of indignant sorting out. During that time, Sheba languishes, forgotten, in the locked car. When Alessandra and her children finally make it to the parking lot, they find Sheba unconscious from heat prostration. (p. 24)

It's worth pointing out several important features of Hot Dog. First, Alessandra's wrongful omission to let Sheba out of the car is fully unwitting. At the time that Sheba languishes,

[117] Note that these are rather *specific* capacities, in contrast with the general capacities that might be lacking in cases of the insane or immature.

Alessandra is unaware that Sheba is in the car, due to the distracting circumstances.[118] Second, it's implausible that Alessandra's actions necessarily express a bad will. Even though forgetting information can sometimes reflect an objectionable evaluative judgment toward others' interests, in this case Alessandra's forgetfulness results from the circumstances. Third, unlike in Reckless Ralph, there's no clear prior recklessness on Alessandra's part. She never appears to consider the risk of forgetting Sheba in the car, and it's not even certain that running that risk would be unreasonable.[119] After all, it's normal for agents to regularly run minor risks, and this is usually justified. Had Alessandra experienced issues with the pickup in the past, then perhaps she should have either taken Sheba with her or made arrangements so that she wouldn't forget. But given her past experiences, it seems reasonable to leave Sheba in the car.[120]

Nevertheless, although Alessandra is neither reckless nor indifferent, capacitarianism contends that she's intuitively blameworthy. The basic reasoning for this judgment is that Alessandra *could* and *should* have remembered that Sheba was in the hot car. In other words, unlike excusing cases, her distracting conditions weren't strong enough to sufficiently undermine her capacity to remember. Santiago Amaya and John M. Doris (2015) describe cases like Hot Dog as *performance* mistakes rather than instances of undermined *competence*.[121] Again, the idea is that Alessandra possesses the relevant capacity to remember that Sheba is in the hot car,

[118] At least, under a certain understanding of 'awareness.' I will have much more to say about the concept of awareness in the next chapter, but for now it should at least be clear that Alessandra isn't actively entertaining the thought that Sheba is in the car.
[119] Nelkin and Rickless (2017) disagree, arguing that "when the consequences of error would be devastating, there arises a stringent duty to take whatever steps would be required to avoid error" (p. 123). I contend that this duty doesn't necessarily apply to cases where the probability of such an error is sufficiently low. However, it's unclear whether Hot Dog passes this threshold.
[120] Although I believe that Alessandra might be justified in running the risk of forgetting Sheba in the car under certain circumstances, the more important point here is that her lack of *awareness* of the risk means that she isn't reckless. Ultimately, the case is under-described in ways that make it difficult to plausibly determine whether Alessandra's actions are justified.
[121] Amaya and Doris (2015) cite the notion of performance mistakes, but it's Brink (2013) who explicitly contrasts performance and competence in the context of the situationist challenge to moral responsibility.

she just fails to exercise this capacity in the moment, despite having a reasonable opportunity to do so. This distinction between performance and competence is key to how capacitarians understand blameworthiness in cases of ignorant wrongdoing.

It's important to note that Alessandra's blameworthiness is seemingly supported by anecdotal and experimental data. As Dana Nelkin and Samuel Rickless (2017) point out, most people in Alessandra's place will blame themselves, where the relevant self-reproach plausibly goes beyond the kind of 'agent regret' that might occur with other kinds of harm associated with one's actions.[122] Moreover, Samuel Murray et al. (2019) recently ran a series of behavioral studies suggesting that "we are disposed to hold others responsible for some of their forgetfulness" (p. 1177). These studies tested intuitions on cases very similar to Hot Dog, and Murray et al. concluded that the results provide some positive evidence in favor of capacitarianism. In this way, capacitarianism appears to support both first-personal and third-personal responsibility judgments. Of course, such data is far from conclusive, but insofar as we want a theory that can explain these responses, reasons-responsiveness views appear to have an extensional advantage over attributionism.

In fact, not only can capacitarianism plausibly capture intuitions in forgetting cases like Hot Dog, but they can also explain cases where the agent is *never* aware of the considerations that make her actions wrong. Consider, for instance, the following case from Randolph Clarke (2017):

> **Unaware Ann:** Ann is driving to a friend's house when she collides in an intersection with another car, killing one of its passengers. Ann has run a stop sign. She didn't see it. She wasn't intoxicated, and she wasn't speeding. But she hadn't driven this route before. And although she was watching the road, she was also thinking about her work; indeed,

[122] Agent regret, a term originally coined by Williams (1981), refers to "a sentiment whose 'constitutive thought' is a subject's first-person thought that it would have been much better had she done otherwise…[and] also requires a certain sort of expression that is different from that of what we might call 'bystander regret'" (Nelkin, 2019). Importantly, agent regret does *not* involve the thought that the subject is at fault or responsible.

she had just realized how to solve a problem that had been bothering her for days. (pp. 238-9)

Unlike Alessandra, who is subconsciously aware that Sheba is in the car as she languishes, Ann is never aware of the stop sign. In this way, Ann is more ignorant than Alessandra, and yet she's arguably just as, if not more blameworthy. Amaya and Doris (2015) distinguish these two kinds of cases in terms of *executive* and *cognitive* mistakes, where the former represent failures to react appropriately to reasons that one actually recognizes (e.g., Hot Dog), and the latter represent failures to recognize that certain moral considerations are germane to one's situation (e.g., Unaware Ann).[123] Crucially, capacitarianism can explain both kinds of cases in terms of unexercised capacities for awareness. Just as Alessandra could (and should) have remembered that Sheba was in the car, Ann could have noticed the stop sign. Indeed, capacitarianism can explain many cases of intuitively blameworthy failures to remember, notice, or realize morally relevant considerations.

2. Justification

According to capacitarianism, fully unwitting wrongdoers can be blameworthy if they could and should have known better. These wrongdoers could have known better if they possessed the capacities to recognize and respond to the relevant considerations that bear on their actions. This account answers basic explanatory questions about when and why ignorant wrongdoers are blameworthy, but it leaves unanswered further questions concerning the justification of some of these claims. Specifically, why should an agent's failure to exercise

[123] According to Fischer and Ravizza's (1998) reasons-responsiveness theory, these mistakes would presumably represent failures of *receptivity* and *reactivity*, although Fischer and Ravizza normally reserve these terms for referring to competence rather than performance.

certain capacities make her liable to blame and perhaps punishment? In order to fully evaluate the capacitarian approach, one needs answers to these questions.

Among reasons-responsiveness theories that embrace capacitarianism, I believe that there are two main justifications for grounding blameworthiness in unexercised capacities for awareness: a backward-looking and forward-looking justification. The backward-looking justification appeals to considerations of *fairness* or reasonable expectations, while the forward-looking one ultimately appeals to the *value* of a certain kind of self-control.[124] In the following section, I explain both types of justifications, citing specific accounts when applicable. Note that each justification is meant to be explanatorily fundamental or basic, so there will sometimes be intermediary explanations that help explain why unexercised capacities license blame. The point of this section, though, is to find the justificatory foundations of capacitarianism.[125]

2.1 I will start with the fairness justification. Although various accounts express the idea in different ways, this grounding of blameworthiness is essentially composed of two main claims: (1) an agent is blameworthy for wrongdoing insofar as she has a fair opportunity to avoid that wrongdoing,[126] and (2) an agent has such a fair opportunity in cases of ignorant wrongdoing insofar as she possesses the capacities to recognize and respond to the relevant considerations that bear on their actions.[127] The former claim explains blameworthiness in terms of the fair

[124] I'm assuming here that these two justifications are mutually exclusive as *ultimate* justifications (or grounds). It's worth noting, however, that the forward-looking justification includes backward-looking considerations, just not as the ultimate justification for blameworthiness. Both justifications are accounts of blame*worthiness* in the sense relevant to moral responsibility as accountability. There might be other considerations that bear on whether an agent ought to be blamed, but do not speak to an agent's responsibility. Such considerations would be neither necessary nor sufficient for blame but might plausibly override a *pro tanto* case for blame.

[125] Perhaps capacitarians might argue that this section mistakenly assumes a certain foundationalist justificatory structure, whereas the theory is actually supported through a kind of wide reflective equilibrium. My analysis is not meant to commit capacitarians to any specific justificatory structure, but only to delineate the fundamental principles that explain why unexercised capacities license blame.

[126] I use 'insofar' as shorthand for 'if and only if and insofar,' indicating that the consequent is a necessary and sufficient condition and has explanatory priority.

[127] Unless otherwise noted, I assume throughout the discussion that there are no situational factors that interfere with the fair exercise of these capacities. This stipulation allows me to focus on the *possession* of the relevant unexercised

opportunity to avoid wrongdoing, while the latter claim explains this fair opportunity in terms of the possession of certain capacities. In this way, we have a two-step grounding of blameworthiness in terms of capacities that is justified in terms of an agent's fair opportunity to avoid wrongdoing.

The first claim in this two-step grounding is represented in several reasons-responsiveness accounts, most obviously David Brink and Dana Nelkin's (2013) *fair opportunity* view. On this view, blame for wrongdoing is appropriate insofar as an agent is responsible, where responsibility is further explained in terms of capacities for reasons-responsiveness ("normative competence") and a fair opportunity to exercise these capacities ("situational control").[128] As Brink and Nelkin (2013) explain, "normative competence and situational control can and should be understood as expressing a common concern that blame and punishment presuppose that the agent had a fair opportunity to avoid wrongdoing" (p. 285). Thus, their account explicitly grounds blameworthiness in fair opportunity by way of a certain conception of responsibility. Like most other reasons-responsiveness accounts, the capacities relevant to fair opportunity are further decomposed into cognitive capacities to recognize wrongdoing and volitional capacities to respond accordingly. Presumably, cognitive capacities are most relevant to cases of ignorant wrongdoing, insofar as they involve abilities to form certain normative beliefs. However, volitional capacities might also be relevant if Amaya and Doris are right that cases like Hot Dog represent a kind of "executive mistake" involving a failure to react appropriately to the normative knowledge that one possesses.[129]

---

capacities. The situational element is less interesting insofar as there is general agreement that sufficient interference with the exercise of an agent's relevant capacities (e.g., duress and coercion) excuses blame.

[128] It's worth noting that Brink and Nelkin (2013) seem to prefer "a conception of normative competence and situational control in which they are potentially interacting, rather than independent dimensions of responsibility" (p. 18), though nothing in the proceeding discussion hangs on this conception.

[129] As I understand the fair opportunity view, normative competence is composed of specifically *normative* capacities. However, ignorant wrongdoing might also involve failures of *non-normative* capacities, such as the

By explicating the fair opportunity to avoid wrongdoing in terms of normative competence (and situational control), the fair opportunity view also represents the second claim in the two-step grounding of blameworthiness in capacities. Assuming that there are no situational factors sufficiently interfering with the exercise of an agent's capacities for reasons-responsiveness, the view entails that the possession of the relevant capacities is sufficient for blameworthiness by providing an agent with a fair opportunity to avoid wrongdoing. Looking back at cases like Hot Dog and Unaware, then, we can see how this fairness justification supports blaming agents. After all, it's usually stipulated that these cases involve situational factors that don't undermine the exercise of one's cognitive and volitional capacities. Therefore, as long as agents like Alessandra and Ann genuinely *possess* these capacities (which seems plausible), they are blameworthy because they had a fair opportunity to avoid wrongdoing.

Having explained the two main claims comprising the fairness justification – and having demonstrated how they might function within a reasons-responsiveness theory – I can now focus on how these claims are justified themselves. Again, Brink and Nelkin (2013) offer an instructive account. In support of the first claim, grounding blameworthiness in the fair opportunity to avoid wrongdoing, Brink and Nelkin point to an analogous view within criminal law theory.[130] This view specifically cites the demand for the fair opportunity to avoid wrongdoing as the rationale for refusing to punish agents in the absence of public notice of a legal requirement and in situations where they did everything within their power to obey.[131] In reference to punishing in the absence of public notice, Brink and Nelkin (2013) explain that such a practice would be

---

capacity to notice a stop sign in the case of Unaware Ann. Still, because these non-normative capacities support an agent's normative capacities, failures of non-normative capacities implicate their associated normative ones.

[130] See, e.g., Hart (1968b), Moore (1997), Morse (1994).

[131] The principle against punishing in the absence of a legal requirement is called *legality* in legal theory. Punishing in situations where agents did everything within their power to obey is referred to as *strict liability*. More precisely, strict liability offenses don't require culpability.

unfair "because it would punish those for failing to conform to behavioral expectations of which they had not been apprised in advance" (p. 307). Likewise, it's plausibly unfair to punish agents in situations where they did everything within their power to obey because it would be unreasonable to expect them to behave any differently. Switching back to moral responsibility, then, one can infer that Brink and Nelkin similarly believe that the reason that fair opportunity justifies blameworthiness is that it would be unreasonable to expect agents to conform to our expectations without such an opportunity. In this way, it's reasonable expectations that ultimately justify grounding blameworthiness in the fair opportunity to avoid wrongdoing.

Importantly, Jan Wieland (2017a) identifies a similar justification for the first claim within the context of a prominent argument that blameless ignorance excuses. The basic argument, which Wieland calls the "fairness explanation," runs as follows: [132]

> (P1) If an agent is (non-culpably) ignorant that she ought to avoid wrongdoing, then she can only avoid it based on luck or akrasia.[133]
>
> (P2) If an agent can only avoid wrongdoing based on luck or akrasia, then it would be unreasonable to expect her to avoid it.
>
> (P3) If the expectation to avoid wrongdoing would be unreasonable, then blame would be unfair, and so the agent isn't blameworthy for that wrongdoing.
>
> (C) Hence: If an agent is (non-culpably) ignorant that she ought to avoid wrongdoing, then she isn't blameworthy for that wrongdoing.

Setting aside the soundness of this argument, (P3) is particularly relevant to the current discussion. In this crucial premise, Wieland explicitly grounds blameworthiness in fairness and explains fairness in terms of reasonable expectations to avoid wrongdoing. Given that the

[132] Wieland claims to be drawing from Levy (2009) and Rosen (2003) in constructing this argument.
[133] The reasoning behind this premise is that if the agent hasn't considered that her actions might lead to wrongdoing, then she can only avoid it by luck; and if she has a false belief about whether her actions are wrong, then she can only avoid wrongdoing by acting akratically. Thus, the two major versions of ignorance – unconsidered belief and false belief – only allow the agent to avoid wrongdoing by luck or akrasia, respectively. As far as forgetting cases, I assume that they would be instances of avoidance by luck if this premise is meant to apply to them.

fairness explanation is a prominent argument explaining why blameless ignorance excuses, then, this reference offers further support for the view that reasonable expectations are the primary justification of the first claim in the fairness justification.[134] Indeed, the fairness explanation provides a citation of reasonable expectations specifically within the context of moral responsibility for ignorant wrongdoing.[135]

Reasonable expectations are also cited in the context of the second claim, which links fair opportunity to the relevant capacities for reasons-responsiveness. Consider, for instance, Clarke's (2017) capacitarian justification for blaming agents for ignorant wrongdoing:

> Given their possession of these capacities and abilities [for awareness], it was reasonable to expect them to have realized that their conduct was wrong, and they were able to avoid it. They then satisfy conditions that plausibly suffice for direct blameworthiness for wrongful conduct despite lacking awareness of its wrongness. (p. 240)

Here, Clarke agrees with Brink and Nelkin that possession of the relevant capacities is sufficient for blameworthiness by providing an agent with a fair opportunity to avoid wrongdoing.[136] Moreover, Clarke asserts that the *reason* why these capacities are sufficient is that it's reasonable to expect an agent to exercise those capacities to avoid wrongdoing. Therefore, absent interfering situational factors, only genuine *incapacity* can undermine fair opportunity. As with the first claim, then, reasonable expectations turn out to be the key consideration justifying the second claim.

---

[134] See also Rudy-Hiller (2020) for an argument that reasonable expectations ground blameworthiness via fair opportunity. According to Rudy Hiller (2020), "fair opportunities are…actually a component of reasonable expectations" (p. 2948).

[135] In particular, the fairness explanation claims that blameworthiness requires fair opportunity and thus reasonable expectations. In this way, fairness is a *necessary* condition for blameworthiness. However, I believe that most who make similar arguments linking blameworthiness and fairness also hold that fair opportunity is *sufficient* for blameworthiness (e.g., Brink & Nelkin, 2013), even if this isn't always explicitly argued for.

[136] Again, assuming that there are no situational factors that interfere with an agent's fair opportunity to exercise these capacities.

In discussing negligence, capacitarian Joseph Raz (2010) seemingly agrees with Clarke that reasonable expectations explain why capacities for reasons-responsiveness are sufficient for a fair opportunity to avoid wrongdoing. First, he argues that consideration of moral duties related to negligence suggests a "conception of responsibility as vested in conduct which results from the functioning, good or faulty, of our capacities of rational agency…" (p. 18). In this way, capacities for reasons-responsiveness are central to his account of moral responsibility. Within this view, Raz (2010) then emphasizes the importance of an agent's "domain of secure competence," which represents the range of possible actions determined by the abilities that an agent securely commands. As he explains, "actions due to malfunction of our capacities of rational agency result from failure to perform acts of which we are masters" (p. 17) – that is, failure to perform acts within our domain of secure competence. Moreover, Raz holds that agents can be held responsible for wrongdoing within this domain *because* of this competence. A plausible interpretation of Raz's view, then, is that an agent has a fair opportunity to avoid wrongdoing in situations where the relevant action falls within the scope of her capacities. Furthermore, we can blame agents for wrongdoing within this scope *because* it's reasonable to expect agents to exercise abilities that they have mastery over. In this way, Raz's justification mirrors Clarke's.

In summary, then, the fairness justification maintains that an agent's failure to exercise certain capacities that she possesses makes her blameworthy because these capacities are sufficient for a fair opportunity to avoid wrongdoing (assuming that the exercise of these capacities is not sufficiently interfered with). Moreover, this fair opportunity justifies blaming the agent because the availability of this opportunity makes it reasonable to expect her to avoid wrongdoing. As previously mentioned, different capacitarian accounts employing the fairness

justification might appeal to different intermediary norms and explanations (e.g., Raz's domain of secure competence), but the fundamental justification of blameworthiness is the same: it's fair to blame agents for ignorant wrongdoing if they possess the relevant capacities for awareness because it's reasonable to expect them to exercise these capacities.

2.2 The second justification, appealing to the value of a certain kind of self-control, also cites fairness and reasonable expectations, but not as the ultimate grounds for blameworthiness. Instead, the blaming *practices* that involve considerations of fairness are further justified by their value to self-control, in a two-tiered system.[137] The most prominent capacitarian account with this justificatory strategy comes from Manuel Vargas (2020) and his work with Samuel Murray (2017, 2020). Like Brink and Nelkin, Vargas's capacitarian account falls out of his broader reasons-responsiveness theory of responsibility. Although there are many interesting components of this theory, Vargas (2020) explains the teleological core as follows:

> When we hold responsible moral considerations-responsive agents (minimally, when we evaluate them in culpability-entailing ways) we participate in a system of practices, attitudes, and judgments that support and improve our responsiveness to moral considerations…Over time, and given psychologies roughly like ours, praise and blame and the related apparatus of responsibility practices performs an important function for us. That is, they sustain and further develop those moral considerations-responsive capacities that seem to naturally occur wherever groups of humans are to be found. (p. 406)

In other words, the function of blame is to help further develop the reasons-responsive capacities that undergird moral agency.[138] Because of this, the view is commonly referred to as the *agency cultivation model*.

---

[137] See Vargas (2022) for a discussion of general theories of moral responsibility that have this two-tiered structure. More broadly, theories that appeal to forward-looking considerations in order to ground blame are often labeled *instrumentalist* (or teleological). There are many different instrumentalist views (both one-tiered and two-tiered) that I could draw from, but I focus on Vargas's (2013, 2020) work for two main reasons: first, as far as I know he's the only capacitarian instrumentalist who focuses on epistemic issues; second, I find his instrumentalist account most plausible overall, and most promising for vindicating capacitarianism.

[138] See Fricker (2016) and McGeer (2013) for a similar (albeit one-tiered) instrumentalist theory that emphasizes the capacity-enhancing effects of blaming practices.

Crucially, this important function of blame doesn't justify blame *within* our practices; individual instances of blame are justified by the same backward-looking justifications as previously discussed. Instead, the entire practice is itself justified by its valuable contribution to agency cultivation. According to Vargas, such cultivation is in turn valuable because we have both individual and social interests in developing capacities for self-control. For example, Vargas (2020) explains that "being seen as incompetent at navigating moral considerations is, minimally, to be marked as untrustworthy in a range of social relations" (p. 407). Therefore, agents have an interest in developing and maintaining such competence to avoid this detrimental designation. Given that blame serves this valuable function, then, we're justified in the blaming practices that fulfill this function.

This teleological theory has the resources to explain why an agent is blameworthy for ignorant wrongdoing insofar as she possesses the relevant capacities for awareness. Specifically, Vargas (2020) argues that "one way we extend our capacities into new contexts is to, at some point, be vulnerable to blame because we had a capacity that went unexercised" (p. 410). Vargas doesn't elaborate much on this point, but one can plausibly fill in the details: being blamed for failing to exercise one's capacities in a certain context provides feedback that increases the likelihood that an agent exercises that capacity in similar contexts in the future. Blaming Alessandra for forgetting Sheba in the car, for example, makes her more likely to remember in the future. Again, the value of this kind of agency cultivation ultimately justifies the blame.

3. Two Problems for Capacitarianism

Overall, I believe that reasons-responsiveness capacitarianism has significant extensional and explanatory advantages over attributionism. Even setting aside cases of fully unwitting wrongdoing, capacitarianism can better capture and explain cases like Difficulty for Dennis and

Reckless Ralph. In these cases, our intuitions are much more plausibly tracking capacities than something like quality of will. Nevertheless, capacitarianism has its own challenges. In this section, I present two main problems for the view: the first relates to the central claim regarding *capacities* – namely, that agents genuinely possess the relevant capacities in cases of ignorant wrongdoing; the second relates to the normative claim that *justifies* blame based on these capacities. Thus, I present challenges to both components of the claim that agents could and should have known better in key cases of ignorant wrongdoing.

3.1 As previously mentioned, capacitarians tend to characterize cases like Hot Dog and Unaware Ann as instances of failures of performance rather than competence because they maintain that these agents possess the necessary capacities for awareness. However, evaluating such capacity claims is notoriously difficult. First of all, pre-theoretical intuitions appear to be divided. Although it seems natural to say that Alessandra had the ability to remember Sheba in the car, for example, Vargas admits that the first-personal phenomenology runs the other way; often, when we fail to notice or remember something, it doesn't *feel* like we had the ability in that moment. Because of this experience, Vargas (2020) suggests that "the capacitarian may need to dismiss phenomenology to save the normative metaphysics by holding that what matters is whether putatively negligent agents in fact have an unexercised capacity to recognize and respond to the relevant considerations" (p. 404). If Vargas is right, then we must look more closely at the normative metaphysics, getting clearer about the nature of these unexercised capacities.[139]

[139] This is not to deny that agents might sometimes feel guilty in cases like Hot Dog and Unaware Ann precisely because they feel that they did have the relevant ability in the moment. My agreement with Vargas is simply that the phenomenology is inconclusive.

One way to ground these capacities is in terms of *past performance*. For instance, Clarke (2017) argues that agents possess the relevant capacities insofar as they "routinely manifest" these capacities in similar situations. Because of this, Clarke claims that Ann has the capacity to notice the stop sign in the case of Unaware Ann because she routinely manifested this capacity in the past. However, even if Clarke can explain cases like Unaware Ann, this account of capacities is problematic, as it faces potential counterexamples from both directions of the analysis. First of all, as Nelkin and Rickless (2015b, 2017) point out, there are plausible cases of capacity possession where the agent failed to routinely manifest the capacity in the past. Imagine, for instance, that Ann routinely failed to notice stop signs because she's often texting while driving. Clarke's account of capacities seemingly implies that Ann doesn't have the capacity to notice stop signs, but intuition instead suggests that her capacity is simply being *masked* by another factor. In this way, lack of past performance doesn't necessarily indicate incompetence.

Conversely, routinely manifesting capacities in the past doesn't necessarily mean that an agent currently possesses those capacities. Suppose that Ann routinely noticed stop signs in the past, but recently had an operation affecting her recognitional abilities. In determining her current capacity to notice stop signs, it seems obvious that the operation is the relevant consideration – not her past performance – and so we should consider her incompetent. Of course, Clarke might reasonably argue that this operation changes the current situation sufficiently such that her past performance is no longer the relevant comparison class. But then how do we determine capacity in instances without relevant past performance? In cases like these, the past performance account of capacities gives the wrong answer or no answer.[140]

[140] This is not to say that past performance might not be relevant to assessing capacity in certain instances, only that past performance can't be a necessary or sufficient condition for capacity. Past performance might play an *evidentiary* rather than constitutive role in a plausible theory of capacities that involves a multimodal conception of evidence. On such a multimodal account, certain sources of evidence for capacity may be masked or otherwise

A better theory of capacities abandons past performance for *counterfactual* performance.[141] Of course, different counterfactual accounts differ in the details, but the basic approach is the same: to determine whether an agent has the capacity in a certain situation, we evaluate her performance in relevantly similar counterfactual scenarios.[142] Such counterfactual views often render this evaluation in possible worlds language. Thus, to determine capacity in a situation, we look to relevantly similar possible worlds; if the agent manifests this capacity in enough of these worlds, then the agent possesses the capacity. For example, to assess whether Alessandra has the capacity for awareness in Hot Dog, we determine whether she remembers Sheba in a suitable proportion of relevantly similar possible worlds.[143]

Still, although the counterfactual approach is better than the past performance account, this conception of capacities is not without its own problems. A thorough discussion of all the issues surrounding capacities (or abilities) is far beyond the scope of this chapter, but I will focus on a central challenge. Carolina Sartorio (2017) refers to this challenge as the *demarcation problem* and describes it below in reference to reasons-responsive capacities:

> …we need some principled reason to single out the aspects of the actual circumstances that we can vary from the aspects of the circumstances that we must held fixed in order to assess an agent's reasons-responsiveness on a certain occasion. Once we acknowledge that *not all* possible worlds where the agents have sufficient reasons to refrain from acting are relevant to their reasons-responsiveness in the actual scenario, we need to say more about which ones are relevant and which ones aren't. (p. 5)

defeasible in certain situations, yet capacity can still be plausibly determined when there's evidence regarding this masking.

[141] See, e.g., Brink (2021), Brink and Nelkin (2013), Fischer and Ravizza (1998), and Vargas (2013).

[142] In particular, there is disagreement regarding the relation between these counterfactuals and the capacities or abilities they explain. For discussion, see McKenna (2022).

[143] According to so-called *abilities-first* views, such counterfactuals are merely evidence of ability/capacity, rather than constitutive of it. This non-reductive approach is meant to address alleged problems with metaphysically grounding abilities in counterfactuals. However, nothing in the following discussion hangs on this dispute.

In other words, we need some principled basis for demarcating the *relevantly similar* possible worlds that ground our capacity judgments.[144] Any counterfactual account obviously appeals to worlds that differ from the actual world, but how exactly do these worlds differ? Without a convincing answer, it seems that we're back to brute and sometimes contradictory intuitions about capacities.[145]

Admittedly, it would be overdemanding to expect the capacitarian to provide a decision procedure for every possible instance of capacity. Nevertheless, it seems fair to ask for reasonably clear demarcation principles for important hard cases. Consider, for instance, the relevantly similar possible worlds that apply to Hot Dog. In assessing whether Alessandra has the pertinent capacity for awareness, one wants to exclude worlds where the conditions don't sufficiently bear on her actual circumstances. For example, imagine a possible world in which her children ask about the location of Sheba, thus jogging her memory. The fact that Alessandra remembers Sheba in this possible world clearly doesn't seem relevant to her capacity in the actual situation. But what about worlds lacking the distracting condition, should we include those? I argue that we shouldn't.[146] After all, this condition appears central to the situation we want to assess for capacity – unlike, for instance, where Alessandra parked, the exact

[144] We also need some account of what constitutes a "suitable proportion" of possible worlds, but I take it that the demarcation issue is more pressing (or at least prior). It's worth noting that the general structure of the demarcation problem applies to counterfactual theories more broadly, such as counterfactual accounts of causation and explanation. For current purposes, I focus on the problem as it applies to capacities. However, it might be worth considering whether capacitarianism can draw on resources from these similar issues in other contexts.

[145] One way to understand this challenge is as a burden of proof argument. Specifically, capacitarian accounts bear the burden of proving that there can be a principled basis for demarcating the relevantly similar possible worlds that explain the relevant capacity judgments. Without discharging this obligation, it's unclear why someone not already committed to capacitarianism should assume that the challenge can be met.

[146] In assessing capacity for awareness, then, we're assessing a fairly *specific* capacity. Presumably Alessandra possesses a general capacity for awareness in that she can remember (or maintain awareness) that Sheba is in the car in many other situations. However, the question is whether she has the capacity in the *current* situation, which demands assessing worlds that are mostly similar to the actual one. See Nelkin and Rickless (2017) for an argument in the context of awareness that "what matters for moral responsibility, surely, is specific ability, rather than general ability" (p. 127).

temperature, which child misbehaved etc. As Sher originally presents the case, he simply stipulates that this distracting condition doesn't undermine capacity; but this will ultimately depend on which possible worlds one assesses, and where one sets the threshold for incapacity. In this way, capacitarians are free to construct cases of genuine performance errors, but they owe a fuller account of how they're evaluating competence.

One notable exception to the general lack of guidance for specifying unexercised capacities is Vargas's (2013; 2020) previously discussed agent cultivation model.[147] According to Vargas's (2013) account, the relevantly similar possible worlds are those given by the standards:

> …an ideal, fully informed, rational, observer in the actual world would select as at least co-optimal for the cultivation of our moral considerations-responsive agency, holding fixed a range of general facts about our current customary psychologies, the cultural and social circumstances of our agency, our interest in resisting counterfactuals we regard as evaluatively irrelevant, and given the existence of genuine moral considerations, and the need of agents to internalize norms of action for moral considerations at a level of granularity that is useful in ordinary deliberative and practical circumstances. (p. 214)[148]

A crucial feature of this account is Vargas's (2020) assertion that "the nature of a capacity is interest-sensitive" (p. 408), and he provides details regarding the relevant interests in the context of moral responsibility. In this way, the agent cultivation model provides a more principled basis for demarcating the relevantly similar possible worlds that ground our capacity judgments.

Despite its advantages, though, Vargas (2013) acknowledges that his account "is more of a recipe for a substantive conclusion than a bold, decisive answer" (p. 222). His response to this limitation is that this generality simply reflects the complexity of our system of responsibility

[147] See also the work of the so-called *new dispositionalists*; e.g., Fara (2008), Smith (2003), and Vihvelin (2004). Overall, I find Vargas's account to be more detailed on the issue of relevantly similar possible worlds, so I use his theory as the current best version.

[148] In the interest of space, I'm simplifying some of the details of this rich account. For instance, Vargas (2013, pp. 214-5) provides an ordering of preferences that structures the ideal observer's choice of possible worlds. As far as I can tell, nothing in the following discussion hangs on these details.

practices. Still, I'm not sure that even this sophisticated theory provides clear enough demarcation principles for important hard cases. It seems possible to argue from this account of capacities to the conclusion that agents like Alessandra and Ann *don't* possess the necessary capacities for awareness, for instance. Indeed, I will make this argument later in the chapter. Yet, if there can be reasonable disagreement about such cases, then capacitarians aren't entitled to the claim that these agents possess the necessary capacities for awareness.

Furthermore, although the agent cultivation model provides a more thorough explanation of unexercised capacities than its counterparts, it does so with commitments that present their own issues. The crucial idea that capacities are interest-sensitive, which helps delimit the relevantly similar possible worlds, certainly has appeal, but it's unclear that it can be separated from a broader teleological conception of moral responsibility. After all, why think that capacities are interest-sensitive unless these interests play some significant role in our theory of moral responsibility? But then the plausibility of the account of capacities relies on the underlying teleological conception of responsibility. Of course, Vargas openly welcomes this result, as his conception of capacities derives from his teleological theory, but those who don't share this conception of responsibility will consider it a drawback that subscribing to the account of capacities requires defending this kind of theory. It's worth considering whether other possible attempts to explain unexercised capacities will introduce similar issues.

3.2 Even granting that agents possess the necessary capacities in relevant cases, though, capacitarianism faces a more fundamental challenge. Specifically, it's dubious that mere possession of such capacities is sufficient for blameworthiness. In other words, even if it's true that an agent *could* have known better, it's not necessarily true that she *should* have known better. In section 2, I outlined two justifications for the claim that an agent's failure to exercise

certain capacities makes her blameworthy. I now raise issues for both approaches, focusing primarily on the fairness justification. Recall that in explaining the fairness justification, I introduced two main claims: (1) an agent is blameworthy for wrongdoing insofar as she has a fair opportunity to avoid that wrongdoing, and (2) an agent has such a fair opportunity in cases of ignorant wrongdoing insofar as she possesses the capacities to recognize and respond to the relevant considerations that bear on her actions. Like most reasons-responsiveness theorists, I'll assume that some version of (1) is correct, especially insofar as fair opportunity is central to reasonable expectations. Instead, I'll focus on the second claim, linking fair opportunity to capacities. After all, this is the claim that is uniquely capacitarian.

The essential problem with the second claim is that mere possession of capacities is insufficient for a fair opportunity (or reasonable expectations). Here, I join Nelkin and Rickless (2017) and Fernando Rudy-Hiller (2019) in arguing that *awareness of risk* is also necessary for fair opportunity; that is, not only does an agent need the relevant capacities to avoid wrongdoing, but she also needs some awareness that her action (or omission) risks wrongdoing.[149] This awareness needn't occur at the precise moment of wrongdoing, but it must at least occur at some suitable prior point. Without any awareness of risk, the agent doesn't have a fair opportunity to avoid wrongdoing.[150]

In order to better illustrate this commitment, return to Hot Dog. While Sheba languishes in the car, Alessandra is unaware that her omission risks harm. Because of this, I maintain that she doesn't have a fair opportunity to avoid wrongdoing during this period. Still, is there some

[149] Rudy-Hiller (2019) also adds a *know-how condition*: "the agent must know how to avoid the risk in question, that is, what to do in order to achieve the desired cognitive state" (p. 724-5). I remain agnostic on the necessity of this kind of condition.

[150] The task of the next chapter will be filling out an account of awareness of risk. For now, I simply use the condition to contrast it with capacitarianism, which denies the necessity of any conception of awareness.

previous time when Alessandra fulfills the awareness of risk condition? The answer depends on how exactly one renders the condition, and how the case is understood.[151] Sher seemingly implies that Alessandra was initially aware that it was a very hot day, and that Sheba would suffer from being left in the car for too long. Thus, one might conclude that she satisfies the awareness of risk condition at that earlier moment. In this version of the case, she should have increased her vigilance or set some sort of reminder if she was going to leave Sheba in the car. Otherwise, she's blameworthy for leaving Sheba to languish.[152] Yet one could easily describe the case so that there's no such prior awareness. Suppose that it wasn't an especially hot day, so it never crossed Alessandra's mind that Sheba could be in danger inside the car.[153] In this version of the case, I contend that Alessandra is plausibly excused for failing to let Sheba out. This contradicts capacitarianism, insofar as Alessandra nevertheless possessed the capacity to become aware that she left Sheba in the car. After all, the relevant facts were sufficiently available to her, and she had the capacity to recognize them. My explanation for the intuition that Alessandra is excused under these conditions is that she fails to satisfy the awareness of risk condition.

The case of Unaware Ann is perhaps more straightforward. At the moment that she runs through the stop sign, she's oblivious of its existence. In this way, she doesn't forget morally relevant information, she never even processes it in the first place. Now, the way the case is originally described, it seems like her preoccupation with her work played some role in her lack of attentiveness. One might argue, then, that she's blameworthy for the accident because she allowed her mind to wander. But this is an insufficient basis for blame. First, an agent's mind can

[151] Among other things, it depends on how one characterizes the necessary awareness, how one characterizes the relevant risk, and how one balances this risk with the reasons for taking it.
[152] See Murray (2017) and Murray and Vargas (2020) for an account of a distinct vigilance capacity.
[153] I assume here that it was still *actually* dangerous to leave Sheba for any extended period, even though this fact wasn't as obvious as it would be on an especially hot day.

wander without her being *aware* that it's wandering. In this case, I contend that the agent lacks the requisite awareness to be held derivatively responsible for her mental preoccupation. Second, even if Ann *were* aware that her mind was wandering, this still isn't enough to necessarily satisfy the awareness of risk condition. Instead, Ann would have to also realize (in some sense) that thinking about this problem might dangerously affect her attention, and there's no indication that this occurs.[154] Because of this, I contend that Ann also lacks a fair opportunity to avoid wrongdoing and is thus excused.

It's important to note that the standard excuses aren't available to Alessandra or Ann.[155] Even though Alessandra's school drama may have distracted her, it's plausible that this situational factor didn't rise to the level of genuine *duress* or any other standard excuse. At least, this distraction wouldn't wholly excuse Alessandra on most theories of moral responsibility, even if it might somewhat mitigate blame. As Amaya and Doris (2015) point out, though, it's commonly held that without such an excuse an agent must be blameworthy. In this way, blameworthiness and excuse are *inverse* concepts.[156] What does this mean for the current view, which demands awareness of risk for blame? I could simply deny that the two concepts are so inversely related and maintain that agents can fail to be blameworthy for wrongdoing despite lacking an excuse. However, I think a more plausible response is to simply argue for an enlargement of the class of standard excuses; specifically, agents are excused when they lack the relevant awareness of risk. Regardless of how we categorize the awareness of risk condition,

[154] I will say much more about this sense of realization in the next chapter.
[155] I borrow the language of 'standard excuses' primarily from Amaya and Doris (2015). Among the standard excuses are insanity, duress, and coercion.
[156] As Moore (1997) says, "excuse is the royal road to responsibility" (p. 548). Brink (2021) adds that "we do well to remember that this is a two-way street" (p. 17).

though, the important point is that capacitarianism fails to recognize that this condition is necessary for a fair opportunity.

Capacitarians would argue that their failure to recognize an awareness of risk condition really amounts to a *rejection* of this additional condition. After all, a defining feature of capacitarianism is the commitment to the possibility of blameworthiness for fully unwitting wrongdoing. Thus, capacitarians would simply deny that fair opportunity requires awareness of risk. Besides pounding the table, though, how might capacitarianism bolster its claim that possession of capacities is sufficient? One potentially promising strategy is to appeal to the forward-looking considerations discussed in section 2.2. Recall that on Vargas's two-tiered account, the individual and social interests in developing capacities for self-control ultimately justify blame. Capacitarians could argue that the value of this self-control tips the balance in favor of blaming in the absence of awareness of risk.

Importantly, appealing to the value of a certain kind of self-control doesn't require capacitarianism to abandon a desert-entailing notion of blameworthiness. This is a key feature of the two-tiered approach. As Vargas (2020) explains, "the teleological element…is a feature of *the system* of first order norms. Desert judgments are judgments at the level of the first order norms and not in conflict with the second-order teleological character of the account" (p. 406). Because of this, the first order justification for blaming agents who possess capacities for awareness can still reference norms grounded in fair opportunity (or reasonable expectations), despite the second order justification of these norms/practices being their role in developing and sustaining a valuable form of self-control.[157] Within this framework, the capacitarian could argue

[157] It's important to point out that Vargas supports *revisionism* when it comes to folk theorizing about moral responsibility, including theorizing about desert. See, e.g., Vargas (2013, 2015, 2022, forthcoming). Thus, although he intends for this account to vindicate most of our ordinary blaming *practices*, he advocates revision at the level of theories of responsibility (and free will).

that our *actual* norms and practices contravene a demand for awareness of risk, and that these practices are further justified by their valuable function. If successful, this argument would shift the burden to the proposed view to explain why we should excuse agents in the absence of awareness of risk, given that our practices support blaming and these practices are valuable.

In response, I concede that I can't offer decisive theoretical considerations favoring an awareness of risk condition to those who don't share the basic underlying intuition. However, I present two main challenges for the previous capacitarian argument. First, appealing to forward-looking considerations doesn't seem to support capacitarianism unless one commits to a teleological theory of moral responsibility that introduces other issues. This is because these considerations are the wrong kind of reasons to justify blame directly.[158] They might be the right reasons to justify the *system* of blaming practices, but this requires committing to our practices being justified in this way. Because of this, capacitarians can't plausibly avail themselves of these forward-looking considerations without committing to a certain theory of moral responsibility. Of course, Vargas would welcome this result, but most other capacitarians reject a teleological theory of responsibility.

Among the reasons for being wary of a teleological theory is that two-tiered accounts don't allow for a certain kind of *internal* criticism that seems intelligible.[159] Specifically, they can't validate critiques of our actual norms and practices that don't ultimately appeal to their role in developing and sustaining a valuable form of self-control. Suppose that it's true that our actual norms and practices contravene a demand for awareness of risk. Still, I might argue that this is a

[158] Vargas (2022) argues that this "wrong kind of reasons" objection against teleological theories doesn't apply to two-tiered theories because blame isn't directly justified by forward-looking considerations on such views. Yet, this structural feature that allows Vargas to potentially evade the objection also explains why capacitarianism can't avail themselves of this strategy without committing to a teleological theory of responsibility. Ultimately, forward-looking considerations are the wrong kind of reasons to justify blame *directly*.

[159] See Vargas (2022) for a survey of major objections to teleological (or instrumentalist) accounts of moral responsibility, some of which his two-tiered account seems poised to address.

mistake by appealing to a certain conception of fair opportunity rather than the disvalue of these norms/practices. How can the teleological view make sense of my criticism? As long as the actual norms and practices promote agency cultivation, then even the two-tiered account must interpret my argument as fundamentally confused.

It's instructive that Vargas (2015; 2022) sometimes compares the normative structure of his teleological theory of moral responsibility to the rules of sports:

> Foul calls in a sport are typically justified by the thought that the safety of the players must be preserved, but this must be balanced with the enjoyment of spectators and players in the flow of the game. However, whether a particular play is a foul or not is clearly not settled by appeal to those framework questions. They are settled internal to the framework, by appeal to the rules of the game. (2022, p. 17)

In the context of sports, then, it doesn't make sense to challenge foul calls that correctly apply the rules of the game. As Vargas (2022) explains, "the rules are the rules, at least until we change them" (p. 18). However, in the context of moral responsibility it *does* make sense to challenge instances of blame even when this blaming conforms with the rules of our blaming practices, and this criticism needn't appeal to external values regarding the entire responsibility practice but can instead reference normative considerations *within* our practice. Moreover, these internal considerations should trump any external ones appealing to individual and social interests. After all, why should we continue to blame agents for wrongdoing that our best normative theory tells us is unfair just because doing so would be somehow valuable? Even if blaming ignorant wrongdoers who possessed the relevant capacities for awareness would help develop and sustain a valuable form of self-control, this doesn't seem to justify blame in the absence of a genuine fair opportunity. Therefore, the fact that our actual practices blame certain agents, and these practices serve a valuable function, is insufficient to justify them if we have strong moral reasons to

refrain from blame. I maintain that we have strong reasons to refrain from blame in the absence of awareness of risk.

Second, even conceding the teleological justification of our norms and practices, it's unclear that blaming agents in the absence of awareness of risk actually develops and sustains a valuable form of self-control. It's ultimately an empirical question whether such blame has this effect, but there are reasons to think that it doesn't. Suppose, for example, that we blame Ann for failing to notice the stop sign, despite possessing the relevant capacities. How could she use this feedback to develop her self-control? If her mind often wanders while driving, then perhaps she could take measures to maintain focus. But imagine it was a one-off slip. Should she now remain hyper-vigilant whenever she drives? Should she carry over this hyper-vigilance to any situation where a lack of awareness is potentially risky? It's plausible that the potential value of such hyper-vigilance would be outweighed by the costs of maintaining this vigilance. More importantly, though, how can one make informed decisions about when to exercise such vigilance without some awareness of the relevant risks? In this case, awareness of risk still plays a central role in our practices.

In response, Vargas might argue that his own teleological theory emphasizes the valuable *systemic* effects of blaming practices on groups of people. Thus, even if blaming Ann might not develop her individual self-control, having the *rule* of blaming agents for similar slips could still serve this valuable function in the aggregate. Yet even though this kind of indirect account has advantages, I still question the mechanism by which a rule of blaming agents for unexercised capacities of awareness develops and sustains self-control. To illustrate this point, consider a case that includes awareness of risk: Reckless Ralph. Blaming reckless agents like Ralph plausibly develops reasons-responsiveness. Because he *wittingly* risked wrongdoing by choosing

to rely on his memory, he can now actively avoid taking such risks in the future. He can recalibrate his balancing of reasons so that he puts more weight on other people's wishes. All this development is possible because his mistake can be traced to some sort of witting action or omission. Without such awareness, however, it's more difficult to see how agents could use negative feedback to help develop their self-control. If the relevant mistake is fully unwitting (and exceptional), then how do you plan to avoid it in the future? The only realistic answer seems to involve a kind of broad vigilance that comes with significant costs.

4. Capacitarian Response

In section 3, I presented challenges to both components of the core capacitarian thesis that agents could and should have known better in key cases of ignorant wrongdoing. Capacitarians will (and do) have responses to both challenges. On the capacities front, I've already mentioned that Vargas's (2013; 2020) rather sophisticated account offers a promising formula for addressing the demarcation problem. Moreover, there are arguments in both moral and legal philosophy refuting the claim that fair opportunity requires awareness of risk.[160] I find these arguments unconvincing, but I must admit that my position is ultimately grounded in rather fundamental normative intuitions – I don't see decisive theoretical considerations on either side. Perhaps, though, I just have faulty intuitions. If it could be established that my judgments regarding cases was non-standard, then this would seemingly undermine my position. In fact, if capacitarianism has one major advantage over its (non-attributionist) rivals, it's precisely this apparent intuitive superiority. Capacitarianism seems to better capture common judgments in relevant cases of ignorant wrongdoing. In this section, I present and evaluate this apparent

[160] See, e.g., Brink (2021) and Hart (1968b).

advantage. Ultimately, I argue that there are reasons to put less weight on the significance of these contradictory responses.

4.1 Focusing on key cases of ignorant wrongdoing where the agent lacks awareness of risk but nevertheless potentially possesses the relevant capacities for awareness (e.g., Hot Dog and Unaware Ann), there's widespread acknowledgement that most people are disposed to blame the agent. Indeed, even those critical of such blame generally admit that dispositions run contrary to their position. Within legal theory, Larry Alexander (2014) and Michael Moore and Heidi Hurd (2011) – who are all critical of punishing for negligence – concede that our retributive reactive attitudes conflict with their view.[161] As Alexander acknowledges, these reactive attitudes persist even when we're convinced that the ignorant wrongdoer had no ill will, and we generally expect the wrongdoer to feel guilty and not just regretful.

This anecdotal evidence is seemingly supported by more rigorous data. As mentioned earlier, Murray et. al (2019) conducted a behavioral study to evaluate "whether and to what extent we judge that people are responsible for the consequences of their forgetfulness" (p. 1177). To do so, they asked participants to read one of four vignettes in which Randy's wife calls to ask him to pick up some ingredients from the store on his way home from work and he ends up forgetting the request.[162] These vignettes varied along two dimensions: (1) the level of care that the agent displays toward performing the planned action, and (2) the level of stress that the agent is under.[163] Participants randomly received one variant, and their reactions were judged by

[161] Importantly, the relevant notion of negligence at play in these discussions might not directly correspond to the concept of ignorant wrongdoing that is the current focus. Indeed, Moore and Hurd (2011) are skeptical that there's even a coherent notion of negligence within the criminal law. Nevertheless, I think it's fair to say that they acknowledge that we are prone to blame agents like Alessandra and Ann.

[162] In order to show that their results were not an artifact of the particular vignettes they used, Murray et al. (2019) ran another study with different vignettes. This study replicated most of the effects of the previous studies, and the differences aren't relevant enough to necessitate discussion of these different vignettes.

[163] Part of the impetus for the study was to provide results that might support one of two competing theories of moral responsibility – *valuationist* (attributionist) or *capacitarian*. Thus, varying the vignettes along dimensions seemingly

questioning them according to the outcome variables of fault, blame, and guilt. The results indicate that "we are disposed to hold people responsible for (some of) their forgetfulness" (p.1196). In fact, Murray et al. found that although stress acted as a mitigating factor on participant's judgments, even high stress situations didn't totally diminish some of the consequences of forgetting.

One must be cautious drawing strong conclusions from a single study, but the Murray et al. results at least provide initial support for capacitarianism. Although they only focused on variants of a single forgetting case, they explicitly modeled this case from similar examples in the philosophical literature. In doing so, they carefully screened off potentially confounding factors that might exist in many real-world situations. It would be interesting to see the results of a similar study focusing on a case like Unaware Ann, or other instances of performance errors without awareness of risk. As previously mentioned, one difference between forgetting cases and other instances of ignorant wrongdoing is that forgetful agents have formerly recognized the relevant considerations. In the study, for instance, Randy recognized that he ought to pick up the ingredients that his wife requests before eventually forgetting. How would participants react to a case where Randy fails to pick up the ingredients because he simply doesn't notice the grocery store as he passes it?

In assessing the import of the study, it's also worth noting that the consequences of forgetting in these vignettes were rather minor. Unlike Alessandra and Ann, Randy's omission doesn't result in any material harm, yet participants were still willing to blame him and thought he should feel guilty. How might participants react to a case with higher stakes? Plausibly, people would be even more likely to blame in situations where the consequences are more

relevant to each theory – care for valuationist theories and stress for capacitarian theories – makes sense as a method for comparing the two.

significant, further supporting the conclusions of the study. However, Murray et al. never actually tested this dimension of blame.

4.2 Despite this evidence, however, there are reasons to put less weight on the significance of these responses. First, there are well-studied cognitive biases that significantly complicate the conclusions one can plausibly take from these reactions. Especially pertinent to the current discussion is what Royzman et al. (2003) call the *I know, you know* bias, which involves overestimating others' epistemic state based on your own knowledge.[164] In applying this bias to cases of resultant luck, Royzman and Kumar (2004) suggest that the tendency to blame unlucky agents more than lucky ones is a product of projecting our own (*ex post*) knowledge onto the unlucky agent (*ex ante*). That is, after learning that an agent's actions resulted in harm, we assume that she was aware of the risk of harm before acting. In the context of a case like Hot Dog, then, it's possible that blaming attitudes toward Alessandra are subject to this bias. After learning that Sheba ends up languishing in the car, there might be a tendency to assume (perhaps subconsciously) that Alessandra was aware of this risk.[165] After all, the way forgetting cases are presented, it's often left implicit that the agent was unaware of any risk of forgetting. This might allow the I know, you know bias to more easily corrupt moral judgments. Moreover, this ambiguity is presumably even more common in everyday moral practice, where we usually have

[164] In support, Nichols et al. (2014) claim that "there is a wealth of independent evidence that we are indeed subject to an egocentric bias in judging others' epistemic situations" (p. 164).

[165] Against this interpretation, Markus Kneer and Edouard Machery (2019) recently ran a series of studies suggesting that "in non-comparative situations, outcome affects blame and wrongness judgments because people view the unlucky agent as more negligent than the lucky agent, and they view the former as more negligent because they consider the probability of the bad outcome's occurrence in the unlucky situation as higher than in the lucky situation" (p. 16). In other words, people tend to judge unlucky agents as more blameworthy because they increase the probability of the outcome in response to its occurrence, not because they ascribe recklessness to the agent. This kind of performance error is referred to as *hindsight bias* (see, e.g., Fischhoff, 1975). Although hindsight bias offers a competing explanation of responses to the relevant cases, I don't believe there's currently conclusive evidence in favor of any explanation. However, even if hindsight bias is the error behind these responses, this still undermines the weight of these responses in supporting capacitarianism; the difference is that hindsight bias doesn't also lend support for an awareness of risk condition.

less epistemic access to other agents, and thus ordinary blaming practices might be even more prone to distortion.

In cases like Unaware Ann, where she simply fails to notice the stop sign, it's more explicit that she isn't aware of any morally relevant risks.[166] Under these circumstances, perhaps it's less likely that the I know, you know bias acts on moral judgments. Still, there are other established biases that might be active. For instance, there is a wealth of social psychology research on the *fundamental attribution error*, the tendency to overemphasize personal characteristics and underemphasize situational factors when explaining others' behavior.[167] As Gilbert Harman (1999) influentially argues, there is good reason to think that *moral* judgments are subject to this error. If this is true, then this bias could go some ways toward explaining blaming attitudes directed at agents like Ann. When we hear that Ann drives into the intersection without noticing the stop sign, we might assume that her inattention derives from a lack of concern for her fellow drivers. Indeed, Matthew Talbert (2017b) makes a similar argument from the fundamental attribution error to support his attributionist theory, claiming, "it's at least possible that we have a tendency to misattribute blame-grounding attitudes in cases of harmful unwitting omissions, particularly when there is no evidence for the absence of these attitudes" (p. 32). Although I disagree that such cases ultimately vindicate attributionism, I can agree that our reactions might plausibly reflect a disposition to project certain attitudes onto agents. [168] Again, such projection is probably even more common in everyday moral practice, where we usually have less access to other agents' attitudes.

---

[166] Although it's explicit that she isn't aware of the risk of running the stop sign, it's *not* explicit that Ann isn't aware that allowing her mind to wander might be risky. Because of this, the I know, you know bias might be influencing responses by generating the assumption that Ann was aware of this risk. I think this bias is less likely in this case, so I consider other possible explanations.

[167] 'Fundamental attribution error' was coined by Ross (1977).

[168] See Brink (2013) for a broader argument that situationist findings like the fundamental attribution error don't compromise the attribution of unexercised capacities to avoid wrongdoing.

What about the tendency for people to feel guilty in situations like Hot Dog and Unaware Ann? Although Murray et al. never asked participants whether *they* would feel guilty in Randy's position, participants did answer that Randy should feel guilty for forgetting. Furthermore, anecdotal evidence suggests that agents like Alessandra and Ann usually blame themselves, even while recognizing that they weren't aware of any relevant risks.[169] Indeed, they often blame themselves even when others aren't prone to blame them. Some, like Michael Moore (2009), who privilege first-personal judgments of responsibility over third-personal judgments, infer from these responses that "the reason we *feel* so guilty in such cases is because we *are* so guilty" (p. 30).[170] In other words, by appropriately focusing on self-blame in relevant situations, we ought to conclude that these agents really are blameworthy.

However, these guilt reactions might instead reflect an asymmetry in the *ethics* of blame. Consider Nelkin (2022), who recently argues for what she calls the *Blame Asymmetry Claim*:

> In a wide variety of cases, it is appropriate for an agent, A, to blame herself to a certain degree, n, at the same time that it would be appropriate for others to blame her to a degree less than n, and there is a systematic explanation for this fact. (p. 98)

Setting aside the systematic explanation at work here,[171] it's rather intuitive that it can sometime be fitting for a person to blame herself more than others should blame her. For instance, it seems appropriate for an agent to blame herself to a certain high degree for hurting a friend's feelings in a way that would be inappropriate for some third party (to borrow an example from Nelkin). This asymmetry might help explain why agents like Alessandra and Ann often blame themselves even when others aren't prone to, but it also highlights that blame responses are often influenced by

[169] I use 'guilt' and 'self-blame' interchangeably here, although some prefer to differentiate the two concepts. For example, McKenna (2022) reserves 'guilt' to refer to the *expression* (or feeling) of self-blame.
[170] Moore (2009) is specifically referring to guilt responses in the context of (resultant) *moral luck*.
[171] For Nelkin (2022), the systematic explanation is grounded in the following principle: "All other things equal, it is harder to justify risking harm to others than to ourselves" (p. 111).

norms that don't directly bear on the blame*worthiness* of the action. Even if it might be appropriate for agents like Alessandra and Ann to blame themselves to some degree, then, they still might not be blame*worthy* in the (desert-entailing, accountability) sense that is distinct from the appropriateness of any individual's blame. Within this framework, one can validate self-blame and related intuitions in situations like Hot Dog and Unaware Ann without conceding that these agents are actually blameworthy in the relevant sense.[172] At the very least, Nelkin (2022) is surely right that "knowledge of a systematic difference in the appropriateness of self- and other-blame…is important when assessing different methodological claims like which blame intuitions to privilege in out theorizing" (p. 116). Because of this, Moore and others are too hasty in deriving blameworthiness from guilt responses, even if these responses are appropriate in a certain sense.

Finally, as Rudy-Hiller (2019) suggests, it's quite possible that intuitions regarding relevant cases of ignorant wrongdoing "are fueled partly by non-desert-entailing varieties of blame" (p. 737). Rudy-Hiller lists five varieties – causal, compensational, role-related restorative, and formational – which can be fitting responses even when desert-entailing accountability blame isn't.[173] Indeed, although Murray et al. attempt to obviate this problem by measuring judgments of fault, blame, and guilt, they admit that "the term 'responsibility' is ambiguous," and that "it is unclear that people use the term 'moral responsibility' in ordinary conversational contexts, and so unclear whether there is a folk concept of moral responsibility to probe or even whether people have a coherent concept of moral responsibility" (pp. 1183-4). If this is right, though, it's hard to see how shifting to the concepts of fault, blame, and guilt will

[172] Another possibility is that they aren't blameworthy to the same *degree*. Although they might be slightly blameworthy, they are much less blameworthy than they actually blame themselves.
[173] Role-related responsibility might be particularly applicable to cases common in the literature that involve paternal, marital, and caretaker relations.

solve the problem, since these concepts presumably inherit at least some of the ambiguity from the notion of responsibility that they're based on. Thus, there is further reason to put less weight on the significance of these responses.

Conclusion

In this chapter I presented capacitarianism as a promising account of moral responsibility for epistemic considerations within the context of reasons-responsiveness theories. However, although I noted that capacitarianism has significant advantages over attributionism in capturing and explaining the influence of these considerations, I argued in section 3 that the approach is ultimately flawed. In claiming that agents possess the necessary capacities in cases of ignorant wrongdoing, capacitarianism faces the demarcation problem; that is, it needs some principled basis for demarcating the relevantly similar possible worlds that ground capacity judgments. More importantly, though, mere possession of such capacities is insufficient for blameworthiness because fair opportunity also requires *awareness of risk*. Incorporating forward-looking justifications of blame won't solve this problem for capacitarianism. First, appealing to forward-looking considerations doesn't seem to support capacitarianism unless one commits to a teleological theory of moral responsibility that introduces its own issues. Second, even accepting a teleological justification of our norms and practices, it's unclear that blaming agents in the absence of awareness of risk actually develops and sustains a valuable form of self-control.

In section 4, I conceded that capacitarianism has responses to both these problems, even if I find these responses unconvincing. Setting aside these issues, though, I presented what I take to be capacitarianism's most significant advantage over its (non-attributionist) rivals – it's apparent correspondence with our reactive attitudes. In cases of ignorant wrongdoing where the agent lacks awareness of risk, but nevertheless potentially possesses the relevant capacities for

awareness, there's both anecdotal and experimental evidence that we're disposed to blame the wrongdoer. In particular, Murray et al. (2019) conducted behavioral studies suggesting a tendency to hold people responsible for their morally relevant forgetfulness. Furthermore, anecdotal evidence indicates that such ignorant wrongdoers normally feel guilty for their actions.

In response, I argued that there are reasons to put less weight on the significance of these responses. First, there are well-studied cognitive biases that complicate the conclusions one can take from our reactions. Particularly relevant is Royzman et al.'s (2003) I know, you know bias, which involves overestimating others' epistemic state based on one's own knowledge, as well as the fundamental attribution error. Second, blame responses are often influenced by norms that don't directly bear on the blame*worthiness* of the action, and thus it would be too hasty to derive blameworthiness directly from guilt responses to wrongdoing. Finally, some of the intuitions regarding blame for relevant cases of ignorant wrongdoing might rely on non-accountability-based forms of blame. After all, it's unclear that there's a coherent folk concept of moral responsibility that's carefully differentiated from other forms of responsibility in ordinary contexts.

Ultimately, then, although capacities are plausibly central to moral responsibility for ignorant wrongdoing, I maintain that accountability blame also requires awareness of risk. In the next chapter, I attempt to elucidate the best version of such an account. Among the issues that must be contended with are two central questions: (1) what kind of awareness is required for moral responsibility, and (2) what kind or risk must an agent be aware of? As it turns out, answering these questions introduces more puzzles than might appear at first glance.

AWARENESS OF RISK

Introduction

In the previous chapter, I argued that awareness of risk is a necessary condition for blameworthiness because fair opportunity requires such awareness. My primary aim in the context of that chapter was simply to contrast my view with capacitarianism, which denies the necessity of such awareness. A more complete account necessitates that I fill in many details of the awareness of risk condition. Explicating this fuller account is the purpose of this chapter. Although certain elements of the view must be left for later investigation, I hope to provide enough information to generate a unique theory of the epistemic dimension. Ultimately, this view endorses a fairly robust awareness condition while avoiding the revisionism of even stronger accounts maintaining that akrasia is necessary for blameworthiness.

In constructing an account of the awareness of risk condition, I rely significantly on pertinent work in the philosophy of law. In the previous chapter, I mentioned that capacitarianism has roots in Hart's account of culpability for criminal negligence. It shouldn't be surprising that criminal law theory has views on how epistemic considerations bear on the appropriateness of punishment. After all, one of the essential elements of any criminal offense is *mens rea* (or guilty mind), which specifies the required mental state of the agent toward the material elements of the relevant offense.[174] Clearly, one of the most important aspects of this mental state is the agent's *epistemic* state, and so there's significant literature regarding the proper characterization of the epistemic state necessary for criminal liability. Of course, criminal and moral liability aren't equivalent, but assuming that criminal punishment at least requires

[174] The other main element being *actus resus* (or guilty act). As mentioned in chapter two, the four grades of mental states that constitute *mens rea* – according to the Model Penal Code (1985) – are purpose, knowledge, recklessness, and negligence.

moral blameworthiness, discussions in criminal law theory have import for analogous disputes in moral philosophy.

My engagement with relevant work in the philosophy of law will thus be unidirectional – drawing on insights from legal theory to help shape my account of the awareness of risk condition without presuming that this account should also apply back to the criminal law. I will also rely on legal theory to help structure the discussion in this chapter. Note that acting with awareness of risk is essentially the concept of *recklessness*, which I briefly explicated in my discussion of Reckless Ralph. In filling in the details of the awareness of risk condition, I will use the Model Penal Code (1985) definition of recklessness as a template. This definition describes acting recklessly as follows:

> A person acts recklessly with respect to a material element of an offense when he consciously disregards a substantial and unjustifiable risk that the material element exists or will result from his conduct.[175]

I choose this definition as my template for two main reasons. First, its prominence in criminal law and legal theory allows me to engage with the relevant literature. Second, it contains the components that are central to a thorough account of the awareness of risk condition, including substantiality and justifiability.

Using the Model Penal Code definition of recklessness, then, I structure this chapter according to the four main elements: (1) awareness, (2) risk, (3) substantiality, and (4) justifiability. In section 1, I articulate the kind of awareness necessary for blameworthiness, both in terms of the type of mental state and the way that state must be entertained. With an account of awareness in hand, I then turn to the content of that awareness. In section 2, I consider how to

[175] The Model Penal Code (1985) goes on to explain that "the risk must be of such a nature and degree that, considering the nature and the purpose of the actor's conduct and the circumstances known to him, its disregard involves a gross deviation from the standard of conduct that a law-abiding citizen would observe in the actor's situation" (§ 2.02(2)(c)).

characterize the risk(s) that an agent must be aware of while acting. An important dispute here is how much to *subjectivize* risk; that is, whether what matters for culpability is simply how the agent appraises the risk, or whether the standard is more objective. Ultimately, I argue for an intermediate account that takes the perspective of a reasonable person with the agent's available evidence. In sections 3 and 4, I address lingering issues concerning justifiability and substantiality. I endorse the orthodox view that an agent needn't believe that the risk she's imposing is unjustified in order to be blameworthy. Nevertheless, I argue that an agent's blameworthiness is sensitive to the degree of unjustifiability of the act (i.e., the substantiality), so she is either excused or scarcely blameworthy for barely unjustified actions.

1. Awareness

In the previous chapter, I used 'awareness' in the awareness of risk condition simply as a placeholder term for some cognitive mental state that is necessary for blameworthiness. Obviously, though, a satisfactory account will need to say much more about the character of this mental state. I believe that I can fill in some crucial details in this section, while temporarily bracketing discussion of the necessary *content* of the relevant awareness. Indeed, I intend each section to build on the previous one, so that the full account takes shape throughout the chapter. At this point, this constructive process dictates that I use examples where the other components don't act as confounding variables. Hopefully readers can suspend certain important considerations for later sections as I focus on awareness.

In the philosophy of law, characterizing the awareness relevant to recklessness in the criminal law is critical, especially insofar as this awareness is essential to distinguishing

recklessness from negligence.[176] Yet, Douglas Husak (2011) and others argue that "the concept of awareness is poorly understood and requires much more elaboration from penal theorists" (p. 214). Unfortunately, things aren't much better in moral philosophy. Although some important distinctions have been made, it's not always clear how these different distinctions interact, and it's possible that we just currently lack the conceptual resources to capture the relevant phenomenology. Nevertheless, there are enough bright lines for me to take a stand on some crucial issues regarding awareness.

1.1 Starting with the type of mental state necessary for blameworthiness, the first significant choice point is whether *knowledge* is necessary. The conditions for knowledge are themselves highly disputed, but there's general agreement that knowledge requires a *true* belief that's somehow *justified*. Given this basic conception, there are compelling reasons to reject a knowledge condition for blameworthiness. First of all, blameworthiness doesn't appear to require that an agent's belief regarding risk be (epistemically) justified. Suppose, for example, that Abby intentionally spoons arsenic into Martha's tea, effectively poisoning her. Now imagine that Abby obtained the supposed arsenic from a dealer with a reputation for passing off inert substances as powerful poisons. Moreover, suppose that Abby tested the substance before spooning it into Martha's tea and it tested negative for arsenic. Even though Abby is seemingly unjustified in believing that her actions risk harming Martha, she's still plausibly blameworthy for the poisoning.[177] Therefore, blameworthiness doesn't require that an agent's belief regarding risk be justified.[178]

---

[176] Notably, Moore and Hurd (2011) argue that awareness isn't what uniquely distinguishes negligence from recklessness. Instead, they claim that "one needs Magruder's objective magnitude of imbalance, in addition to awareness, to mark this significant breakpoint in culpability" (p. 150).

[177] This example is adapted from Rosen (2008).

[178] Stark (2020) presents another interesting type of case wherein agents appear blameworthy despite having unjustified/unreasonable beliefs about risk. In these *overestimation* cases, "(i) it would have been reasonable for the defendant to believe that the probability of harm x was lower than what she, in fact, believed, but (ii) the risk of

Secondly, and perhaps more controversially, blameworthiness doesn't appear to require that an agent's belief regarding risk be true. Cases supporting this view include instances of what the criminal law calls "abstract endangerment." In basic terms, abstract endangerment involves conduct that, although not *actually* (or "concretely") risky, is hypothetically risky in some way.[179] Consider the following example from Findlay Stark's (2020) discussion of recklessness:

> Parker was a lodger (an informal subtenant) in a semidetached Council house (a form of social housing) leased to Smith. Smith was dissatisfied with her accommodation. In a misguided attempt to help her, Parker set fire to Smith's sofa (couch) whilst she was out, hoping to render the house uninhabitable and cause the Council to rehome her. Parker did not check if his neighbours in the connected property were home before starting the fire. Assume, for the purposes of this article, that Parker believed that his neighbours might be home and in danger of being killed in the fire. (pp. 10-11)[180]

Now suppose that Parker's neighbors were actually on vacation for the week. Parker's conduct seems blameworthy, even though his belief that he was imposing a risk was false.[181]

If cases of abstract endangerment don't elicit the intended intuition, consider instances of (unreasonable) underestimation of risk.[182] Suppose, for example, that Parker's neighbors were actually home when he set fire to Smith's sofa, but Parker unreasonably and falsely judged that

---

harm x that it would have been reasonable to believe existed would still have been unjustified to take, in all the circumstances" (Stark, 2020, p. 28). For example, imagine that Carrie unreasonably overestimates the risk of harm of driving 20 mph over the speed limit in a school zone. Even if that belief is unreasonable – perhaps because there is no one in sight – Carrie is still plausibly blameworthy for driving over the limit if the reasonable estimation of risk is also unjustified. Such cases represent an exception to Stark's general rule that recklessness requires reasonable belief.

[179] As Stark (2020) explains it, abstract endangerment "involves a situation where, although an interest was not put 'concretely' in the path of harm by Φing, Φing is the type of activity that tends toward 'concrete' endangerment, even if each token of that activity does not involve it" (p. 12).

[180] Stark adapts this example from the English case of *R v Parker*. For the purposes of the example, assume that Parker's uncertainty about his neighbors being home is insufficient to undermine the ascription of a genuine belief that they were home. One way to make this assumption explicit would be to replace 'might be' with 'were' in the description of his mental state.

[181] In criminal law terms, he would be guilty of abstract reckless endangerment, assuming that this is a substantial and unjustifiable risk. Of course, Parker is guilty of actual reckless endangerment regarding the neighbors' property, but the focus here is on the risk to the neighbors' personal safety. I don't believe that this concrete risk necessarily confounds the case. Imagine (for some reason) that Parker instead released a poisonous gas that he knew might travel to the neighbors' property. In this case, his conduct clearly still seems blameworthy, even though there's no actual reckless endangerment regarding the neighbors' property under these conditions.

[182] I will return to underestimation cases in section 2.2, where I use them to argue against the subjective account of risk.

his actions posed almost no risk to them. Plausibly, Parker's false belief about the *magnitude* of the risk he was imposing wouldn't excuse him from blame. If it did, then many paradigmatic cases of recklessness would similarly be excused.[183] Instead, I argue that intuitions suggest that true belief, and thus knowledge, isn't required for blameworthiness.

1.2 If the foregoing is correct, then the type of mental state necessary for blameworthiness is simply belief; that is, an agent need only properly believe that her actions risk something unjustifiable. Yet, if beliefs are just basic propositional attitudes, then there are many ways that an agent might believe something. Unfortunately, the taxonomy of belief is often unclear, but I will do my best to articulate an account of *how* an agent must be aware of risk in order to be blameworthy. In doing so, I hope to remain as neutral as possible on the metaphysics of belief – i.e., whether beliefs are best understood in terms of representations, dispositions, functional states etc.[184] However one understands the nature of belief, there are different ways that an agent can have this attitude.

In attempting to usefully categorize different types of belief, I first posit that beliefs have three main relevant properties: (1) temporality, (2) saliency, and (3) availability. These properties interact in important and interesting ways, but I will try to discuss them in isolation first. The temporality of beliefs references the fact that beliefs come and go. If I read a sign that says, 'road work ahead,' then I might form the corresponding belief in the moment. Obviously, I didn't have this belief before reading the sign, and I might lose it later. In the literature regarding the epistemic dimension, it's common to make a relevant distinction between *occurrent* and

[183] For Stark (2016, 2020), cases where there the agent unreasonably *under*estimates the relevant risk are instances of (possibly culpable) *negligence*. This is why Stark (2020) doesn't include such cases in his above discussion of recklessness. I prefer to understand unreasonable underestimation of risk as a form of recklessness, but what's most important for current purposes is that we agree that such unreasonable beliefs aren't necessarily excusing.
[184] Cf. Stark (2016), who explicitly defends a dispositional account of belief for his conception of awareness of risk in the context of the criminal law.

dispositional beliefs. Occurrent beliefs are usually characterized as *conscious* beliefs in the moment. Yet this description obscures important distinctions, as consciousness isn't an essentially temporal feature of beliefs. I might believe that it's raining either consciously or unconsciously in the moment. Moreover, I can consciously believe that it's raining three weeks ago or at this very the moment. Because of this, it would be better to simply define occurrent beliefs as beliefs in the moment and omit any mention of consciousness in the definition.

Given this revised conception of occurrent beliefs, the relevant contrast class is just beliefs that *aren't* held in the moment. Unfortunately, this isn't how dispositional beliefs are usually defined. Instead, they're often negatively characterized as nonconscious beliefs, which clearly won't work here. Moreover, insofar as these dispositional beliefs involve dispositions, this invites further problems. First of all, this definition would rule out non-occurrent beliefs that aren't dispositions to believe. But it's unclear that all beliefs that aren't held in the moment are dispositions to believe; and thus, the dispositional element makes the category too narrow to contrast with occurrent beliefs. Secondly, *dispositionalism* is a general theory of belief that characterizes beliefs as essentially dispositions. If dispositional beliefs are contrasted with occurrent beliefs, then, occurrent beliefs would seemingly be ruled out by a strict dispositionalism.[185] Yet, presumably even dispositionalists want to account for the temporality of beliefs. Because of these problems, I propose that the better temporal distinction is simply between occurrent and non-occurrent beliefs, where this indicates whether the belief is manifested in the moment.

[185] Schwitzgebel (2019) makes similar point: "In fact, a strict dispositionalism may entail the impossibility of occurrent belief: If to believe something is to embody a particular dispositional structure, then a thought or judgment might not belong to the right category of things to count as a belief. The thought or judgment, P, may be a *manifestation* of an overall dispositional structure characteristic of the belief that P, but it itself is not that structure" (Occurrent Versus Dispositional Belief, para. 5).

Having hopefully clarified the taxonomy of temporality, I can progress to a discussion of saliency.[186] As previously mentioned, occurrent beliefs are often problematically described as conscious beliefs in the moment, even though consciousness actually relates to another dimension of belief – one that I term saliency.[187] This dimension is hazier than temporality, but it essentially involves an agent's *level* of awareness regarding particular content. I will further elucidate this definition shortly, but hopefully the basic phenomenon is familiar. While driving to the grocery store, Alanis might recognize that she ought to turn right at the next light as she recites her grocery list in her head. The grocery list is currently foregrounded in her awareness, even though she has genuine beliefs about both objects. One way that philosophers often explain this distinction is in terms of *implicit* and *explicit* beliefs. The kind of beliefs relevant to routine or habitual actions (like driving a familiar route) are usually implicit, whereas explicit beliefs occur at a higher level of awareness.[188]

Given that consciousness relates to saliency, it might be tempting to identify explicit belief with conscious belief. Unfortunately, like many issues involving consciousness, things aren't that simple. Not everyone agrees that consciousness is a unitary concept. Ned Block, for instance, makes an influential distinction between *phenomenal* and *access*-consciousness. As Block (1995) explains it, "phenomenal consciousness is experience; the phenomenally conscious aspect of a state is what it is like to be in that state. The mark of access-consciousness, by contrast, is availability for use in reasoning and rationally guiding speech and action" (p. 227). If

[186] Saliency is closely related to what Moore and Hurd (2011) term "vividity," although my overall taxonomy of belief differs from theirs.
[187] Although I describe temporality and saliency as orthogonal properties, saliency is technically an aspect of occurrent beliefs. This is because saliency refers to the level of awareness regarding particular content *at a given moment*.
[188] Ultimately, I prefer to render saliency as a more scalar concept that admits of multiple levels of awareness, rather than in terms of the implicit/explicit binary. Nevertheless, it's common to refer to the highest level of awareness as explicit belief and the lower levels as implicit beliefs.

this type of distinction makes sense – and most agree that it does[189] – then it appears to thwart the attempt to distinguish explicit and implicit beliefs in terms of either notion of consciousness. After all, Alanis is plausibly access-conscious of *both* her grocery list and the fact that she ought to turn right at the next light. Moreover, an agent can have implicit beliefs regarding things that she's phenomenal consciousness of. Alanis might be phenomenally conscious of the color of the car in front of her as she drives, for example, and yet that consciousness might be in the background of her awareness.

A better explanatory concept is *attention*. The precise nature of attention is disputed, but the basic phenomenon involves a kind of mental selectivity.[190] Importantly, this kind of selectivity is distinct from consciousness, even if the two concepts are closely related. As Christopher Mole (2008) explains:

> The commonsense view is that everything to which one pays attention is, necessarily, something of which one is conscious, not because commonsense is committed to the view that consciousness is a necessary *prerequisite* for attention, but because attending to something is treated by commonsense as *a way of* being conscious of it. (p. 89)[191]

Regardless of whether or not implicit beliefs involve a certain kind of consciousness, then, the commonsense view entails that they don't necessarily involve attention. This account appears to accurately describe Alanis; her attention is on her grocery list, instead of the fact that she ought to turn right at the next light. In this way, attention is characteristic of explicit beliefs.

Finally, I turn to availability. Recall that Block maintains that access-consciousness involves a mental state's *availability* for use in reasoning and rationally guiding speech and action. Others have recognized a similar, significant property of mental states more broadly. For

[189] A similar dichotomy can be found in Dennett's (1969) distinction between *aware*$_1$ and *aware*$_2$, and Moore's (1993) distinction between the *conscious* and the *preconscious*.
[190] For an overview of the literature on attention, see Mole (2021).
[191] Mole (2008) goes on to defend the commonsense view against prominent objections.

instance, Neil Levy's (2014) own theory of the epistemic dimension requires that "contents that might plausibly ground moral responsibility are *personally available* for report (under report-conducive conditions) and for the driving of further behavior…" (p. 31), clarifying that "information is personally available…when the agent is able to effortlessly and easily retrieve it for use in reasoning *and* it is online" (p. 33).[192] Setting aside the issue of what it means for information to be online,[193] Levy's reference to retrieval implicates memory. This is another important feature of beliefs – they can be stored in memory and recalled later at various levels of awareness. Even if I rarely have an occurrent explicit belief that Boston is the capital of Massachusetts, for example, it's available to me in memory. This might not be the only way a belief can be available, but it's an important one.

Just like saliency, availability comes in degrees. Levy's definition of personal availability requires effortless and easy retrieval, but both concepts are plausibly scalar. For example, I have a belief about who won the 1948 United States presidential election, but it's much easier to retrieve my belief about who won the most recent election. A belief can also be temporarily unavailable. I might not be able to recall the capital of Paraguay in time at a trivia night, even if I would eventually remember. Permanently forgotten beliefs aren't available, though, even if they're somehow still stored in memory.

Importantly, the retrieval process that is characteristic of this notion of availability is *non-inferential*, meaning that it doesn't rely on other beliefs to occurrently manifest the pertinent belief. Indeed, I maintain that this non-inferentialism is characteristic of the possession of beliefs

[192] Levy (2014) admits that his own account is close to Block's notion of access consciousness, although he prefers to use 'awareness' for the relevant mental state "because [access consciousness] builds into its definition availability to a broad range of consuming systems" (p. 36).
[193] According to Levy (2014), "any state that actually guides an agent's behavior is online, but, notoriously, states that guide behavior may be personally unavailable to the agent…" (p. 32).

in general. If asked which has more letters – the last name of the 1948 United States presidential election winner or the name of the capital of Paraguay – I could reason to an answer based on my component beliefs, but this process would be forming a new belief, not retrieving an available one. This difference in the etiology of beliefs will be relevant later in the paper, but for now it's enough to simply emphasize that the property of availability at hand doesn't apply to beliefs that require inferences to manifest.

To briefly summarize, then, beliefs have three main relevant properties: (1) temporality, (2) saliency, and (3) availability. In terms of temporality, beliefs are either occurrent or non-occurrent, indicating whether the belief is manifested in the moment. In contrast, saliency comes in degrees, depending on the level of attention by the agent. Implicit beliefs are those with little or no attention, whereas explicit beliefs involve high levels of attention. Finally, availability represents the ease of retrieval of a given belief, where retrieval is a non-inferential process that leads to the manifestation of a stored belief.

1.3 Now that the theoretical structure is in place, I can better explain the kind of awareness necessary for blameworthiness. Recall that the Model Penal Code definition of recklessness characterizes the necessary mental state in terms of *conscious* disregard. As demonstrated in the preceding discussion, consciousness is a poor and incomplete description of an agent's awareness. In the context of the criminal law, this obscurity makes it difficult to demarcate recklessness, especially in relation to negligence. A better definition of recklessness would be more thorough about the type of mental state necessary for this level of *mens rea* culpability. Another related problem with characterizing recklessness in terms of conscious disregard is that, in the absence of disabusing language, one might be tempted to read consciousness as occurrent explicit belief. Such an interpretation would set a high bar for this level of culpability, which

would be particularly troublesome insofar as negligence is a problematic desert basis for punishment (either for particular offenses or more generally).

I aim for a much clearer and detailed account of the relevant mental state in the awareness of risk condition. Moreover, if the Model Penal Code conception of recklessness is understood as involving occurrent explicit belief, I intend a significantly weaker condition. Specifically, although I also hold that occurrent belief regarding risk is required for blameworthiness, such belief needn't be explicit on my account. This means that an agent can be blameworthy for imposing an unjustified risk even if she's only occurrently aware of this risk at some lower level of awareness, as her attention is on other matters.[194] Indeed, most blameworthy recklessness plausibly takes this form, especially insofar as explicit awareness of risk involves an agent attending to the fact that her actions carry specific risks.[195]

However, although my account of awareness is significantly weaker than the above Model Penal Code version, it's still controversial. One notable upshot of the view is that it excuses certain agents who forget morally relevant information. Consider this example from Fernando Rudy-Hiller (2017):

> Jill is preparing a small birthday party for her five-year-old son. She's about to bake a cake and other treats for the children when it occurs to her to check with other parents whether any of the kids has a food allergy. Only one of them, Rob, tells Jill that his daughter is severely allergic to nuts. However, as soon as she hangs up the phone and turns her attention to other tasks, Jill forgets what Rob has just told her and so when she's finally mixing up the ingredients for the cake she isn't aware that she shouldn't put nuts in it. (p. 414)

[194] Importantly, the agent's implicit awareness of risk must involve sufficient availability. Thus, even if her attention is on other matters, she should be able to turn her attention easily and effortlessly to the relevant risk. I discuss this issue in response to an objection later in the section.

[195] Realistic instances of recklessness involve awareness of multiple risks, but I simplify cases for ease of exposition. Still, blameworthiness only requires requisite awareness of *one* unjustifiable risk. Of course, an agent will only be blameworthy for the outcomes of her recklessness if these outcomes appropriately match the risk that she was aware of (see section 2.3 for relevant discussion). Thus, an agent who is only aware of one risk of her conduct isn't blameworthy for a different risk materializing.

Assuming that Jill's lack of awareness involves the absence of an occurrent beliefs that baking the cake with nuts risks triggering the allergy, then, she isn't blameworthy for putting nuts in the cake on the proposed account.

Importantly, Jill *would* be blameworthy if her attention merely shifted away from the allergy while she was mixing up the ingredients. In this case, she would still have had an occurrent implicit belief about the allergy, which is sufficient for blame. Moreover, Jill would also be blameworthy for putting nuts in the cake if she was previously (occurrently) aware of the risk of *forgetting* and failed to take steps to prevent it.[196] Here, Jill would resemble Reckless Ralph from chapter two. It's difficult to know how often people are aware of the risk of forgetting in these types of situations in the real world. Regardless, though, this awareness needn't be explicit to justify blame.

Setting aside these variations of the case, though, many have the intuition that Jill is blameworthy in the original example, and so the proposed account appears underinclusive with respect to blameworthiness. I will address this challenge later in the chapter. Before turning to the next component of the awareness of risk condition, though, I want to consider an argument that the current account of awareness is *over*inclusive in a certain way. The argument begins with the observation that there's a substantive difference between the driver Alanis and the version of Jill where she has an occurrent implicit belief about the allergy as she mixes up the ingredients. Specifically, the driver who recognizes that she ought to turn right at the next light is acting out of habit, whereas Jill is not. Because of this, the reasoning goes, it's not clear that Jill's implicit

[196] In this case, we can *trace* Jill's blameworthiness for putting nuts in the cake to her culpability for recklessly omitting to take steps to prevent forgetting. It's these kinds of cases that Nelkin and Rickless's (2017) *Opportunity Tracing view* primarily attempts to capture. I aim to also explain cases of ignorant wrongdoing where this tracing strategy is unavailable.

belief is accessible in the right way; and without such accessibility, Jill lacks a fair opportunity to avoid wrongdoing.[197]

In response, I first agree that there's a real difference between habitual actions and actions involving other kinds of implicit beliefs. When an agent performs a habitual action, she has performed it multiple times before, often initially with accompanying explicit beliefs. At some point, she no longer needs to focus her attention on these beliefs to perform the action, as the information has become hardwired. In contrast, Jill just found out about the allergy and is presumably attempting to bake the cake without nuts for the first time – her recent belief about the allergy hasn't been similarly hardwired. Furthermore, in cases of habitual actions, the agent's implicit beliefs are also guiding her actions, whereas for Jill the whole problem is that her implicit beliefs are *not* manifested in her actions.

Nevertheless, although there's a real difference between habitual actions and actions with other kinds of implicit beliefs, I maintain that *both* involve the availability (or accessibility) required for blame. One reason this availability might seem dissimilar is that it's more obvious with habitual actions. Because an agent's implicit beliefs are manifested in her behavior with habitual actions, we assume that these beliefs are easily retrieved. If the relevant street was blocked off, for instance, we assume that Alanis would shift her attention to the fact that she's supposed to turn there and now must find another route. But implicit beliefs like Jill's can be just as available. We're all familiar with the phenomenon of being distracted, only for some pertinent fact to return to our explicit attention. There's no reason to think that this couldn't happen in a case like Jill's. Similarly, there's no reason to think that Jill wouldn't have remembered the allergy had she been appropriately prompted to retrieve the belief. In this way, the *capacity* to

[197] Credit to Dana Nelkin for suggesting this line of argument.

retrieve implicit beliefs can be just as strong in cases of non-habitual actions, even though agents fail to exercise it.

Now, one might worry that my response here problematically reveals my account to be capacitarian, despite arguing against capacitarianism in the previous chapter. However, my disagreement with capacitarianism isn't simply based on the presence of capacities. Indeed, I believe that any plausible theory of moral responsibility appeals to capacities. Instead, my disagreement with capacitarianism is based on the *absence* of awareness. Thus, even though I appeal to capacities here, it's still within an overall account that requires awareness of risk.

2. Risk

Now that I outlined an account of how an agent must be aware of risk, I can move on to the content of this awareness. The first main component of this content is the risk itself, divorced from its justifiability or substantiality.[198] Although the risk element might seem like the most straightforward component, a comprehensive account introduces a host of issues. Indeed, characterizing the risk involved in cases of culpable recklessness is the core of any theory of awareness of risk. At the outset, it's worth acknowledging that although I will usually be discussing *the* risk, realistic scenarios involve multiple risks – because every candidate action has multiple possible outcomes. Thus, determining the justifiability and blameworthiness of a given course of action involves somehow aggregating these risks. I leave this issue of aggregation to others, given the many other fundamental questions I aim to answer in this section.[199] Instead, I focus on the somewhat contrived case of an action with an isolated risk.

[198] Oberdiek (2017) argues that "the concept of risk relevant to a moral framework of risk imposition cannot be a matter of non-normative fact but must itself be moralized" (p. 48). I disagree. At least conceptually, it's both possible and preferable to separate risk from its justifiability and substantiality.

[199] Alexander and Ferzan (2009), for example, propose the following account: "Suppose, as will ordinarily be the case, that a given act increases by varying amounts the risks to various legally protected interests. So the actor might

2.1 The first task is just to define risk as a concept, or to specify the relevant conception for our purpose. As others have pointed out, we use 'risk' to refer to different things in different contexts.[200] Sometimes 'risk' refers to an event, as in, "contracting salmonella is a risk associated with consuming raw meat." Other times, 'risk' refers to the *probability* of some event, as in, "playing football risks getting a concussion." It's this latter conception that's relevant to the current discussion. Specifically, I will define risk as the probability of some negative event, where the moral valence of the event is determined by an independent moral theory. For instance, risk for a consequentialist would be something like the probability of bad consequences. Although it's coherent to discuss the probability of positive events, I take it that risks conventionally involve something unwanted or negative; and given that I'm primarily concerned with blameworthiness, my focus is on this negative side anyway.

My definition of risk as the probability of a negative event is far from revisionary. Indeed, it's a fairly standard conception of risk in this context.[201] Still, it's worth mentioning that it's probably more common in moral philosophy to understand risk as the probability of harm specifically. I render my account in terms of negative events for two reasons. First, I want to be more inclusive and leave room for the possibility of wrongs that aren't harms. Second, couching things in terms of harm would leave my account hostage to a particular theory of harm, which I hope to avoid. Overall, I only want to make substantive theoretical decisions when necessary.

2.2 Given this definition of the concept of risk, the next issue is how to delineate *the* risk that an agent must be aware of when acting. This task instantly leads to complications. Suppose, for

---

believe that act A increases the risk to legally protected interest I_1 by R_1, increases the risk to legally protected interest I_2 by R_2, and so on. His culpability for A is a function of the sum of the risks he imposes on those interests…Even if no one of the risks, viewed in isolation would render the [act A] reckless, the sum of them might" (p. 47).

[200] See Hansson (2022) for an overview of different uses of 'risk'.

[201] See, e.g., Hansson (2022).

example, that a construction worker blindly throws a brick off a roof, striking a pedestrian on the street below.[202] On the proposed account, the worker is blameworthy if he was aware of the (unjustifiable) risk of his actions – that is, the probability of the relevant negative event. But this vague claim requires further clarification. In particular, how should we understand the relevant probability? There are several plausible theories of probability, but the most common interpretation in the literature on risk is *frequentist*. Broadly speaking, frequentism identifies probabilities with relative frequencies of events within some reference class. Yet this immediately leads to a problem: what is the unique event and reference class that could generate this frequency? Should the reference class be bricks thrown off roofs, or bricks thrown from heights more generally? Should the event be described simply as hitting a pedestrian, or hitting a pedestrian in a particular part of the body? As Stephen Perry (2001) argues, "there is no correct or canonical answer to the question of which reference class and attribute we should chose" (p. 98).

Hans Reichenbach (1949) termed this dilemma the *reference class problem*, but its history in probability theory predates even him.[203] Although the issue is particularly stark with frequentist interpretations of probability, Alan Hájek (2007) and others persuasively argue that every plausible interpretation of probability faces the problem at some level. In fact, Hájek (2007) maintains that there are really two problems, a *metaphysical* and *epistemological* version:

> The former problem arises because it seems that there should be a fact of the matter about the probability of X; what, then, is it? The latter problem arises as an immediate consequence: a rational agent apparently can assign only one (unconditional) probability to X: what, then, should that probability be? (p. 565)

[202] Example adapted from Perry (2001), which I believe is itself adapted from an example by Hart (1968b).
[203] Hájek (2007) attributes the origin of the problem to Venn (1876).

Given that the epistemic dimension of moral responsibility is the current focus, the epistemological problem appears most relevant. Regardless of the actual risk (in some metaphysical sense), we want to know what probability an agent ought to assign to a negative event for it to count as *the* relevant risk.

One straightforward response to the epistemological reference class problem is just to identify the risk with whatever probability the agent *actually* assigns to the pertinent event. For example, if the construction worker (implicitly) believes that there's a low probability of hitting a pedestrian with a brick, then that's the relevant risk for purposes of moral assessment.[204] In the context of criminal culpability, Larry Alexander and Kimberly Ferzan (2009) notably advocate this kind of *subjective* account of risk, contending that "there is no gap between the actor's subjective estimate of the risk and the 'true' or 'objective' risk because the latter is either illusory…or arbitrary…" (p. 31). In other words, they argue that objective accounts of risk face a dilemma: if they identify risk with the probability that a rational agent assigns to an event under *full* information, then events will only have a 0 or 1 probability; [205] yet if they identify risk with the probability that a rational agent assigns to an event with incomplete information, then there are any number of seemingly arbitrary perspectives to choose from. [206] As the construction worker prepares to throw the brick off the roof, for instance, there might be several onlookers who see him doing so and yet differentially assess the probability of him hitting a pedestrian,

[204] Indexing risk in this way won't necessarily generate a discrete probability if agents have *vague* or *imprecise* credences. If agents have imprecise credences then the risk will be a range of probabilities, and if agents have vague credences then the edges of that range will be "fuzzy" or indeterminate.
[205] The reasoning here appears to be that in a causally deterministic universe one could accurately predict events based on knowledge of antecedent events and the laws of nature. Alexander and Ferzan (2009) are careful to qualify, "leaving aside quantum events" (p. 29).
[206] It's worth noting that Alexander and Ferzan's use of 'objective' here appears slightly nonstandard. At least according to common usage, objective accounts of risk are totally *perspective-indifferent*, to use Oberdiek's (2017) terminology. However, Alexander and Ferzan seem to conceive of objective accounts as representing some omniscient (or otherwise authoritative) perspective. Although they ultimately argue that risk is an essentially epistemic concept, it seems inappropriate to construe 'objective' as an epistemic notion from the start.

based on their unique vantagepoint. Whose perspective should be taken as authoritative? Alexander and Ferzan argue that every answer is arbitrary and thus favor a subjective account.

If the foregoing is right, then there are two significant reasons to prefer the subjective account of risk.[207] First, insofar as *the* risk is just whatever probability the responsible agent assigns to the event, the subjective account appears to solve the reference class problem rather easily.[208] Second, objective accounts – the competing class of views – face the horns of a dilemma; either they're false or arbitrary. However, although the subjective account plausibly solves the reference class problem, the argument against objective accounts is flawed. Even if Alexander and Ferzan are right that any rational agent with full information would only assign probability 0 or 1 to every event, they fail to adequately support the claim that any perspective other than the agent's is arbitrary. Crucially, the alleged arbitrariness of selecting a perspective other than the agent's is different from the arbitrariness related to the reference class problem.[209] The former notion pertains to the possibility of an *authoritative* perspective (or evidence base), while the latter issue concerns the selection of *any* perspective that can generate a single probability representing the risk. Once these two problems are differentiated, it doesn't follow from the fact that there are multiple possible perspectives that only the agent's perspective is

[207] It's worth noting that the subjective account would still need to address the issue of whether the agent's (*ex ante*) conception of the event matches the actual outcome. For example, if the construction worker believes that there's a 1% chance of hitting a pedestrian with a brick, is this the probability that we should assign to the outcome of hitting two pedestrians via ricochet? That is, how *specific* must an agent's conception of the risked outcome be in order to properly count as awareness of risk; must they believe that their actions risk some outcome in all its detail, or is more general belief sufficient? I will return to this broader issue later in the chapter, but it's importantly distinct from the reference class problem. Furthermore, because every theory of awareness of risk faces this issue of specificity, it's not a disadvantage of the subjective account.

[208] As Hayek (2007) explains, "it thus appears that there is not any interesting reference class problem for the radical subjectivist. The probability that you assign to E is whatever it is. Qua nothing. This is a benefit, if that's the right word for it, of the radical subjectivist's permissive epistemology" (p. 576).

[209] Alexander and Ferzan (2009) aren't always clear about this and seem to run the two issues together. Although they might be interconnected, they're nevertheless different problems.

authoritative. At least, it's dialectically open for an objective account to argue for the authority of a certain perspective that might also solve the reference class problem.

One such candidate perspective is common in criminal and tort law: the *reasonable person* perspective.[210] Although the Model Penal Code (puzzlingly) references the perspective of a "law-abiding person" in its definition of recklessness, it directly mentions the reasonable person in its definition of *negligence*. Specifically, according to the Model Penal Code (1985), an agent's failure to perceive the relevant risk in cases of negligence must "involve a gross deviation from the standard of care that a reasonable person would observe in the actor's situation" (§ 2.02(2)(d)). Thus, the Model Penal Code recognizes the significance of the reasonable person perspective, even if it doesn't use the notion in the definition of recklessness that I use as a template for my own theory. In the current context of defining the relevant risk, an objective account based on the reasonable person perspective would obviously identify the risk with whatever a reasonable person would assign to the pertinent event. For example, if a reasonable person would assign a 75% chance to the construction worker hitting a pedestrian from throwing a brick off the roof, then that's the risk at issue. If the construction worker happens to believe that the risk is much lower, this is immaterial to his blameworthiness.

Of course, the plausibility of this reasonable person approach depends significantly on how one elaborates the details of this perspective. In this way, the concept of the reasonable person is more of a framework than a practical account. One recent, rather sophisticated version

[210] Hart and Honoré (1959) advocate a variation of the reasonable person perspective based on the conceptualization and common knowledge available to ordinary persons. This *ordinary person* approach is also endorsed by Perry (2001), who maintains that "the appropriate characterization of risk is based on the level of knowledge and ability to assess probabilities that an ordinary person in the defendant's position could be expected to possess" (p. 119). However, I think Oberdiek (2017) convincingly argues that "the perspective that matters is not that of an ordinary person understood in purely descriptive terms…but of a person understood in normative and indeed moral terms" (p. 48).

is John Oberdiek's (2017) *evidence-relative* account,[211] according to which, the correct perspective for assessing risk is a reasonable person with the same available evidence as the agent imposing the risk. In reference to the brick-throwing example, this theory would seemingly invalidate the risk assessments of both an onlooker on the street and the construction worker himself. In the case of the onlooker, her assessment would be invalid due to her access to evidence that the construction worker lacks. Whereas, for the construction worker, his risk assessment would be invalid because he unreasonably underestimates the probability of harm. Insofar as this is an accurate application of Oberdiek's account, then, it's initially appealing due to its ability to capture these intuitions.

But can the evidence-relative reasonable person account overcome the reference class problem? Oberdiek seems confident that it can by relying on *moral* constraints built into the reasonable person perspective that narrow the reference class. Specifically, Oberdiek cites both *demandingness* and *contractualist* norms. Regarding the former, Oberdiek (2017) explains that "the set of facts that might be thought to shape the characterization of a particular risk…will be limited at the outset by the demandingness constraints generated by the reasonable person perspective" (p. 61). In other words, because the reasonable person perspective is limited by the *available* evidence, this constraint rules out characterizations of risk that rely on facts that would be unreasonably demanding to discover (i.e., unavailable). For instance, when using a gas stove, the reasonable person's assessment of risk needn't factor in evidence that would require careful

[211] Oberdiek (2017) draws from Derek Parfit's (2011) distinction between fact-, belief-, and evidence-relative conceptions of wrongness. According to Parfit (2011), an act is "*wrong* in the *fact- relative* sense just when this act would be wrong in the ordinary sense if we knew all of the morally relevant facts…*wrong* in the *belief-relative* sense just when this act would be wrong in the ordinary sense if our beliefs about these facts were true…*wrong* in the *evidence-relative* sense just when this act would be wrong in the ordinary sense if we believed what the available evidence gives us decisive reasons to believe, and these beliefs were true" (p. 150-1). Oberdiek adapts the notion of evidence-relativity for his own purpose.

examination of the entire fuel system.[212] Oberdiek (2017) then supplements this demandingness constraint with the contractualist norm that "risk must be characterized to be least objectionable to the individual to whom it is most objectionable" (p. 7).[213] According to him, these two constraints entail that "*the* risk will be the maximum credence that a reasonable person would assign to it" (p. 64). As a maximum, this yields a unique probability, thus solving the reference class problem.

Insofar as Oberdiek's account represents proof of concept, then, an objective account of risk shouldn't be ruled out from the outset. But is an objective account actually preferable to a subjective one? Both approaches have costs and benefits. In terms of theoretical considerations, the subjective account is certainly simpler than Oberdiek's, and it relies on fewer moral commitments to generate a unique risk assessment. At the same time, though, the subjective account appears both over- and underinclusive regarding blameworthiness.[214] In the overinclusive direction, it might seem harsh to always blame agents for their *perceived* recklessness, even when they're not actually imposing unjustifiable risks. Consider this example from Alexander and Ferzan (2009):

> David wants to get home in time to watch the Lakers game on television. He accelerates until his speedometer reads ninety miles per hour, a speed that he believes creates a very substantial risk of death, serious bodily injury, or property damage to other drivers, passengers, and pedestrians. In fact, his speedometer is broken, and he is going only fifty-five miles per hour, a reasonable speed given the road and traffic conditions. (p. 27)

[212] This example comes from Thomson (1986), although she uses it to draw a different conclusion.
[213] Oberdiek (2017) calls his account *epistemic contractualism* and explains that "epistemic contractualism is an extension of moral contractualism in that it requires characterizations of risk to be mutually justifiable, that is, justifiable to both agents and those individuals who might be affected by the agent's risky conduct" (pp. 59-60).
[214] Similarly, Robinson (2003) argues that "this view that subjective risk-taking ought to be the only focus of criminal law that has caused both the regular overgrading of recklessness offenses where no risk in fact has been created and the regular undergrading – indeed, exclusion from liability – of culpable inattentiveness that results in the creation of a prohibited risk" (p. 27).

Even though David's willingness to put others in danger for a clearly insufficient reason reflects poorly on his character, it's not clear that his *actions* are blameworthy. After all, no one was actually put in danger by David's driving. Indeed, his actions seem even less blameworthy were he unreasonable in his beliefs regarding risk. Imagine, for instance, that David knew that his speedometer often malfunctioned, and he had just driven by a radar speed sign indicating that he was driving 55 mph. Under these conditions, David doesn't appear blameworthy to many.[215]

In the underinclusive direction, the subjective account seems to excuse too many agents who don't believe their actions are (unjustifiably) risky. Suppose that our construction worker judges the risk of hitting a pedestrian by throwing a brick off the roof to be infinitesimal. According to the subjective account, this unreasonable judgment determines his culpability; were he to hit a pedestrian, the construction worker would be excused on the basis of the slight risk he believed to be taking.[216] Yet, I believe that most people would be disposed to blame the construction worker for his carelessness. In general, considered judgments suggest that agents can be wrong about the risk of their actions, and at least some of the time they're blameworthy for acting on these false beliefs.

Indeed, regardless of whether one agrees with the details of Oberdiek's account, it appears that what's missing from the subjective account is precisely some notion of *reasonableness*. Although considered judgments might acknowledge some subjectivity in terms

[215] In section 1.1, I used Stark's (2020) example of Parker the lodger to explain why true belief regarding risk isn't required for blameworthiness. One might wonder, then, why David should be excused on the basis of his false belief. My honest answer is that I actually agree with Alexander and Ferzan that David is blameworthy, but I think that most people would disagree. In the current context, I'm attempting to assess the costs and benefits of objective and subjective accounts of risk from a more neutral perspective, which doesn't rely on other commitments. Furthermore, I believe it's really the *under*inclusiveness of the subjective account that drives me (and others) away, so I don't need to rely on the overinclusiveness to justify going more objective.

[216] I discuss justifiability in the next section, but I assume here that taking very minor risks is justifiable in most contexts, as long as there is some appropriately countervailing reason. In this case, we might suppose that throwing the bricks without checking below is much more expedient.

of the evidence-relative nature of risk, it's also intuitive that evidence should constrain belief. Given that the construction worker has evidence that pedestrians might be walking under the roof, it's unreasonable for him to believe that throwing a brick off it imposes infinitesimal risk. In this way, the reasonable person perspective better captures intuitions than the subjective account in a range of cases.

Still, the reasonable person approach has its own issues. First, it's unclear that it can convincingly solve the reference class problem. Oberdiek's (2017) ingenious strategy to "plumb the account's normativity" (p. 40) relies on introducing controversial moral commitments that require their own defense. Ideally, one could identify a single authoritative perspective without demanding acceptance of a contractualist moral framework. More importantly, though, there's still Alexander and Ferzan's lurking arbitrariness worry; specifically, the reasonable person perspective might appear *morally* arbitrary. Even if it's true that the reasonable person would assign a higher risk to throwing a brick off the roof than the construction worker, for example, one might question why that should affect the *worker's* blameworthiness?[217] Perhaps intuitions regarding such cases are actually motivated by a hidden assumption: anyone who acts so carelessly must lack concern for the interests of others. Crucially, however, insufficient concern can come apart from risk misjudgment. In such cases, the reasonable person approach needs a convincing argument for why such misjudgment should be blameworthy.

Providing such an argument for the reasonable person approach is beyond the scope of this chapter. Nevertheless, given the counterintuitive implications of the subjective account, I

[217] I believe that the evidence-relative reasonable person account vitiates some of the force of the original arbitrariness worry. Unlike the fact-relative approach, the evidence-relative account is indexed to part of the agent's actual perspective. Nevertheless, there remains a kind of arbitrariness worry.

believe that some version of a reasonable person account must be right.[218] Indeed, such a view might not even need to rely on all the controversial moral commitments that Oberdiek does. Although, as a metaphysical issue, we might be concerned that our theories don't generate a unique probability of some event, this result isn't as worrying at the epistemological level (at least in the current context). Imagine that the perspective of a reasonable person with the agent's available evidence only generates a *range* of probabilities regarding some outcome. Still, as long as that range is completely above the threshold for unjustifiability, it seems that we still have a conclusive answer as to whether the agent should take the risk.[219] In this case, the reasonable person approach avoids the worst implications of the reference class problem, even if it doesn't solve it.

Where does this all leave the awareness of risk condition? I've argued that *the* risk relevant to blameworthiness is determined by the reasonable person perspective,[220] and thus an agent who misjudges the risk of her actions can still be held accountable. But what then must an agent be aware of to satisfy the awareness of risk condition? Obviously, she needn't be aware of the risk relevant to blameworthiness, otherwise misjudgment would necessarily excuse. Instead, my account simply demands *any* occurrent belief about the probability of some negative outcome of one's actions. Even if the construction worker falsely believes that there's a low probability of hitting a pedestrian with a brick, then, such awareness is sufficient for

[218] I also rule out any objective account that argues that an agent's assessment of risk must match the *actual* risk, however that's defined. This kind of fact-relative account is indeed morally arbitrary; it's unfair to blame an agent on the basis of facts that she doesn't have appropriate access to when acting.

[219] A more difficult kind of case is one where the range straddles the line between justifiability and unjustifiability. I would argue that one is potentially blameworthy as long as the range isn't totally over the threshold for justifiability. However, as I will explain in section four, it might be that this kind of close call is either excused or involves significant mitigation of blame.

[220] Rather than taking the reasonable person perspective, one could argue that the risk relevant to blameworthiness is the risk that the *actual* agent could reasonably be expected to judge, given her evidence. For current purposes, I take these two notions to be equivalent. The fundamental point is that there must be a reasonableness constraint on risk assessment, indexed to the available evidence.

culpability.[221] I contend that this account provides the agent with a fair opportunity to avoid wrongdoing, while also recognizing a reasonableness constraint on the risk relevant to blameworthiness.

2.3 Finally, every account of awareness of risk must address the issue of how *specific* an agent's belief about risk must be.[222] If the construction worker ends up hitting two pedestrians with a brick via ricochet, for example, is he blameworthy for this outcome if he was only aware of the more general risk of hitting a pedestrian?[223] The more specific we require the agent's awareness to be, the less often she'll be blameworthy. Yet, a totally general belief about the riskiness of one's actions seems insufficient for blaming for all outcomes. The challenge is appropriately delineating the required specificity of risk.

Once again, Alexander and Ferzan's account represents one end of a spectrum of views. When faced with the reference class problem, their solution is to deny any conception of risk other than the agent's; similarly, when faced with the specificity challenge, their solution is simply to deny the significance of results.[224] On Alexander and Ferzan's (2009) view, an agent's blameworthiness is totally captured by his *choices*, and the consequences of those choices are irrelevant, "assuming that the actor has taken the last step he believes necessary to unleash the risk and that the relevant level of risk is now beyond his control" (p. 192). Given this theory, the

[221] A difficult kind of case is one in which the agent judges that the risk is *zero*. In this case, it seems accurate to say that the agent doesn't actually believe that there's a risk. For example, I think it's accurate to say that I don't believe that there's a risk that two could be greater than one. Now, it seems irrational to judge anything other than logical inconsistencies as having zero risk, but there might be such irrational agents. My intuition regarding these rare cases is that these agents don't satisfy the awareness of risk condition.
[222] For discussion, see Fischer and Tognazzi (2009) and Vargas (2005)
[223] I'm setting aside the issue of the reasonable person perceptive for the moment to simplify the current discussion. Assume that the construction worker has reasonable beliefs regarding risk going forward.
[224] In other words, Alexander and Ferzan (2009) reject *resultant luck*, i.e., luck in the way things turn out. See Hartman (2017) and Moore (1997, 2009) for arguments in favor of resultant luck.

construction worker would be equally blameworthy for recklessly throwing the brick off the roof, whether it hit one, two, or even *zero* pedestrians.

Alexander and Ferzan's "results don't matter" view fits with their subjective account of risk, as well as their broader theory of culpability. As they consider *insufficient concern* as the "essence of culpability," it makes sense that the construction worker would be equally blameworthy whether he hit a pedestrian or not.[225] After all, the results of the action are immaterial to the attitude of the construction worker toward the potential of harm to pedestrians. As with the reference class problem, denying the significance of results also dissolves the problem about specificity, given that there's no need to link outcomes to awareness. Again, though, this solution comes at a significant cost, generating counterintuitive implications in a number of situations. Most people share the intuition that results matter for blameworthiness in some way, so denying their significance seems like a last resort strategy.[226] If there's another account that doesn't contravene these considered judgments, it's preferable *ceteris paribus*.

My own solution is a mixed account that recognizes different requirements at different levels of specificity. Recall that in the earlier discussion of awareness, I maintained that occurrent belief of risk is necessary for blameworthiness (although it needn't be explicit). The more complete version of this view is that, in order to be blameworthy for some outcome, the agent must have: (1) an *occurrent* belief of the *general* risk of her conduct, and (2) a *disposition*

[225] I assume here that the construction worker expresses insufficient concern in recklessly throwing the brick off the roof. Given Alexander and Ferzan's subjective account of risk, this means that his own assessment of the risk was (objectively) unjustifiable. This assumption isn't necessary on my view; the construction worker can be blameworthy even if his actions don't express insufficient concern.

[226] An interesting strategy proposed by Robichaud and Wieland (2017) involves the distinction between the *degree* and *scope* of blameworthiness. The basic idea is that results might increase the scope of blameworthiness but not the degree. This strategy would capture the intuition that results matter without entailing that the agent is more blameworthy for results that seem out of her control.

to believe the *specific* risk of her conduct.[227] For example, the construction worker is blameworthy for hitting two pedestrians with a brick if he was occurrently aware that throwing a brick off the roof risked harm and disposed to believe that throwing the brick risked hitting two pedestrians via ricochet. In this way, the account requires both a manifested belief and a disposition to believe.

The occurrent element of the mixed account should be fairly clear, but the dispositional element requires elaboration. Recall that I mentioned earlier that occurrent beliefs are frequently contrasted with "dispositional beliefs," which are usually negatively characterized as nonconscious beliefs. Setting aside the issue of whether nonconscious beliefs should be labeled dispositional (they shouldn't), it's important to point out that these beliefs are fundamentally different from the dispositional element of my account. The dispositionalism figuring into my account is best captured by what Robert Audi prefers to call "dispositions to believe." According to Audi's (1994) distinction, the essential difference between dispositional beliefs and dispositions to believe is "between accessibility of a proposition by a retrieval process that draws on memory and its accessibility only through a belief formation process" (p. 420). In other words, unlike dispositional beliefs, disposition to believe require some *mediating process* to actually manifest in thought. For example, consider again the case of Jill and the cake. If asked whether anyone who might eat the cake has a nut allergy, she need only retrieve the forgotten proposition from her memory that Rob's daughter is allergic.[228] But if asked if the cake is vegan, she might have to do some thinking if she has never considered this proposition. She might recall

[227] Although I use the phrase "*the*…risk" here, this just refers to any probability of the relevant negative event. This is consistent with my claim in section 2.2 that the awareness of risk condition only requires *any* occurrent belief about the probability of some negative outcome of one's actions. Furthermore, note that these two awareness conditions are necessary but not sufficient for blameworthiness. The relevant risk must also satisfy the justifiability and substantiality criteria. In the case of the construction worker, I assume that these criteria are met. Later, when I discuss these criteria in detail, this assumption will be confirmed.

[228] Assuming that she only temporarily forgot about the allergy and the belief is sufficiently available.

both the definition of veganism and the ingredients in her recipe, and then infer that it's not vegan. In this case, Jill only has a disposition to believe the cake isn't vegan, because although she has the raw materials (or grounds) for forming this belief, some mediating process was necessary to manifest it.

In the case of the construction worker, he only needs a disposition to believe that his conduct risks the specific outcome that materializes in order to satisfy this level of awareness of risk. Now, the precise nature of dispositions is notoriously fraught, and certainly beyond the scope of this chapter, but I hope to fill out some of the details of this disposition to believe. First of all, the grounds for this disposition are at least partly beliefs. For instance, the construction worker presumably has basic beliefs about physics, the weight of a brick, the effect of being hit by a brick etc. He also has *perceptual access* to his surroundings, even if he doesn't form beliefs based on his experience.[229] From these raw materials, then, the construction worker could have a disposition to believe that his conduct risks a specific outcome, even if he hasn't actually formed the belief.[230] One hypothetical check for such a disposition would be a counterfactual test of some sort. Imagine, for instance, that we asked the construction worker whether some specific

[229] Note that both Agule (2022) and Stark (2016) recognize the significance of these background beliefs and perceptions in their discussions of culpability for negligence. For Stark, these beliefs and perceptions ground blameworthiness when the agent's failure to form the relevant belief about risk demonstrates insufficient concern for the interests of others. For Agule, these beliefs and perceptions ground blameworthiness when the agent's *executive processes* engage with these beliefs and perceptions, and these processes play a role in the agent's action or omission. Unlike my account, neither Stark nor Agule require actual awareness of risk at any level of attention, hence they defend culpability for negligence. However, both views are rather similar to my own, especially Agule's. The fundamental difference between my view and Agule's (2022) seems to be his claim that "there is no reason we should privilege the particular working of attention over the other executive functions" (p. 243). I disagree, as I believe that attention is necessary for fair opportunity.

[230] As Audi (1994) explains, "dispositions to believe are higher-order properties belonging to us by virtue of a much smaller number of first-order psychological properties…" (p. 431).

outcome was possible as a result of his conduct. If he would answer in the affirmative, then this indicates the possible presence of the relevant disposition to believe in the actual situation.[231]

An alternative solution to the challenge of specificity is simply requiring some intermediate (occurrent) awareness of risk.[232] However, this strategy faces a dilemma: to capture intuitions in most cases, the necessary awareness of risk will have to be fairly general, yet such a general conception of risk opens the door for counterintuitive cases of blameworthiness. For example, imagine that Scott lets his inexperienced friend target shoot in the woods behind his suburban home, knowing that this risks damage and/or injury. In order to capture the range of negative outcomes that Scott is plausibly blameworthy for, the necessary awareness of risk can't be too specific. Now suppose that something fluky happens; unbeknownst to Scott, the cartridges he purchased for his friend to shoot were incorrectly loaded, causing the gun to explode and injure his friend. Given that Scott was aware that letting his untrained friend target shoot risked damage and/or injury, a general awareness of risk condition would judge him blameworthy for his friend's injuries. Yet, even if Scott is reckless for letting his inexperienced friend target shoot, it seems counterintuitive that he's blameworthy for this fluky accident.[233] Of course, one could try to narrow the awareness of risk condition to rule out flukes like this, but this move might imperil the ability to blame Scott for negative outcomes that he's plausibly blameworthy for.

---

[231] Importantly, I'm only suggesting this counterfactual test as a device for helping to determine whether an agent has the relevant disposition to believe. There are well-documented problems with the strategy of trying to explain dispositions in terms of counterfactuals, and I don't wish to get involved in these disputes.

[232] This is the strategy of Nelkin and Rickless (2017). Although their account doesn't give a fully detailed description of the kind of awareness required, they're clear that the requisite awareness needn't involve full attention.

[233] Assume that Scott was totally unaware of the possibility of this fluky accident, meaning that he didn't even have a disposition to believe that the gun could explode from incorrect loading of the cartridges. In this case, it seems more plausible that whoever incorrectly loaded the cartridges is *fully* responsible for his friend's injuries. Contrast this case with a version where Scott was aware of this risk. Under these circumstances, Scott clearly shares at least some responsibility.

The fundamental problem with this alternative account is that an agent's occurrent awareness of risk can't wholly determine the specific outcomes that she's blameworthy for. This single variable is simply too coarse-grained to account for the subtle relation between an agent's awareness and her culpability. For this reason, I argue that another component is necessary; specifically, one that reflects her dispositions to believe in specific risks. Not only is this multidimensional account extensionally superior, but it also has explanatory power. Part of the reason why Scott isn't blameworthy for his friend's accident is that it would be impossible to recognize this risk from his (*ex ante*) perspective.[234] In other words, his total psychology doesn't contain the necessary grounds for the disposition to believe the specific risk of his conduct. It's this potential for recognition, when paired with an awareness of the general risks of one's actions, which supports blameworthiness for specific outcomes. Without such potential, it's difficult to see how an agent has a fair opportunity to avoid wrongdoing.

3. Justifiability

I turn now to the issue of a particular risk's justifiability. I inevitably breached the topic in previous sections already, but I'm now prepared to offer a fuller account. First, I hope it's uncontroversial that imposing risks are sometimes justified. For example, an ambulance driver speeding to the hospital with someone in critical condition appears justified in her actions, even though speeding imposes risks of serious harms on others. The usual explanation for this

[234] One might reasonably wonder whether Scott would pass the proposed counterfactual test. After all, if asked whether the gun could explode from incorrect loading of the cartridges, might he not answer in the affirmative? This question brings up an important constraint on any such counterfactual test; namely, the agent can't *gain* new information upon being asked about the possibility of the specific risk. For instance, suppose that upon being asked about the possibility of the gun exploding from incorrect loading, Scott *realized* (i.e., forms a belief) that cartridges can be incorrectly loaded in a way that is dangerous. This counterfactual scenario would not correctly model his (*ex ante*) perspective, where this newly formed belief is absent. Thus, in conceiving of the counterfactual test we must (perhaps unrealistically) imagine the agent as approaching the relevant question without learning any new information from the question itself. This contrived scenario is the only accurate test for dispositions to believe that I can imagine.

justifiability is that the risks are outweighed by the *reasons* for imposing them. In this case, the risks of speeding are plausibly outweighed by consideration of the person in critical condition.

If justifiability is essentially about the balance of risks versus reasons, then moral theory should ultimately explain how we ought to make this calculation, given that justifiability concerns the rightness or wrongness of actions. Oberdiek (2017), for instance, offers an interesting contractualist explanation, according to which, "permissible risking is permissible if and only if it is acceptable to each affected person taken individually" (p. 12). I won't attempt to provide such an account. Instead, I'll focus on the significance of an agent's *beliefs* about the justifiability of her actions. In this way, the current concern remains the epistemic dimension of blameworthiness.

3.1 Having set aside the issue of what exactly makes a given risk imposition justified, I first reiterate that not all unjustified risk impositions are blameworthy. In particular, I've argued that awareness of risk is necessary for such blame. The real question, then, is whether blameworthiness requires that agents *believe* themselves to be imposing an unjustified risk. Imagine, for example, that Becca decides to drive home drunk, knowing that this imposes a significant risk of harm on others, while falsely believing that she's justified because it would otherwise be expensive to get home. Supposing that Becca meets all other conditions for awareness of risk, does her false belief regarding the justifiability of her actions excuse her from blame?

I contend that such false beliefs clearly don't excuse agents. Indeed, this might be one issue regarding recklessness where there seems to be near universal agreement. Even Alexander and Ferzan (2009), who hold heterodox views on several issues, maintain that "unlike risk, justification is objective…In general, the actor's mistaken belief that his reason X justifies his

act's risk R is immaterial to his culpability" (p. 59). According to their theory of culpability, actions based on mistaken beliefs about the weight of reasons paradigmatically reflect insufficient concern for others' interests and are thus blameworthy.[235] Yet, one needn't accept their theory of culpability to come to the same conclusion. As long as an agent is aware of the risks of her actions, and the reasons at play, then she plausibly possesses a fair opportunity to avoid wrongdoing.

3.2 Before moving on, there are two related, complicating issues worth discussing. First, although I've argued that false beliefs about the *weight* of reasons don't excuse agents from blame, this form of ignorance regarding justification isn't exhaustive. Consider, for example, an extremely selfish man, Louis, who often fails to recognize that others' interests provide him with reasons. Now, suppose that Louis dangerously cuts someone off in traffic in an attempt to shorten his commute. We can imagine that Louis is so selfish that, in appraising the justifiability of his prospective action, it doesn't even register that the other driver's well-being might give him a reason not to attempt the risky maneuver. In this context, Louis believes that only his interests are relevant to the justifiability of his actions. Given that Louis is ignorant of the very *reasons* that factor into the relevant justifiability calculation, is he also blameworthy?

George Sher (2006) refers to cases like Louis as instances of lack of moral insight or imagination, and notes that agents like Louis usually seem blameworthy. Even though Louis is insensitive to the reason-giving nature of others' interests in certain contexts, he differs from certain psychopaths who are literally *incapable* of recognizing that others' interests could

[235] An interesting sort of case for views like this is one in which the agent correctly assesses the weight of reasons and yet acts in accordance with this assessment on the basis of *other* reasons. For example, imagine that Kant's shopkeeper recognizes that it would be morally wrong to overcharge, but only does it because of the risk that it might ruin his shop's reputation. The question is whether correct assessment of weights plus right action is *sufficient* for praise.

provide reasons.[236] Instead, Louis's deficiency is failing to realize that interests that he acknowledges in other circumstances apply to this one. He's so selfishly focused on his own interests in these situations that he fails to even consider the moral significance of others. In this case, it's difficult to see a morally significant difference between Louis and Becca. In both cases, their ignorance regarding the justifiability of their actions isn't plausibly excusing.

Another issue that this discussion brings up is the relation between factual and moral ignorance. As it stands, the current account of the awareness of risk condition is slightly asymmetric. Specifically, an agent can be excused for a false but reasonable belief regarding the riskiness of her actions, but not for any false beliefs regarding the justifiability of those actions.[237] As Alexander and Ferzan (2009) point out, though, "our perceptions of the strength of reasons, just like out beliefs about matter of fact, including facts that bear on risks, come on us 'unbidden'" (p. 153, fn. 76). Becca doesn't *choose* to misjudge the weight of reasons for and against drunk driving, for instance, any more than someone who misjudges the riskiness of his actions due to a lack of information. What then is the relevant difference between the two kinds of epistemic mistakes?

One kind of explanation for the significance of moral ignorance is that such misjudgment is *constitutive* of a bad or indifferent will (or insufficient concern), whereas mistakes of fact are not. However, whether this is true or not,[238] I argued in chapter two that having a bad will is neither necessary nor sufficient for blameworthiness.[239] Instead, a more plausible explanation

---

[236] Alexander and Ferzan (2009) also acknowledge that "sometimes mistakes regarding justification reveal culpability-negating insanity rather than culpability" (p. 154). Setting aside what constitutes 'insanity', the psychopath I have in mind suffers from this kind of culpability-negating condition, whereas Louis doesn't.

[237] I'm assuming here that reasonable beliefs about risk can nevertheless be false. This makes particular sense given the evidence-relative reasonable person account, where reasonableness is indexed to available evidence.

[238] See, e.g., Wieland (2017b) for an argument that "moral ignorance and lack of good will might come apart" (p. 163).

[239] My considered view is that quality of will is *relevant* to blameworthiness, without being necessary or sufficient. For instance, recklessness grounded in insufficient concern seems *more* blameworthy than recklessness without

appeals to the fair opportunity to avoid wrongdoing. In general, agents like Louis and Becca – who are *purely* morally ignorant[240] – retain this fair opportunity insofar as they possess the necessary reasons-responsive capacities and are relevantly aware of the risks of their actions. In other words, knowledge of the moral significance of one's actions isn't necessary for blameworthiness. Situational factors like social indoctrination might undermine fair opportunity by undermining the exercise of these capacities, but moral ignorance isn't excusing by itself.

A natural follow-up question is *why* fair opportunity requires awareness of risk but not awareness of unjustifiability? At this point, the dialectic runs into explanatorily foundational moral claims. As discussed in chapter two, reasonable expectations often ground fairness considerations. I argue that the asymmetry between moral and non-moral awareness is ultimately based on these expectations. Although it's unreasonable to expect agents to meet moral demands without awareness of certain facts about their situation, such expectations aren't necessarily unreasonable once they meet this requirement. Absent interfering situational factors, a reasons-responsive agent can reasonably be expected to meet demands. If such an agent fails in this regard, then she's deserving of blame.

4. Substantiality

One potential worry about the current account is that it's overly punitive to agents who must make hard choices based on the balance of reasons. For example, imagine that Liz is a researcher considering testing a new drug on animals. The drug could benefit a significant number of people, but the testing could also harm the animals. Now suppose that Liz correctly

insufficient concern. In this way, attributionists are right that quality of will is morally significant, they're just mistaken about its role in moral responsibility.

[240] By *purely* morally ignorant, I mean that their moral ignorance is not a result of factual ignorance. For instance, Louis isn't (factually) ignorant that dangerously cutting someone off in traffic puts them at risk of harm.

(or reasonably) assesses the relevant probabilities, but slightly underestimates the moral weight associated with the animals' well-being. Because of this, she judges the testing to be morally justifiable, when in fact it's barely unjustifiable.[241] Assuming that this represents a hard choice, it might seem harsh to blame Liz for deciding to go ahead with the testing. After all, although she misjudged the weight of reasons, her mistake was relatively minor. When the true balance of reasons is so close, it seems plausible that blameworthiness should be affected.

In order to account for these intuitions, a theory might try to take into account how *badly* an agent misjudges the balance of reasons. For instance, recall that the Model Penal Code definition of acting recklessly involves consciously disregarding a "*substantial* and unjustifiable risk." One way to render the substantiality requirement here is independently of the unjustifiability requirement. On such an account, unjustifiable risks must pass some threshold for level of risk in order to count as reckless. Yet this is implausible. Imposing even miniscule risks can be blameworthy in the absence of justifying reasons. Imagine a combatant shooting at a civilian from a great distance purely out of boredom. Even if there's vanishingly low probability of hitting the civilian, the combatant is obviously culpably reckless.[242]

A better account *combines* the substantiality and unjustifiability requirements so that an action is culpably reckless only if the risk is substantially unjustifiable. On this view, although Becca is reckless for drunk driving, Liz isn't reckless for testing on animals. Indeed, this integrated account is one way of understanding the Model Penal Code (1985) condition on acting recklessly that "the risk must be of such a nature and degree that…its disregard involves a *gross deviation* from the standard of conduct that a law-abiding person would observe in the actor's

[241] For ease of exposition, I'm assuming a broadly consequentialist theory. A more realistic account would include deontological norms that bear on justifiability.
[242] See Alexander and Ferzan (2009) for a similar argument that "even very tiny risk impositions can be culpable if imposed for insufficient or misanthropic reasons…" (p. 27).

situation." (§ 2.02(2)(c), my emphasis). Substituting in the superior reasonable person standard, we might say that Liz isn't blameworthy because her decision to test on animals isn't a gross deviation from how a reasonable person would act in this situation (unlike Becca). Her judgment of the balance of reasons, though wrong, is reasonable. There might be other ways to set the threshold for substantial unjustifiability, but some sort of reasonable person standard seems plausible.

Still better is an integrated account that incorporates a *scalar* dimension into the relation between substantiality, justifiability, and blameworthiness. There are at least two ways this could be implemented. First, justifiability could be rendered as a fully scalar property, whereby blameworthiness is inversely related to the unjustifiability of the risk. Second, justifiability could remain a *binary* property – represented as the ultimate balance of reasons – but blameworthiness could still be inversely related to the substantial unjustifiability of the risk (i.e., how finely balanced the reasons are on each side).[243] This latter option seems like a better representation of the concept of justification, and perhaps has theoretical benefits, but the results are the same. On either account, Becca is very blameworthy, whereas Liz is barely blameworthy. If it still seems counterintuitive that Liz is even minutely blameworthy, though, the threshold and scalar account could be hybridized. On this kind of view, there's some threshold for substantial unjustifiability past which recklessness becomes blameworthy, but the degree of blameworthiness increases as the risk gets further from that threshold.

Among other things, a scalar account seems compatible with the plausible view that blameworthiness is partly a function of *difficulty*. Dana Nelkin (2016), for instance, argues that "there is good reasons to think that difficulty is a factor in determining degrees of

[243] Thanks to Dana Nelkin for alerting me to this second, more plausible conception of justifiability.

blameworthiness and praiseworthiness…" (p. 373). One way of explaining why Liz is less blameworthy than Becca is that Liz's judgment about the justifiability of her actions is much more difficult to get right. In general, it's plausible that moral judgments about the balance of reasons are more difficult to get right as the weight of those reasons approach each other. There may be exceptions to this rule,[244] but it lends support to the more important point that substantial unjustifiability is at least a mitigating factor for blameworthiness.

One important caveat to the above picture involves the issue of *stakes*. In the case of both Liz and Becca, the stakes at hand are rather high because their actions involve the well-being of humans and animals. Under such circumstances, it's plausible that there's greater demand to correctly determine the balance of reasons. Even if it's difficult for Liz to judge whether she ought to test the drug on animals, then, we expect her to make a reasonable effort. If she failed to make such an effort, it seems that we would blame her more than someone who made the same effort in a situation with less at stake (holding fixed the margin of the balance of reasons). Of course, some of this difference in blameworthiness is simply explained by the stakes themselves, as we tend to blame agents for their recklessness more when there's more on the line. But the stakes also appear to have an indirect influence on blameworthiness via the aforementioned effect on the demands to correctly determine the balance of reasons. Unfortunately, I lack the space to fully explore the impact of stakes on these demands, or the nature of these demands more generally. For current purposes, I simply want to note that the issue of substantiality (and

[244] One interesting kind of possible exception are cases where the slight balance of reasons is easy to discern. For example, imagine the choice between feeding a village of 1,000 people versus a village of 1,001. All else being equal, the minor benefit to the village of 1,001 clearly seems preferable. Credit to Dana Nelkin for suggesting this kind of case. Undoubtedly, there will be other kinds of exceptional cases, but there's still a general relation between the balance of reasons and difficulty.

its interaction with justifiability) turns out to be rather complicated. Nevertheless, I believe these complications are both less fundamental and less vexing to an awareness of risk theory.

Conclusion

In this chapter, I filled in the details of the awareness of risk condition that I argue is necessary for blameworthiness. In constructing this account, I used the Model Penal Code definition of recklessness as the structure to build from, relying on relevant work in the philosophy of law addressing recklessness and the notion of risk more broadly. In particular, I compared and contrasted my own view with similar theories from Alexander and Ferzan (2009) and Oberdiek (2017). Ultimately, unlike these other two theories, my own account is grounded in the commitment that blameworthiness entails that the agent has a fair opportunity to avoid wrongdoing.

I agree with Husak (2011) that the concept of awareness requires much more elaboration in the literature, and so I tried to give the most thorough account of the kind of awareness pertinent to the awareness of risk condition. To this end, I argued that in order to be blameworthy for some outcome, the agent must have: (1) an occurrent belief of the general risk of her conduct, and (2) a disposition to believe the specific risk of her conduct. As far as the occurrent belief, the agent must entertain a belief about the riskiness of her actions at the time of wrongdoing, but this doesn't mean she must explicitly entertain that belief. As for the disposition to believe, I drew on Audi (1994) to articulate a notion whereby such dispositions are grounded in other beliefs and perceptions and require some mediating process to form the relevant belief. A possible theoretical check for such a disposition would be a counterfactual test in which the agent is asked whether some specific outcome was possible as a result of her conduct. If she would answer in

the affirmative, then this indicates the presence of the appropriate disposition to believe in the actual situation.

Finally, I argued that as long as an agent is aware of the risks of her actions, and the reasons at play, then she plausibly possesses a fair opportunity to avoid wrongdoing. Because of this, an agent can be blameworthy for imposing an unjustifiable risk that she believes is justified. One addendum to this account is that an agent can also be blameworthy if she fails to recognize the moral significance of her conduct, where this failure isn't due to any general incapacity such as certain versions of psychopathy. Although this account might seem particularly harsh to agents who make mistakes about the balance of reasons in difficult cases, I temper this result by advocating a scalar account that apportions blame based on the substantiality of the unjustifiable risk. This theory seems consistent with the plausible view that blameworthiness is partly a function of difficulty, as close calls about the justifiability of one's actions are usually more difficult to get right.

REMAINING ISSUES

Introduction

In contrast with attributionism and capacitarianism, I maintain that awareness of risk is necessary for blameworthiness. In the previous chapter, I filled in the details of this awareness of risk condition by focusing on each major component of the Model Penal Code definition of recklessness. As mentioned before, my theoretical aim is to develop an account that limits revisionism overall, even if judgments may vary regarding certain cases. In this way, the outlines of my awareness of risk condition are shaped by considered judgments at all levels of moral theorizing in a method resembling *reflective equilibrium*.[245] Now that I have this rather detailed theory in hand, I will use this chapter to address remaining issues regarding the epistemic dimension. Some of these issues involve specific kinds of cases, whereas others involve broader questions regarding moral and legal responsibility. Hopefully, by the end, I have answered the most significant lingering questions about the implications of my theory.

1. Pure Epistemic Recklessness

The first issue that I want to discuss involves a specific kind of case that is relevant to the epistemic dimension. This is a case of *pure* epistemic recklessness leading to harm. In these cases, agents are aware that their actions risk ignorance but unaware that this ignorance risks harm. Now, hopefully it should be clear that my account blames agents who, though unaware at the time of wrongdoing, were previously aware that their actions risk harm.[246] For example, as

[245] For discussion of the method of reflective equilibrium in moral theory, see Daniels (1996), Rawls (1971), and Scanlon (2002).

[246] In this way, my account incorporates a *tracing* or *blame transfer* component. Robichaud and Wieland (2017) discuss a number of possible blame transfer views that delineate the scope and degree of blameworthiness between the omission to inform oneself (B1) and the ignorant wrongdoing and its consequences (B2). I agree with Robichaud and Wieland (2017) that the best view appears to be the following: "The degree of B1 and of B2 determines the overall degree of blameworthiness; the degree of B1 is insensitive to downstream factors; the degree of B2 is sensitive to downstream factors, though it is constrained by the degree of B1" (p. 296).

mentioned earlier, Jill would be blameworthy for putting nuts in the cake if she was earlier aware of the risk of forgetting and failed to take steps to prevent it. I assume that her awareness of the risk of forgetting would also include awareness of the risk of harm that might result from forgetting. At least, this is how I intended the case to be understood. Under this description, the case is just a slightly more complicated version of standard awareness of risk cases. The only difference is that the content of Jill's awareness of risk is more indirect, as she believes that there's a direct risk of forgetting, which might then risk harm. One might say, then, that Jill is being epistemically reckless by culpably mismanaging her beliefs, but that her fault is ultimately *moral* and not just epistemic because she's aware that this mismanagement carries a risk of harm to others.[247]

The more interesting cases of epistemic recklessness are ones in which the agent's awareness of risk is purely directed at her beliefs. For example, suppose that Edgar consumes much of his news media from a charismatic charlatan, Alex, who hawks supplements and other wellness products. Edgar genuinely believes Alex's health claims, but he's also aware that receiving all his advice from a single source risks ignorance. Nevertheless, because he can't bear the thought that Alex could be wrong, he chooses not to investigate the validity of these claims. After all, he reasons, the worst that could happen is he ends up consuming ineffective products at some cost. Therefore, Edgar considers his reckless mismanagement of these beliefs to be morally irrelevant.[248] Now imagine that Edgar ends up taking a supplement touted by Alex that has a record of adverse reactions, which would have prevented Edgar from consuming it had he been

[247] The case of Reckless Ralph in chapter two is also an example of this kind of slightly more complicated version of a standard awareness of risk case. In chapter two I described Ralph as epistemically reckless but contrasted him with instances of *pure* epistemic recklessness. In this section I focus on these pure cases.

[248] In other words, Edgar doesn't believe that he's obligated to investigate the validity of Alex's health claims and so does nothing wrong in omitting to investigate. Although he's incorrect in this instance, agents aren't (morally) obligated to always improve their epistemic situation, and so it's not unreasonable to think that sometimes one needn't scrutinize one's belief.

aware. Instead of merely being ineffective, these supplements make Edgar sick enough that he must be hospitalized, draining precious medical resources during a global pandemic.

Assuming that it's wrong to needlessly drain medical resources during a pandemic, is Edgar excused for his ignorant wrongdoing?[249] Specifically, is he excused if he was aware that receiving all his advice from a single source risks ignorance but *not* aware that this ignorance risks wrongdoing? Ultimately, I'm not sure where intuitions stand regarding cases like this, where there's a disconnect between awareness of risk of ignorance and the resulting risk of wrongdoing. My guess is that people's willingness to blame is proportional to the perceived foreseeability that the relevant ignorance would lead to harm.[250] But what would my proposed account say about such cases of pure epistemic recklessness? On the one hand, Edgar doesn't believe that getting all his health advice from Alex risks any wrongdoing, and so it might seem that he's straightforwardly excused from blame. On the other hand, though, Edgar is aware that his conduct risks ignorance, and this ignorance is inherently dangerous in the circumstances, even if Edgar doesn't recognize the risk. Given that agents can be blameworthy on my account even if they judge that their conduct is justified, then, Edgar might similarly be blameworthy for failing to realize the moral significance of his ignorance.[251]

---

[249] Note that I'm not asking whether Edgar is culpable for his ignorance, although I assume that he is. I don't provide an account of culpable ignorance because it's not clear what epistemic obligations/duties agents are under such that their ignorance could even be wrongful. Ultimately, ignorance is a mental state, not an action, and requires independent inquiry regarding responsibility. My concern is with responsibility for actions.

[250] There's a worry that people's intuitions aren't genuinely registering the stipulation that the agent is ignorant of the moral significance of her ignorance. It might seem implausible that Edgar could actually fail to realize that getting all his health advice from Alex risks any wrongdoing, for instance, and so there could be a tendency to project knowledge of the moral significance of one's actions onto such cases.

[251] Another possibility is that Edgar is blameworthy for needlessly draining medical resources during a pandemic because he's culpable for the ignorance that he acts from. After all, he's aware that relying on Alex for all his health advice risks ignorance, so perhaps he's blameworthy because he refused to improve his epistemic situation. Yet, mere awareness of potential ignorance in the absence of its moral significance can't be enough to necessarily condemn an agent. It's unrealistic and overdemanding to require that agents must constantly try to improve their epistemic situation. Therefore, even if Edgar is culpable for his ignorance, this doesn't necessarily entail that he's blameworthy for his ignorant wrongdoing.

Interestingly, cases of pure epistemic recklessness like this pull apart and highlight the subjective and objective components of my account. The view requires that some elements of an agent's perspective must be considered when assessing blameworthiness, but not all elements. The foundational consideration that unifies these components is whether the agent has a fair opportunity to avoid wrongdoing. In Edgar's case, I argue that he doesn't possess such an opportunity. Although we might criticize him for his epistemic practices, he's not blameworthy for his ignorant wrongdoing. In order to have a fair opportunity, agents must at least be aware that their ignorance risks some sort of wrongdoing. Without such awareness of the potential moral significance of their epistemic situation, they can't reasonably be expected to correct or inform themselves.[252]

2. Negligence

Given that my account of the awareness of risk condition uses the Model Penal Code definition of recklessness as a template, and draws on pertinent work from legal theory, one might naturally wonder about the implications for negligence. After all, negligence is an adjacent class of elemental *mens rea*, and the legitimacy of punishing for criminal negligence is a major topic in recent work in the philosophy of law.[253] What does my account have to say about this debate? Following my criticism of capacitarianism, and subsequent demand for an awareness of risk condition, one might already anticipate the answer: I'm generally skeptical of blaming for negligence. However, this broad claim requires further elaboration and qualification. In the

[252] One might compare Louis and Edgar and wonder why Louis is blameworthy if Edgar isn't. After all, isn't Louis also ignorant of the potential moral significance of his actions? Unlike Edgar, however, Louis is aware that his actions risk *harm* to others – it's just that he doesn't see this fact as reason-giving. In this way, Louis is only *de dicto* and not *de re* morally ignorant; that is, he's aware of his action's wrong-making *features*, even though he doesn't conceive of them as wrong. In contrast, Edgar is also *de re* morally ignorant – he's not even aware of the features that make his actions wrong.

[253] See, e.g., Cowley and Crebs (2020).

following section, I explain more precisely what conception of negligence I reject as culpable. Hopefully this clarification will help assuage worries that might arise from my general commitment that negligence isn't blameworthy.

Because the focus of my project is the epistemic dimension of moral responsibility, I haven't been particularly worried about properly defining recklessness or negligence. Instead of demarcating these categories, my attention was simply on necessary conditions for blameworthiness. Insofar as I reject culpability for negligence, however, everything hinges on the definition of negligence. Unsurprisingly, there are many conceptions of negligence within legal theory and the criminal law, but a particularly influential version is the Model Penal Code's (1985) definition:

> A person acts negligently with respect to a material element of an offense when he should be aware of a substantial and unjustifiable risk that the material element exists or will result from his conduct. The risk must be of such a nature and degree that the actor's failure to perceive it, considering the nature and purpose of his conduct and the circumstances known to him, involves a gross deviation from the standard of care that a reasonable person would observe in the actor's situation. (§ 202(2)(d))

Setting aside the second sentence, which sets a threshold for culpable risk,[254] the definition has two key features: (1) the lack of an awareness condition and (2) the claim that the agent should have been aware. The first feature distinguishes negligence from recklessness, while the second feature distinguishes negligence from *strict liability*.[255]

At first glance, it might appear that the Model Penal Code definition of negligence is rather clear and straightforward – negligent wrongdoing is conduct that the agent wasn't aware

[254] Alexander and Ferzan (2009, p. 43) and others accept this gross deviation condition for culpability. As discussed in the previous section on the substantiality element of awareness of risk, I'm also sympathetic to some version of this condition. However, as I'm generally skeptical of blaming for negligence because of the absence of *awareness* instead, this element of the Model Penal Code definition isn't relevant to the current discussion.

[255] Strict liability is liability without *mens rea*. Unsurprisingly, strict liability statutes in the criminal law are rare and controversial, especially for more serious crimes.

was unjustifiably risky, even though she should have been. But this definition is actually rather ambiguous, especially as it pertains to the distinction between negligence and recklessness. First of all, the *content* of the negligent agent's mental state is under-described. As Michael Moore and Heidi Hurd (2011) point out, it's unclear what an agent must be aware of to be reckless, rather than negligent; must she be aware of "(a) a risk of this type? (b) the substantiality of such a risk? (c) the unjustifiability of taking a risk of this magnitude, including an awareness of both the seriousness of the harm risked and the value of the reason for running the risk?" (p. 153). For their part, Moore and Hurd maintain that awareness of (a) is sufficient for recklessness, but others disagree.[256] Even if this issue was settled, though, the Model Penal Code definition is also vague regarding the *level* of awareness that separates negligence from recklessness. Moore and Hurd (2011) argue that "any scintilla of awareness" (p. 150) isn't enough to cross the threshold from negligence to recklessness, and so there can be *advertent* negligence, but others disagree.[257]

Clearly, then, there's no univocal conception of negligence for my account to address – even the Model Penal Code definition is ambiguous enough to admit of multiple interpretations. Of course, I could always define negligence as wrongdoing that fails to meet my awareness of risk condition and then just reject culpability for negligence under this definition, but it might be more helpful to consider a few cases that clarify the kinds of wrongdoing my account considers blameworthy or excused. Especially pertinent are cases that seemingly fit with common understandings of negligence, but which my account still deems blameworthy. This discussion will hopefully clarify my position regarding negligence for the majority who don't have their own worked out account of the concept.

[256] See, e.g., Husak (2011), who maintains that "the reckless defendant, unlike the negligent defendant, believes that he is creating a substantial and unjustifiable risk" (p. 208). See also Stark (2016, 2020), who categorizes cases of unreasonable underestimation of risk as (possibly culpable) instances of negligence.
[257] See, e.g., Alexander and Ferzen (2009).

One type of case that many people often consider to be culpable negligence involves some sort of *tracing* element. For example, imagine that Martha drives down the highway with a large couch tied to the roof of her SUV. Because she failed to properly secure it, the couch slides off and causes an accident with minor injuries. Now, because Martha couldn't actually see the couch while driving, she had totally forgotten it was on the roof at the time of the accident. Nevertheless, she knew when she shoddily secured it that she would be taking a risk by driving down the highway. Martha's wrongdoing here might seem like a classic case of culpable negligence. Indeed, to the degree that most people have an intuitive notion of negligence, the paradigm is often dangerously clueless agents like Martha, where this intuitive conception fails to take into account whether these agents had prior awareness of risk. On my own account, however, this prior awareness makes Martha (culpably) *reckless* rather than negligent.[258] Thus, although all parties here agree that Martha is blameworthy, this consensus might initially be obscured by different conceptions of negligence.

Another significant kind of case involves *reduced* awareness. Imagine, for example, that Paul decides to race a car that passes him on the highway, leading to an accident with a third car that results in minor injuries for the innocent bystander. Now suppose that Paul never *explicitly* considered the dangers of his conduct, as his attention was focused on the race, but that he was implicitly aware that this racing was generally dangerous and disposed to believe that he could have an accident like the one that materialized. On my account, Paul is blameworthy for his recklessness. Yet, because his awareness of risk was only implicit, I think that some might classify this as a case of culpable negligence. Indeed, this categorization is seemingly

[258] Technically, I never provided a formal account of recklessness in my discussion of awareness of risk. But my theory of the awareness of risk condition can be read as a theory of recklessness insofar as recklessness represents the lowest category of culpability.

encouraged by the Model Penal Code definition of recklessness, which describes recklessness in terms of "conscious disregard." As mentioned previously, this (problematically) suggests that negligence includes any kind of awareness of risk that is less than occurrent and explicit. Regardless of where the bar is set for awareness, though, I agree with those who find agents like Paul blameworthy.

Lastly, there might be confounding cases involving agents' beliefs regarding the *justifiability* of their actions. Recall the case of Louis from chapter three: an extremely selfish man who often fails to recognize others' interests and dangerously cuts someone off in traffic in an attempt to shorten his commute. As mentioned before, Louis might be so selfish that it doesn't even register that the other driver's well-being could give him a reason not to attempt the risky maneuver, and thus he's totally unaware of the unjustifiability of his conduct. For some, this lack of moral insight might read like a case of culpable negligence, whereas on my account Louis would be blameworthy for recklessness instead. Once again, then, I agree with the prevalent culpability intuition, even if I might characterize the wrongdoing differently.

Each of the previous three cases involves a factor that might suggest negligence: (1) tracing, (2) reduced awareness, and (3) lack of awareness of justifiability. Furthermore, all three cases are instances of intuitively blameworthy wrongdoing. The main argument in this section is that my account of the awareness of risk condition agrees with the intuition that these three kinds of agents are blameworthy, even though I'm generally skeptical of culpable negligence. The possible incongruity here is simply the result of different conceptions of the awareness that separates negligence from recklessness. Insofar as recklessness requires the kind of awareness of risk that my account asserts regarding moral responsibility, then these three cases wouldn't count

as negligence on my view. But the more important point is that my theory isn't as revisionary as my general skepticism toward culpable negligence might appear at first.

Finally, it's worth mentioning that even theorists who support blaming for negligence usually narrow their definition of culpable negligence quite significantly.[259] One recent example is Craig Agule's (2022) sophisticated account,[260] which claims that an agent is responsible for negligence when:

> (1) the agent has a minimal working set of executive functions;
> (2) the agent acts or fails to act in some negligent fashion; and
> (3) the agent's executive processes play a role in the agent's action or omission, in the case of negligence by being engaged with (i) background beliefs relevant to the risk involved in their behavior and (ii) perceptions of the features of the situation relevant to the risk produced by that action or omission. (p. 246)

Without delving into all the interesting aspects of Agule's account, note that although he doesn't explicitly require any kind of *awareness* for culpability, he does require a certain kind of "executive engagement" with particular contents. In this way, culpable negligence involves a *positive* mental state, rather than merely the absence of one. Ultimately, then, my general skepticism about negligence isn't even so divergent from some developed theories of negligence that don't share this skepticism.[261]

3. Returning to the Revisionist Argument

In the opening chapter, I discussed a revisionist argument based on the epistemic dimension that implies that agents are almost never morally responsible. As I mentioned then, this argument has shaped much of the literature on the epistemic dimension, as various theories

[259] See Alexander and Ferzan (2009, p. 71) for a similar point. With these narrowed accounts, it's not always clear whether they're limiting *culpable* negligence or negligence more broadly. For current purposes, all that matters is that they limit culpability; if these accounts allow for cases of *non-culpable* negligence, then it's merely a terminological matter whether I also classify them as negligence.
[260] Agule's (2022) account draws from both Hirstein, Sifferd, and Fagan (2018) and Stark (2016).
[261] Another class of views that narrows the definition of culpable negligence asserts some version of a quality of will (or attributionist) condition – see, e.g., Simons (1994) and Tadros (2005)

respond to it in different ways, usually attempting to refute it. Yet, I explained that my own approach to the epistemic dimension wouldn't be guided by this argument for two main reasons: (1) this framing often obscures more than it clarifies, and (2) there are many interesting questions outside of this framework. Despite this shift in focus, though, I pledged to return to the revisionist argument once I elucidated my theory of the epistemic dimension. In this section, I make good on that promise by engaging with the argument, explaining how my theory avoids the revisionist conclusion.

First, recall the basic structure of the argument as Rosen (2003) explains it. He starts off with the seemingly plausible claim that an action done from non-culpable *factual* ignorance is itself non-culpable. The flip side of this claim is that factually ignorant wrongdoing is blameworthy only if (and insofar as) the ignorance is itself culpable. But what makes factual ignorance culpable? According to Rosen, culpability is initially a matter of epistemic irresponsibility, wherein an agent mismanages her epistemic situation. Yet, because such mismanagement is *itself* an action (or omission), it's ultimately blameworthy only if: (1) it wasn't performed under ignorance, or (2) it was performed under culpable ignorance. Hopefully, it's clear that trying to establish (2) begins a regress, as we would need to determine whether *that* ignorance involved blameworthy epistemic irresponsibility. The only way out of the regress is for some ultimate instance of wrongdoing that *wasn't* performed under ignorance.[262]

Once we get to this point in the argument, it becomes clear that we need to know what constitutes ignorance, or the inverse of ignorance: *awareness*. It's here that Rosen makes a particularly bold claim. On his account, the requisite awareness to ground blameworthiness is

[262] To be clear, this is the only way to establish *blameworthiness* within the regress structure. Obviously, another way to terminate the regress is to establish that the relevant ignorance was non-culpable. This would involve ignorance that wasn't the result of any epistemic irresponsibility.

full akrasia; that is, occurrent and conscious awareness that one's actions are all-things-considered wrong. Anything less than this constitutes ignorance. This conception of awareness generates two related results. First, because it requires awareness that one's actions are all-things-considered-wrong, the revisionist argument retroactively applies to both factual *and* moral ignorance. Thus, although Rosen starts with an intuition based on only factual ignorance, he purports to show that the implications are even broader. Secondly, and derivatively, this means that *all* blameworthy wrongdoing must ultimately be grounded in full akrasia. Any factual or moral ignorance not resulting from akrasia is excusing. Insofar as ignorance is common, then, blameworthiness is rare.

In order to make this revisionist argument even clearer, it's worth presenting a more formal account. Others have constructed their own versions,[263] but this is my formalization based on Rosen (2003):

> (P1) Factually ignorant wrongdoing is blameworthy only if (and insofar as) the ignorance is itself culpable.
>
> (P2) Ignorance is culpable only if it results from blameworthy epistemic irresponsibility.
>
> (P3) Epistemic irresponsibility is blameworthy only if: (1) it wasn't performed under ignorance, or (2) it was performed under culpable ignorance.
>
> (P4) Wrongdoing isn't performed under ignorance (or is performed with awareness) if and only if the agent is fully akratic; that is, occurrently and consciously aware that her actions are all-things-considered wrong.
>
> (C) Hence: Wrongdoing is blameworthy only if (and insofar as) the agent is fully akratic.

This formalization might require some support premises to make it valid, but it captures the main claims and structure of Rosen's argument. In order to further draw out it's clear revisionist

[263] See, e.g., Rudy-Hiller (2022) and Wieland (2017).

implications, I could have added a premise that wrongdoing *isn't* usually fully akratic and concluded that wrongdoing is rarely blameworthy.

Given my characterization of the awareness of risk condition, it's hopefully clear that I reject premise four of the above argument. Although Rosen (2003) and other revisionists don't share my taxonomy of the properties of awareness, it's clear that they demand a much stronger form of saliency for blameworthiness. For example, they would excuse Paul – who was only implicitly aware that his street racing was dangerous – whereas I would blame him. Even more starkly, the revisionists require the belief that one's action is all-things-considered wrong, whereas my account rejects excuses based on false beliefs regarding justifiability. In this way, my theory of the awareness of risk condition is asymmetric where the revisionist conception of awareness is symmetric – moral ignorance isn't excusing in the same way as factual ignorance is excusing.

As mentioned in my original discussion of the revisionist argument, rejecting only the awareness premise of the argument is a rather conservative response. Attributionism, for instance, potentially rejects *every* premise of the argument. After all, there's no necessary conceptual connection between expressing a certain evaluative orientation and believing certain things. Different attributionists accounts will reject different premises based on the details of their views,[264] but there's seemingly universal agreement that an agent can be blameworthy despite moral ignorance. Similarly, capacitarianism universally rejects the first premise and

[264] One major point of contention is premise one. Specifically, there's disagreement among attributionists whether factually ignorant wrongdoing can be blameworthy in the absence of culpable ignorance. Based on Smith's (2005, 2007, 2017) assertion that forgetting certain things can express a bad will, it shouldn't be surprising that she rejects premise one. However, attributionists like Harman (2011) seemingly accept premise one and instead argue against other claims.

potentially rejects others.[265] As explained in chapter three, capacitarianism asserts that factual ignorance isn't excusing if the agent could and should have been aware of the relevant fact(s).

Nevertheless, even if I only reject the awareness premise, I don't believe that concedes too much to revisionists. Although accepting the other premises plausibly leads to more excuses for ignorant wrongdoing, it's really premise four that drives broader skepticism about moral responsibility. Just by weakening the standard for both the kind and content of awareness, then, many intuitively blameworthy cases of wrongdoing are captured. If this is right, then I don't think attributionism or capacitarianism gains theoretical advantages by potentially rejecting more of the revisionist argument.

[265] For example, there's disagreement about premise two. Fernando Rudy-Hiller (2017) rejects it because he argues that an agent can be *directly* culpable for her ignorance, given certain capacities. Randolph Clarke (2017), on the other hand, ostensibly accepts it but argues that "substandard" awareness can still ground blameworthiness in the absence of culpable ignorance.

REFERENCES

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

Agule, C. (2022). Minding negligence. *Criminal Law and Philosophy,* 16(2), 231-251.

Alexander, L., Ferzan, K., & Morse, S. (2009). *Crime and culpability: A theory of criminal law.* Cambridge: Cambridge University Press.

Alexander, L. (2014). Hart and punishment for negligence. In C.G. Pulman (Ed.), *Hart on responsibility*. *Philosophers in depth* (pp. 195-205). London: Palgrave Macmillan.

Amaya, S., & Doris, J. (2015). No excuses: Performance mistakes in morality. In J. Clausen & N. Levy (Eds.), *Handbook of neuroethics* (pp. 253–272). Dordrecht: Springer.

Arpaly, N. (2002). *Unprincipled virtue: An inquiry into moral agency*. Oxford: Oxford University Press.

Arpaly, N. & Schroeder, T. (2013). *In praise of desire*. New York: Oxford University Press.

Audi, R. (1994). Dispositional beliefs and dispositions to believe. *Noûs*, 28(4), 419-434.

Björnsson, G. (2017). Explaining away epistemic skepticism about culpability. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility* (Vol. 4, pp. 141–164). Oxford: Oxford University Press.

Brink, D., & Nelkin, D. (2013). Fairness and the architecture of responsibility. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility* (Vol. 1, pp. 284-313). Oxford: Oxford University Press.

Brink, D. (2021). *Fair opportunity and responsibility*. Oxford: Clarendon Press.

Brink, D., & Nelkin, D. (2022). The nature and significance of blame. In J. Doris & M. Vargas (Eds.), *The Oxford handbook of moral psychology* (pp. 177-196). Oxford: Oxford University Press.

Clarke, R. (2014). *Omissions: Responsibility, agency, and metaphysics*. Oxford: Oxford University Press.

Clarke, R. (2017). Ignorance, revision, and commonsense. In P. Robichaud & J. Wieland (Eds.), *Responsibility: The epistemic condition* (pp. 233–251). Oxford: Oxford University Press.

Cowley, C., & Krebs, B. (Eds.). (2020). Special issue on recklessness and negligence. *Criminal Law and Philosophy*, 14(1).

Daniels, N. (1979). Wide reflective equilibrium and theory acceptance in ethics. *Journal of Philosophy*, 76(5), 256–82.

Dennett, D. (1969). *Content and consciousness.* New York: Routledge.

Edwards, J. (2018). Theories of criminal law. In E.N. Zalta (Ed.), *The stanford encyclopedia of philosophy*.

Fara, M. (2008). Masked abilities and compatibilism. *Mind*, 117(468), 843–65.

Fischer, J., & Ravizza, M. (1998). *Responsibility and control: A theory of moral responsibility*. Cambridge: Cambridge University Press.

Fischer, J., & Tognazzini, N. (2009). The truth about tracing. *Noûs*, 43(3), 531-556.

Fischhoff, B. (1975). Hindsight is not equal to foresight: The effect of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance*, 1(3), 288.

FitzPatrick, W. (2017). Unwitting wrongdoing, reasonable expectations, and blameworthiness. In P. Robichaud & J. Wieland (Eds.), *Responsibility: The epistemic condition* (pp. 29-46). Oxford: Oxford University Press.

Frankfurt, H. (1969). Alternate possibilities and moral responsibility. *The Journal of Philosophy*, 66(23), 829–839.

Frankfurt, H. (1971). Freedom of the will and the concept of a person. *The Journal of Philosophy*, 68(1), 5–20

Fricker, M. (2016). What's the point of blame? A paradigm based explanation. *Nous*, *50*(1), 165–183.

Ginet, C. (1966). Might we have no choice? In K. Lehrer (Ed.), *Freedom and determinism* (pp. 87-104). New York: Random House.

Guerrero, A. A. (2007). Don't know, don't kill: Moral ignorance, culpability, and caution. *Philosophical Studies*, 136(1), 59-97.

Guerrero, A. A. (2017). Intellectual difficulty and moral responsibility. In P. Robichaud & J. Wieland (Eds.), *Responsibility: The epistemic condition* (pp. 199–218). Oxford: Oxford University Press.

Haji, I. (1997). An epistemic dimension of blameworthiness. *Philosophy and Phenomenological Research*, 57(3), 523-544.

Hájek, A. (2007). The reference class problem is your problem too. *Synthese*, 156(3), 563-585.

Hansson, S. (2022). Risk. In E.N. Zalta (Ed.), *The stanford encyclopedia of philosophy.*

Harman, G. (1999). Moral philosophy meets social psychology: Virtue ethics and the fundamental attribution error. *Proceedings of the Aristotelian Society*, 99, 315-331.

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

Hart, H. L. A. (1968a). Legal responsibility and excuses. *Punishment and responsibility* (pp. 28–53). Oxford: Clarendon Press.

Hart, H. L. A. (1968b). Negligence, mens rea, and criminal responsibility. *Punishment and responsibility* (pp. 136–157). Oxford: Clarendon Press.

Hartman, R. (2017). *In defense of moral luck: Why luck often affects praiseworthiness and blameworthiness*. Abingdon: Routledge.

Hirstein, W., Sifferd, K.L., & Fagan, T. (2018). *Responsible brains: Neuroscience, law, and human culpability*. Cambridge: MIT Press

Hieronymi, P. (2008). Responsibility for believing. *Synthese*, 161(3), 357-373.

Husak, D. (2011). Negligence, belief, blame and criminal liability: The special case of forgetting. *Criminal Law and Philosophy*, 5(2), 199–218.

Levy, N. (2005). The good, the bad, and the blameworthy. *Journal of Ethics and Social Philosophy*, 1(2), 1-16.

Levy, N. (2009). Culpable ignorance and moral responsibility: A reply to FitzPatrick. *Ethics*, 119(4), 729–741.

Levy, N. (2011). *Hard luck: How luck undermines free will and moral responsibility*. Oxford: Oxford University Press.

McGeer, V. (2013). Civilizing blame. In J. D. Coates and N.A. Tognazzini (Eds.), *Blame: Its Nature and Norms* (pp. 162–188). Oxford: Oxford University Press.

McKenna, M. (2008). Putting the lie on the control condition for moral responsibility. *Philosophical Studies,* 139(1), 29-37.

McKenna, M. (2012). *Conversation and responsibility*. Oxford: Oxford University Press.

McKenna, M. (2013). Reasons-responsiveness, agents, and mechanisms. In D. Shoemaker (Ed.), *Oxford studies in agency and responsibility* (Vol. 4, pp. 151– 184). Oxford: Oxford University Press.

McKenna, M. (2022). Guilt and self-blame within a conversational theory of moral Responsibility. In A. Carlsson (Ed.), *Self-Blame and Moral Responsibility* (pp. 151-174). Cambridge: Cambridge University Press.

McKenna, M. (2022). Reasons-responsiveness, Frankfurt examples, and the free will ability. In D. Nelkin and D. Pereboom (Eds.), *The Oxford Handbook of Moral Responsibility* (pp. 27–52). Oxford: Oxford University Press.

McGrath, M. (2021). Being neutral: Agnosticism, inquiry and the suspension of judgment. *Noûs*, 55(2), 463-484.

Mele, A. (2019). *Manipulated agents: A window to moral responsibility*. New York: Oxford University Press.

Mole, C. (2021). Attention. In E.N. Zalta (Ed.), *The stanford encyclopedia of philosophy*.

Moore, M. (1993). *Act and crime: The philosophy of action and its implications for criminal law.* Oxford: Oxford University Press.

Moore, M. (1997). *Placing blame*. Oxford: Clarendon Press.

Moore, M. (2009). *Causation and responsibility: An essay in law, morals, and metaphysics*. Oxford: Oxford University Press.

Moore, M., & Hurd, H. (2011). Punishing the awkward, the stupid, the selfish, and the weak: The culpability of negligence. *Criminal Law and Philosophy*, 5(2), 147–198.

Morse, S. (1994). Culpability and control. *University of Pennsylvania Law Review*, 142(5), 1587-1660.

Murray, S. (2017). Responsibility and vigilance. *Philosophical Studies*, 174(2), 507–527.

Murray, S., Murray, E. D, Stewart, G., Sinnott-Armstrong, W., & De Brigard, F. (2019). Responsibility for forgetting. *Philosophical Studies*, 176(5), 1177-1201.

Murray, S., & Vargas, M. (2020). Vigilance and control. *Philosophical Studies*, 177(3), 825-843.

Nelkin, D. (2011). *Making sense of freedom and responsibility*. New York: Oxford University Press.

Nelkin, D. (2015a). Psychopaths, incorrigible racists, and the faces of responsibility. *Ethics*, 125(2), 357-390.

Nelkin, D., & Rickless, S. (2015b). Review of R. Clarke's Omissions: Agency, metaphysics, and responsibility. *Notre Dame Philosophical Reviews.*

Nelkin, D. (2016). Difficulty and degrees of moral praiseworthiness and blameworthiness. *Nous*, 50(2), 356–378.

Nelkin, D., & Rickless, S. (2017). Moral responsibility for unwitting omissions: A new tracing view. In D. Nelkin & S. Rickless (Eds.), *The ethics and law of omissions* (pp. 106-130). Oxford: Oxford University Press.

Nelkin, D. (2019). Moral Luck. In E.N. Zalta (Ed.), *The stanford encyclopedia of philosophy*.

Nelkin, D. (2022). How much to blame?: An asymmetry between the norms of self-blame and other-blame. In A. Carlsson (Ed.), *Self-blame and moral responsibility* (pp. 97–116). Cambridge: Cambridge University Press.

Nichols, S., Timmons, M., & Lopez, T. (2014). Using experiments in ethics – ethical conservatism and the psychology of moral luck. In M. Christen, C. van Schaik, J. Fischer, M. Huppenbauer, & C. Tanner (Eds.), *Empirically informed ethics: Morality between facts and norms* (pp. 159–176). Dordrecht: Springer.

Oberdiek, J. (2017). *Imposing risk: A normative framework*. Oxford: Oxford University Press.

Parfit, D. (2011). *On what matters: Two-volume set*. Oxford: Oxford University Press.

Peels, R. (2011). Tracing culpable ignorance. *Logos and Episteme*, 2(4), 575-582.

Peels, R. (2014). What kind of ignorance excuses? Two neglected issues. *Philosophical Quarterly*, 64(256), 478-496.

Pereboom, D. (2001). *Living without free will*. Cambridge: Cambridge University Press.

Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford: Oxford University Press.

Perry, S. (2001). Responsibility for outcomes, risk, and the law of torts. In G.J. Postema (Ed.), *Philosophy and the law of torts* (pp. 72-130). Cambridge: Cambridge University Press.

Rawls, J. (1971). *A theory of justice: Original edition*. Cambridge: Belknap Press.

Raz, J. (2010). Responsibility and the negligence standard. In J. Raz (Ed.), *From normativity to responsibility* (pp. 255–269). Oxford: Oxford University Press.

Reichenbach, H. (1949). *The theory of probability*. Berkeley: University of California Press.

Robichaud, P. & Wieland, J. (Eds.). (2017). *Responsibility: The epistemic condition*. Oxford: Oxford University Press.

Robinson, P. (2003). Prohibited risks and culpable disregard or inattentiveness: Challenge and confusion in the formulation of risk-creation offenses. *Theoretical Inquiries in Law*, 4(1), 1-30.

Rosen, G. (2003). Culpability and ignorance. *Proceedings of the Aristotelian Society*, 103, 61–84.

Rosen, G. (2004). Skepticism about moral responsibility. *Philosophical Perspectives*, 18(1), 295–313.

Ross, L. (1977). The intuitive psychologist and his shortcomings: Distortions in the attribution process. In L. Berkowitz (Ed.), *Advances in Experimental Social Psychology* (pp. 173-220). New York: Academic Press.

Royzman, E., Cassidy, K., & Baron, J. (2003). "I Know, you know": Epistemic egocentrism in children and adults. *Review of General Psychology*, *7*(1), 38–65.

Royzman, E., & Kumar, R. (2004), Is consequential luck morally inconsequential? Empirical psychology and the reassessment of moral luck. *Ratio*, 17, 329-344.

Rudy-Hiller, F. (2017). A capacitarian account of culpable ignorance. *Pacific Philosophical Quarterly*, 98(1), 398–426.

Rudy-Hiller, F. (2019). Give people a break: Slips and moral responsibility. *Philosophical Quarterly*, 69(277), 721-740.

Rudy-Hiller, F. (2020). Reasonable expectations, moral responsibility, and empirical data. *Philosophical Studies*, 177, 2945-2968.

Rudy-Hiller, F. (2022). The epistemic condition. In E.N. Zalta (Ed.), *The stanford encyclopedia of philosophy*.

Sartorio, C. (2018). Situations and responsiveness to reasons. *Noûs* 52(4), 796-807.

Scanlon, T.M. (1998). *What we owe to each other*. Cambridge: Harvard University Press.

Scanlon, T.M. (2002). Rawls on justification. In S. Freeman (Ed.), *The cambridge companion to Rawls* (pp. 139–167). Cambridge: Cambridge University Press.

Schwitzgebel, E. (2019). Belief. In E.N. Zalta (Ed.), *The stanford encyclopedia of philosophy*.

Sher, G. (2006). Out of control. *Ethics*, 116(2), 285-301.

Sher, G. (2009). *Who knew?: Responsibility without awareness*. Oxford: Oxford University Press.

Sher, G. (2017). Blame and moral ignorance. In P. Robichaud & J. Wieland (Eds.), *Responsibility: The epistemic condition* (pp. 101–116). Oxford: Oxford University Press.

Shoemaker, D. (2011). Attributability, answerability, and accountability: Toward a wider theory of moral responsibility. *Ethics*, 122(3), 602–632.

Shoemaker, D. (2013). Qualities of will. *Social Philosophy and Policy*, 30(1–2), 95–120.

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

Smith, A. (2008). Control, responsibility, and moral assessment. *Philosophical Studies*, 138, 367–392.

Smith, A. (2017). Unconscious omissions, reasonable expectations, and responsibility. In D. Nelkin & S. Rickless (Eds.), *The ethics and law of omissions* (pp. 36-60). Oxford: Oxford University Press.

Smith, H. (1983). Culpable ignorance. *Philosophical Review*, 92(4), 543-571.

Smith, M. (2003). Rational capacities, or: How to distinguish recklessness, weakness, and compulsion. In S. Stroud & C. Tappolet (Eds.), *Weakness of will and practical irrationality* (pp. 17-38). Oxford: Clarendon Press.

Sripada, C. (2016). Self-expression: A deep self theory of moral responsibility. *Philosophical Studies*, 173, 1203–1232.

Stark, F. (2016). *Culpable carelessness: Recklessness and negligence in the criminal law*. Cambridge: Cambridge University Press.

Stark, F. (2020). The reasonableness in recklessness. *Criminal Law and Philosophy*, 14(1), 9-29.

Strawson, P.F. (1962/1993). Freedom and resentment. In J. Fischer & M. Ravizza (Eds.), *Perspectives on moral responsibility* (pp. 45-66). Ithaca: Cornell University Press.

Talbert, M. (2008). Blame and responsiveness to moral reasons: Are psychopaths blameworthy? *Pacific Philosophical Quarterly*, 89(4), 516-535.

Talbert, M. (2012). Moral competence, moral blame, and protest. *The Journal of Ethics*, 16(1), 89-109.

Talbert, M. (2017a). Akrasia, awareness, and blameworthiness. In P. Robichaud & J. Wieland (Eds.), *Responsibility: The epistemic condition* (pp. 47–63). Oxford: Oxford University Press.

Talbert, M. (2017b). Omission and attribution error. In D. Nelkin & S. Rickless (Eds.), *The ethics and law of omissions* (pp. 17–35). Oxford: Oxford University Press.

Talbert, M. (2019). Moral responsibility. In E.N. Zalta (Ed.), *The stanford encyclopedia of philosophy*.

Talbert, M., (2022). Attributionist theories of moral responsibility. In D. Nelkin & D. Pereboom (Eds.), *The oxford handbook of moral responsibility* (pp. 53-70). Oxford: Oxford University Press.

Timpe, K. (2011). Tracing and the epistemic condition on moral responsibility, *The Modern Schoolman*, 88(1–2), 5–28.

van Inwagen, P. (1975). The incompatibility of free will and determinism. *Philosophical Studies*, 27, 185– 199.

Vargas, M. (2005). The trouble with tracing. *Midwest Studies in Philosophy*, 29, 269– 291.

Vargas, M. (2013). *Building better beings: A theory of moral responsibility*, Oxford: Oxford University Press

Vargas, M. (2015). Moral responsibility and desert: Social, scaffolded, and revisionist. *Philosophical Studies*, forthcoming.

Vargas, M. (2020). Negligence and social self-governance. In A. Mele (Ed.), *Surrounding self-control* (pp. 400-20). Oxford: Oxford University Press.

Vargas, M. (2022). Instrumentalist theories of moral responsibility. In D. Nelkin and D. Pereboom (Eds.), *The Oxford Handbook of Moral Responsibility* (pp. 3–26). Oxford: Oxford University Press.

Vargas, M. (forthcoming). Revisionism. In J. Campbell and K. M. Mickelson, and V.A. White (Eds.), *A Companion to Free Will*. Hoboken: Wiley-Blackwell.

Vihvelin, K. (2004). Free will demystified: A dispositional account. *Philosophical Topics*, 32, 427–50.

Wallace, J. (1994). *Responsibility and the moral sentiments*. Cambridge, MA: Harvard University Press.

Watson, G. (1975). Free agency. *Journal of Philosophy*, 72(8), 205-20.

Watson, G. (1996). Two faces of responsibility. *Philosophical Topics*, 24, 227–248.

Wolf, S. (1990). *Freedom within reason*. Oxford: Oxford University Press.

Wieland, J. (2017a). Introduction: The epistemic condition. In P. Robichaud & J. Wieland (Eds.), *Responsibility: The epistemic condition* (pp. 1–28). Oxford: Oxford University Press.

Wieland, J. (2017b). What's special about moral ignorance? *Ratio*, 30(2), 149–164.

Wieland, J. & Robichaud, P. (2017). Blame transfer. In P. Robichaud & J. Wieland (Eds.), *Responsibility: The epistemic condition* (pp. 281-298). Oxford: Oxford University Press.

Williams, B. (1981). Moral luck. In B. Williams (Ed.), *Moral luck* (pp. 20–40). Cambridge: Cambridge University Press.

Zimmerman, M. (1997). Moral responsibility and ignorance. *Ethics*, 107(3), 410–426.