Moral Responsibility and Motivating Reasons

On the Epistemic Condition for Moral Blameworthiness

Thomas A. Yates

Doctor of Philosophy
Philosophy
The University of Auckland
February 5th, 2021

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in Philosophy, at the University of Auckland, 2021.

Abstract

In the last decade, there has been a surge of interest in what is epistemically required for an agent to be morally responsible for an action (viz., the "epistemic condition" for moral responsibility). According to a prominent account known as "volitionism," responsibility or blameworthiness for wrongdoing requires that the act, or some act in its causal history, was done in full awareness of its wrongfulness. Since such a view makes blameworthiness hard to come by, volitionists have argued that rarely, if ever, are we justified in pronouncing wrongdoers blameworthy. Not surprisingly, a sizeable literature on the epistemic condition has emerged in response.

This thesis defends a novel account of the epistemic condition which avoids the volitionist's revisionary implications and provides a diagnosis of where volitionism goes wrong. According to my proposal, an agent satisfies the epistemic condition on blameworthiness for wrongdoing if and only if she had right and outweighing motivating reasons to avoid wrongdoing that were explicit or at least consciously accessible (through what I call "deliberative attunement"), and she had these reasons either at the time of the act, or at some earlier time in which she failed to take a precaution against her later wrongdoing. I argue that cases that satisfy this description of the epistemic condition, even if they are not cases involving *fully* advertent wrongdoing, may nevertheless be intuitive cases of blameworthiness. And because there are many more cases of this kind, my account sidesteps volitionism's revisionary implications.

After a few chapters laying the foundation for my discussion and setting out volitionism, I argue that a proper understanding of the nature of blameworthiness and responsibility (as consisting in responses to normative reasons) supports a critical presupposition made by volitionists—the "culpability internalist" view that blameworthiness for an act requires that the agent has beliefs or credences concerning the act's moral significance. Following a defence of this presupposition against those who deny it, I turn to a more detailed discussion of different varieties of culpability internalism, where I argue that volitionism is ultimately too strong a variety of internalism and that my proposal is the way forward.

For Ruby

Acknowledgements

I would like to begin with a huge thank you to my supervisor, John Bishop, who consistently and patiently offered support and invaluable feedback throughout my doctoral journey. I have profited hugely from his philosophical wisdom and his attention to detail in reading through drafts of each chapter at various points along the way, especially in the lead up to my submission. I would also like to thank my co-supervisor, Fred Kroon, who provided precious feedback on portions of my material and who, like John, provided stimulating exchange over lunch and coffee catchups. Staff and former staff in the Philosophy Discipline at the University of Auckland also deserve thanks for many helpful comments and suggestions along the way, including Glen Pettigrove, Andrew Withy, Emily Parke, Raamy Majeed, Jeremy Seligman, Tim Dare, and Timothy Mulgan (whose graduate class on free will and moral responsibility five years ago was the catalyst for my interest in moral responsibility).

For giving me helpful and poignant feedback on conference material presented in New Zealand, Australia, and the United States, I would especially like to thank Elizabeth Harman, Neil Levy, Christine Swanson, Graham Oppy, Carolina Sartorio, Matthew Flannagan, Chloe Wall, Tucker Sigourney, Marshall Thompson, and especially Erik Franklin, who was the commentator on my conference paper, "Blameworthiness, Better Alternatives, and Motivating Reasons," at the Florida State University Free Will, Moral Responsibility, and Agency Conference 2020. I would also like to thank Stephen Setman who I met there and who subsequently offered helpful comments on a section of Chapter Two.

For stimulating discussion, good company, and helpful feedback on ideas presented at graduate seminars at the University of Auckland, I would like to thank my friends and (former) graduate students, Lars Ivarrsson and Mackenzie Groff especially, but also Matteo Ravasio, Conor Leisky, David Kelley, Adam Dalgleish, Sidney Carls-Diamante, Vladimir Kirstic, Sun Liu, and Isabella McAllister. I would also like to thank my friends, Danyon Chong, Daniel Lett, Josiah Bovill, Ellen Long, Andy Mack, and Dominic Spary for engaging conversation over the years, and Cameron Surrey for a stimulating podcast discussion.

But perhaps those to whom I owe the most gratitude are those who provided the constant spiritual, emotional, and material support that kept me going. For that, I thank my loving parents Michael and Michelle Yates, my siblings Reuben (and Rosy), Anna, and Joshua, as well as my in-laws, Susan and Earl Simpson, and Thomas, Olivia, and Harry. I also owe a huge debt of gratitude to the University of Auckland for the Faculty of Arts Masters Thesis Scholarship (2016), the University of Auckland Doctoral Scholarship (2017-2019), and the opportunity to teach as a Graduate Teaching Assistant (and guest lecturer) for the full duration of my study. But most of all, I wish to express my utmost gratitude to the person who was there for me every step of the way, who willingly went through all the "ups and downs" with me, and who never ceased to offer patience, love, and support (not to mention, plenty of examples to think about)—my precious and beautiful wife, Ruby Yates.

Contents

Chapter 1
Introduction

1.1 An Historical Illustration: The Józefów Massacre

On July 13th 1942, Reserve Police Battalion 101 was ordered by the SS to perform an egregious "special action." Their task was to enter the village of Józefów, Poland, haul every young-to-middle-aged Jewish man into trucks heading for concentration camps, and execute the rest. All Jewish women, children, and elderly were to be round up and shot. Tragically, 1500 Jews were killed that day (Browning 1992).[1]

Two peculiar features of this tragic incident merit attention. The first is that many of the men were given the opportunity to step out, right from when the task was announced early that morning until later in the day. Otherwise the men were quite capable of slipping away unnoticed. Initially, their battalion leader, Wilhelm Trap, gave the older men the chance to step out if they "did not feel up to the task" and to "await a further assignment" (p. 57). And some captains were happy to relieve their men throughout the day, with the exception of the dyed-in-the-wool SS member Captain Wohlauf (p. 62). Another peculiar feature is that many members of Battalion 101 had a motive to opt out. They were—as historian, Christopher Browning, describes them—"ordinary men," many of whom family men with wives and children back in Germany, but most importantly, "anti-Nazi."

> [These] were men who had known political standards and moral norms other than those of the Nazis. Most came from Hamburg, by reputation one of the least nazified cities in Germany, and the majority came from a social class that had been anti-Nazi in its political culture. These men would not seem to have been a very promising group from which to recruit mass murderers on behalf of the Nazi vision of a racial utopia free of Jews. (p. 48)

[1] I will rely solely on Christopher Browning's (1992) account of the incident.

And yet the vast majority carried out the action. Only a dozen out of a battalion of five hundred took the initial opportunity to opt out, and no less than eighty percent of the shooters continued to shoot until the end of the day (pp. 73-4).

In part explaining this shocking failure to opt out, many shooters clearly believed that it was right, obligatory, or permissible to carry out the action. Nazi Captain Wohlauf has already been mentioned, but another disturbing case was of a particular policeman from Bremerhaven, who killed children only because it "'[soothed his conscience] to release [Ger. "redeem/save"] children unable to live without their mothers'" (p. 73). One policeman reflects on the battalion's shared moral plight: "'Only years later did any of us become truly conscious of what had happened then… Only later did it first occur to me that [it] had not been right'" (p. 72). And many of the shooters carried on in their campaign to rid Poland of the Jews. As Browning says, "When the time came to kill again, the policemen did not 'go crazy.' Instead they became increasingly efficient and calloused executioners" (p. 77).

Other members of the battalion had objections. One man who initially opted out said that he opposed the action because he was "a great friend of the Jews" (p. 75). Another man, Lieutenant Buchman, apparently could not murder innocent women and children. The shooters who opted out after some executions explained that they could no longer bear it. One policeman, Franz Kastenbaum, reflected that:

> The shooting of the men was so repugnant to me that I missed the fourth man. It was simply no longer possible for me to aim accurately. I suddenly felt nauseous and ran away from the shooting site. (pp. 67-68)

Indeed, the mood in general was grim that day (p. 69). But in later reflections, "the vast majority" of the shooters did not cite "ethical or political principles" against their conduct, rather a kind of "physical revulsion" over what they had to do and see (especially after their comrades' sloppy aiming; p. 74). Browning is perspicuous about their failure to acknowledge anti-Semitism:

> What is clear is that the men's concern for their standing in the eyes of their comrades was not matched by any sense of human ties with their victims. The Jews stood outside their circle of human obligation and responsibility. (p. 73)

The Battalion left the incident full of "resentment and bitterness over what they had to do" (p.

76), but these observations suggest that they were not (consciously at least) resentful or bitter over the *moral* reprehensibility of what they had to do.

1.2 The Epistemic Condition for Moral Blameworthiness & Responsibility Revisionism

It is undeniable that the shooters were *perpetrators of moral evil*. It is also plausible that they were *moral wrongdoers* (although this is arguable).[2] But should we think that the shooters were *morally blameworthy* for the shootings? I ask the question, here, of whether each shooter was worthy of blame, individually, and at least to some degree, for the executions that he performed. This question should strike one as an interesting and difficult question. On first impression, one might feel that the shooters, or many of them, *were* blameworthy. They were given a clear choice about whether or not to participate in the shootings and later in the day they could excuse themselves from the shootings. They were not, for all we know, being controlled by anyone or anything (e.g., an evil demon or a mad scientist) to act, despite some "pressure for conformity" (Browning 1992, 71). And as "ordinary men," it is reasonable to assume that they were generally morally competent or possessing the ability (despite the occasional failure) to tell right from wrong. On some accounts of the conditions under which an individual is blameworthy for an action, the shooters would therefore have been morally blameworthy for the executions.

But according to a prominent view about what knowledge or awareness is required for moral blameworthiness (Zimmerman 1997b, 2008; Rosen 2003, 2004, 2008; Levy 2009, 2011; Ginet 2000), only a handful of the shooters, if any, would have been blameworthy for the executions. On this view, which is sometimes called "volitionism" (Robichaud 2014; Rudy-Hiller 2018),[3] it is not enough that the shooters were merely *able* to tell right from wrong. In order for the shooters to have been morally blameworthy for the executions, they would have had to perform the executions either in *full awareness* of the wrongfulness of the executions at the time, or in spite of *culpable ignorance* of the wrongfulness of the executions. Moreover, the policemen could have been culpably ignorant only if their

[2] As we shall see, some think that wrongdoing requires blameworthiness, but the shooters' blameworthiness is precisely what is in question.

[3] This should be distinguished from other positions called "volitionism." E.g., Neil Levy (2005) uses the term for the wider position that holds that voluntary control, decision, or choice is necessary for responsibility (and blameworthiness).

ignorance was itself the foreseeable upshot of past wrongdoing in full awareness of *its* wrongfulness. Thus, according to volitionism, all culpable or blameworthy[4] wrongdoing is or is traceable to fully "advertent" or "clear-eyed" wrongdoing,[5] and so any given execution at the Józefów Massacre could have been culpable only if it was or was traceable to fully advertent wrongdoing. ("Full awareness" here means having the true occurrent belief that the act was morally wrong all-things-considered, based upon occurrent beliefs in the features of the act which make made it wrong—e.g., its killing of innocents. Beliefs are "occurrent" at some time when they occur as thoughts or feelings to the agent at that time, as opposed to when they are merely "dispositional.") But if blameworthiness requires such full awareness, it does not seem likely that many of the Battalion 101 shooters were blameworthy for the executions. Bring to mind the fact that at least *eighty percent* of the shooters continued to shoot until fifteen hundred were dead. Recall the policeman who reported that the men were not "truly conscious" of what they did until well after it happened—indeed that it took a long time for it to "occur" to him that the massacre had not been "right." The fact that the shooters could step out also suggests that those who continued to shoot did not have moral reservations against doing so. But most importantly, recall that the "vast majority" of shooters (later interrogated) did not cite "ethical or political principles" as reasons to opt out of the shootings. Thus, it is doubtful that the shooters did so while in full awareness of its wrongfulness—which is to say, for the volitionist, that they were not "directly" or "originally" blameworthy for the executions.

Now, the shooters could still have been "indirectly" or "derivatively" blameworthy for their executions if, argues the volitionist, they did so from culpable ignorance of the moral wrongfulness of doing so. Still, it is unlikely that the volitionist would find grounds to blame them for their ignorance, given that their culpable ignorance must have been the foreseeable upshot of another instance of fully advertent wrongdoing for which they were to blame. Could the shooters have knowingly and culpably failed to take an opportunity in the past to prevent becoming ignorant about the wrongfulness of executing the Jews? Given that they had "known political standards and moral norms other than those of the Nazis" and had come from a "social class that had been anti-Nazi in its political culture", it may be reasonable to suppose that some of them faced situations in which they, by their lights, allowed Nazi ideology too much light of day. But it is difficult to have confidence in the claim that, even in

[4] I use "culpable" and "blameworthy" interchangeably in this thesis.
[5] The view is often described as requiring that the fully advertent act is *akratic* (Rosen 2004; Levy 2011), but since I argue (in Chapter Four) that the conditions do not necessarily entail akrasia, I avoid this terminology.

these situations, the men would have occurrently believed that entertaining Nazi thought was all-things-considered morally wrong, and for the reasons that make it wrong, whilst also foreseeing that their doing so could lead to mass murder. Thus, not only is it unlikely for the volitionist that the shooters were directly blameworthy for their murders, it is unlikely that they were *indirectly* blameworthy as well, and so it is unlikely that they were blameworthy at all for them.

The rationale for volitionism is complex, but it is worth noting in this introduction how well the volitionist's principles hold up in cases of "factual ignorance," from which we can reason analogically to the application of those principles to cases of "moral ignorance," such as the Józefów Massacre. By cases of factual ignorance, I mean cases involving ignorance of the wrongfulness of the act due only to ignorance of some *non-moral fact*, in contrast to cases of moral ignorance where the ignorance of the act's wrongfulness is due only to ignorance of some *moral principle*. To take another example from World War II, consider Rik Peels' (2017, 1–2) case of Sir Arthur Coningham who initiated the bombing of enemy ships without knowing that there were carrying 7800 prisoners of war and camp survivors. Coningham was factually ignorant because his failure to recognise the wrongfulness of the bombing was owing to his false beliefs that "there were only SS men and other German soldiers on board" (p. 1). Now volitionism may seem to yield the right verdict about how we should assess whether Coningham was blameworthy. Since Coningham was ignorant of the fact that there were "friendlies" on board, he cannot have been directly blameworthy for the order to bomb the ships. Nevertheless, he could well have been indirectly blameworthy for the order, if he "should have known" that there were friendlies on board—if he knowingly and culpably failed to (say) arrange intelligence on the possibility of friendly fire, or deceived himself into thinking that there were none on board.[6] However, it would probably have been unlikely for Coningham to have performed these "benighting acts"—as Holly Smith (1983) famously called them—given that they would have constituted a clear breach of duty, and so, as with the Józefów Massacre, volitionism yields the verdict that it is unlikely that Coningham was blameworthy for the bombing.

But whatever we might think about the volitionist's verdict about this case of *factual* ignorance, the claim that the *morally* ignorant Battalion 101 shooters were unlikely to be blameworthy strikes many, and myself included, as mistaken. For many writers,

[6] Cf. W. K. Clifford's (1886, 1) classic case of the shipowner, who sends a ship full of emigrants to sea in the "sincere and comfortable conviction that his vessel was thoroughly safe and seaworthy" after deceiving himself of its seaworthiness.

factual ignorance is more likely to excuse than *moral* ignorance (Moody-Adams 1994; E. Harman 2011). Indeed, the volitionist's likely verdict about the Józefów Massacre is symptomatic of a general feature of volitionism—or at least of the way that volitionists have drawn out its implications[7]—that I and many others find problematic. It strongly suggests a radical form of "revisionism" about blameworthiness and responsibility ascriptions, akin to other forms of scepticism about responsibility (cf. G. Strawson 1994). If, as the volitionist says, culpable wrongdoing is necessarily either fully advertent or traceable to fully advertent wrongdoing, then culpable conduct is actually quite hard to come by. Rarely is anyone ever blameworthy (Zimmerman 1997b; Levy 2011). Or else their blameworthiness is nearly impossible to tell (Rosen 2004). But as Fernando Rudy-Hiller (2018) has observed:

> it seems to most philosophers that something has gone awry at this point, for (the assumption seems to be) it can't be the case that the requirements of the [epistemic condition on responsibility] are so stringent that most ordinary wrongdoers, particularly *morally* ignorant ones… fail to meet them and so turn out to be blameless after all.

Revisionism *per se* is not, I think, the problem. The problem is that such a view grates against our very strong intuitions in favour of the commonness and relative accessibility of blameworthiness for wrongdoing. Not surprisingly, then, there has emerged a considerable literature on this problem, with philosophers in characteristic fashion offering objections, encountering rejoinders, sharpening views, and exploring further questions occasioned by the debate concerning blameworthiness' "epistemic condition." Although an apparent majority feel that "something has gone awry," there is, true to form in philosophy, no clear consensus on where volitionism goes wrong.

1.3 My Proposal & the Literature

It is the object of this thesis to contribute to this debate with a novel account of the epistemic condition for moral blameworthiness, an account which winds up avoiding the volitionist's revisionist implications. According to my account, blameworthiness for

[7] Carl Ginet (2000) does not draw out these implications.

wrongdoing need not depend, at the time of acting or in the past, on *fully* advertent wrongdoing but only on *partially* advertent wrongdoing, at least of a certain kind. In particular, I propose that an agent may be *directly* blameworthy for wrongdoing if, at the time of acting, the agent had right and outweighing motivating reasons to refrain from wrongdoing that were explicit or at least consciously accessible in a certain way—through what I call "deliberative attunement" (as I explained later). Otherwise the wrongdoer may be *indirectly* blameworthy if, at an earlier time, they had right and outweighing motivating reasons against that act, or its general type, which were explicit or consciously accessible, and they foresaw this wrongdoing as a risk of performing some "benighting act" against which they also had right and outweighing reasons (that were explicit or consciously accessible). An account of the full meaning of having "right and outweighing motivating reasons to refrain that are explicit or consciously accessible" I will save until later. Suffice to say for now that they are the reasons that the agent sees herself as having which are sensitive to the features that make the act wrong and cumulatively weigh in favour of the avoidance of wrongdoing.

Although my account shares certain similarities with volitionism, a consequence of my account is that more of the Battalion 101 shooters were likely to be blameworthy for the executions. Reflecting once more on this case, I think that this is just as it should be. Consider again that many of them had "motive" (i.e., reasons) to opt out. Many of them had come from an anti-Nazi political culture. Some had been friends with Jews. Lieutenant Buchman was enlightened enough to recognise murder for what it truly is. We might also doubt whether their physical revulsion was not in some way sensitive to the morally heinous nature of what they had to do. Put it this way: it is doubtful that the shooters would have been so embittered and resentful had they been rounding up and executing *pigs* all day. Many (although it is debatable whether *most*—cf. Browning 1992, 72, xviii) recognised that they had a choice. And in light of these considerations among others, there were surely many who, although they did not believe that what they had to do was all-things-considered morally wrong, were nevertheless *morally uncertain* about the nature of their action, and recognised that a more cautious choice would be to avoid participating in the massacre. In these ways, I think that it is plausible to suppose that many had outweighing reasons to opt out of the shootings which were sensitive to the wrongfulness of doing so, and that were at least consciously accessible if they were not occurrent or explicit. And that is only to consider whether they were directly blameworthy for the executions. My account of indirect blameworthiness is

more likely to account for their blameworthiness for previously allowing Nazist ideology too much light of day. That, at any rate, is a snapshot of how my account is supposed to work.

The rationale for this alternative to volitionism is also quite complex. The case for the view, in particular, has a number of interwoven threads. Apart from direct intuitions of blameworthiness and responsibility, I develop and tease out the implications of an account of *blame* for wrongdoing as morally faulting for wrongdoing, an account of *moral responsibility* as the quality of "respondence" to normative reasons, and a theory of *fair moral expectations* on agents to avoid wrongdoing as requiring right and outweighing motivating reasons to avoid it. For a taster, it is worth mentioning the way in which the proposed view accounts for certain "failure-to-notice" cases which generate a strong intuition of blameworthiness but which volitionists struggle to account for. Consider cases in which a child or pet suffers an accident while the parent gets atypically distracted (Sher 2009), someone fails acknowledge a good friend's birthday after uncharacteristically forgetting the date (A. Smith 2005), or a guest suffers an allergic reaction after the baker forgets their allergies (Rudy-Hiller 2017). The "uncharacteristically" qualification is intended to indicate that these are cases such that no *tracing* of blameworthiness to an earlier time can account for the agent's blameworthiness for their wrongdoing, and yet we still get an intuition of blameworthiness. But they are also cases in which the agent does not act contrary to full awareness of wrongdoing. My view accounts for them on the grounds that they have right and outweighing motivating reasons against their omissions that are not *explicit* (because forgotten) but still *consciously accessible* in the relevant way—either at the time of the omission or indeed at earlier time when they were attuned to their situation in the right way, even though they were unconscious of the risks.

How does my account square with the literature? The fact that it requires, for blameworthy wrongdoing, "partial" advertence to that wrongdoing signals it as belonging to a broader family of approaches which require for blameworthiness at least *some* advertence to wrongdoing, or at least some beliefs or credences concerning the act's moral significance.[8] Call these views *culpability internalist* views. They include volitionist views, but also what Rudy-Hiller (2018) calls "weakened internalist"

[8] This qualification is intended to include views which do require awareness of wrongdoing for blameworthiness, but hold that it is *sufficient* for blameworthiness that one acts contrary to one's *false* moral beliefs or credences (see, e.g., Zimmerman 1997a).

views—the family of views to which mine joins and notably include P. Robichaud's (2014) non-decisive reasons view, C. Sartorio's (2017) awareness of moral significance view, A. Guerrero's (2007) moral risk view, D. Nelkin's and S. Rickless' (2017) risking future omissions view, and the dispositional belief-in-wrongdoing view defended by, among others, I. Haji (1997), K. Timpe (2011), D. Husak (2011), and R. Peels (2011). In Chapter Seven and Eight I discuss these views and argue that they are all left wanting in a way that my view is not. The only other form of culpability internalism is found in some versions of what Rudy-Hiller (2018) calls "quality-of-will" views, according to which blameworthiness for conduct is a function of the agent's displaying ill will in their conduct (e.g., Arpaly 2002, 2015; Harman 2011, 2015; A. Smith 2005; H. Smith 2012; Talbert 2013; Littlejohn 2014). According to some of these views (Harman 2011 and Talbert 2013; *contra*, e.g., A. Smith 2005), for the agent to display ill will, they must have some awareness of the morally significant features of the act, the features that they culpably disregard. Given what I will take to constitute the nature of blameworthiness, the internalism that I promote will have more in common with the "weakened internalists" than the quality of will internalists.

Standing in contrast to culpability internalist views are culpability *externalist* views, which do not require any advertence, even partial advertence, to wrongdoing. Those views that Rudy-Hiller (2018) calls "epistemic vice" and "capacitarian" views count among them, along with those quality of will views which do not require advertence to wrongdoing or wrong-making features at all (Smith 2005). J. Montmarquet (1992, 1993, 1999) and W. FitzPatrick's (2008, 2017) epistemic vice accounts hold that blameworthiness for actions is a function of being traceable to their display of epistemic vices (or epistemically vicious attitudes), which, they argue, does not depend on the agent's morally significant beliefs or credences. The "capacitarians," G. Sher (2007), F. Rudy-Hiller (2017), R. Clarke (2014, 2017), and S. Murray (2017), argue that blameworthiness for wrongdoing does not require awareness of wrong-making features of the act but merely the *capacity* for awareness of those features (as well as the "opportunity" to exercise that capacity, and a norm prescribing that awareness). I discuss these externalist views in Chapter Six, where I raise specific issues for each form of externalism and run my accounts of blame, responsibility, and moral expectations against them.

In light of this literature, the guiding question for this thesis is best put as the question of *whether, and if so, in what way, an agent's blameworthiness for their conduct*

(actions or omissions) requires having beliefs or credences concerning the moral significance of their conduct.[9] If blameworthiness does require these beliefs or credences, then culpability internalism is true. If it does not, then externalism is true. "If so, in what way" then asks for a specific version of culpability internalism.

It should be clear by now that this question is strictly about the *epistemic condition* for moral blameworthiness, or at least, about the epistemic condition concerning whether and what "awareness of moral significance" (Rudy-Hiller 2018) is required for blameworthiness. But blameworthiness is often also thought to require awareness of one's *conduct*, awareness of its *consequences*, awareness of *alternatives*, and awareness of *how to meet one's obligations* (Rudy-Hiller 2018; Peels 2014). But since these forms of awareness relate to features that are themselves morally significant, a theory about whether they are required should depend (partially) on a theory of whether and what awareness of the moral significance of one's action is required for blameworthiness. That will at least be the case for my own internalist account of the epistemic condition (where, e.g., I will argue that awareness of alternatives is required in Chapter Three and Seven).

Since the thesis' guiding question is about the epistemic condition, I should note also that the thesis does not directly implicate other conditions on blameworthiness and moral responsibility, such as the control or freedom condition. Of course, the classic literature on moral responsibility revolves around this condition, centred especially on whether some form of determinism or *in*determinism is compatible with the often-proposed freedom necessary for moral responsibility.[10] Consider, for instance, how common it is in philosophy to hear the phrase "free will and moral responsibility." But since wholly opposing views on the epistemic condition may nevertheless agree on what constitutes the freedom or control condition, or see eye-to-eye on a certain form of compatibilism or incompatibilism, I will not wade into the debate about whether determinism or indeterminism is compatible with responsibility-relevant freedom.

[9] See n. 8 for why this is put in terms of "beliefs and credences concerning the moral significance of the act" rather than advertence to, or awareness of, its moral significance.
[10] See N. Levy & M. McKenna (2009) for a good review of this literature.

1.4 So What?

Why does inquiry into the epistemic condition on blameworthiness for conduct matter? I can think of four reasons why it matters. The first is that the issue is interesting in its own right. Just as figuring out the molecular components of a cell is a worthy intellectual pursuit, so is figuring out the epistemic requirements for culpable wrongdoing. Surely there is something "to be known" about the matter, at least relative to the prevailing way we have come to think about morality.[11] The second reason is that inquiry into the epistemic condition for culpable action matters for questions about blameworthiness and moral responsibility more widely—including questions about the epistemic condition on blameworthiness for character traits and attitudes like beliefs, theories of collective blameworthiness, and, of course, questions about the conditions on *praiseworthiness* with respect to each of the categories of things that can merit blame. Relatedly, investigating the epistemic condition for culpable actions has implications (of interest to me) for the conditions under which religious believers are blameworthy or blameless for their doxastic or sub-doxastic "faith ventures," since it appears that faith ventures can sometimes themselves be regarded as "permissible" or "impermissible" (Bishop 2007).[12] A third reason is perhaps the most important: this inquiry matters for *how we should treat others*, both individually and institutionally.[13] If, as I will argue, blameworthiness is a condition on the fairness of certain kinds of adverse treatment (e.g., behaviour modification, verbal condemnation, sanctions, or legal punishment), then whether or not these forms of treatment are *fair* or *deserved* depends on how we should characterise the epistemic condition. Notably, the more stringent the epistemic condition, the less likely are people going to be the fair objects of this treatment. This is related to a fourth and final reason why this inquiry matters. It matters for what we should say to certain issues being discussed in criminal law—the issues, for example, of whether or not *negligence* is a justifiable mode of criminal liability (Alexander and Ferzan 2009), of how to determine the *mens rea* in cases of *forgetting* that one has created a substantial and unjustifiable risk (Husak 2011), of what mental state should be

[11] I will comment on meta-ethical issues in Chapter 3: in general my approach in this thesis is to remain as neutral as possible on meta-ethical debates, at least consistently with a position that holds that there is such a thing as "moral truth" which can be reasoned about.

[12] As far as I know, only one other theorist about the epistemic condition is thinking about this application. E. Harman has a paper in progress entitled, "Does False Religious Belief Exculpate?" http://www.princeton.edu/~eharman/.

[13] See this point with regard to responsibility for belief and our social institutions in R. Van Woudenberg (2009).

enough to acquit sex offenders (Nottelmann 2007, 2–16), of how to determine commanders' criminal liability for their subordinate's war-crimes (Nerlich 2007; Robinson 2017), of how to redefine and rehabilitate the insanity defence (Vinocour 2020), and of the conditions under which ignorance of the *law* exculpates (Husak 2016). Indeed, Douglas Husak (2016) notes, to his surprise, that while:

> the past several years have witnessed an explosion of interest among moral philosophers about the foundations of blameworthiness… few parallel treatments have been undertaken by philosophers of criminal law. (p. 146)

On the assumption (I share) that "criminal law should conform to morality presumptively" (p. 146), one gets the sense that work on the epistemic condition will become increasingly significant for the theory and practice of criminal law. There are thus good reasons to be thinking about the epistemic condition.[14]

1.5 Thesis Methodology

Before I turn to setting out the thesis' structure, it would be worth motivating the thesis' methodology. My approach is a strong *theory-first* approach to the epistemic condition, for reasons that will become clear. A section on the methodology employed will also be of use in explaining the structure of the thesis. Like most approaches in contemporary analytic ethics and metaethics, my project will attend to "considered judgments" or intuitions that I have about particular "cases" or "thought-experiments" which are relevant to the matter at hand.[15] Only those intuitions in which I have considerable degree of confidence—or which I deem to be "plausible"—shall be accepted for further reflection and refinement. Based on these intuitions, or in order to account for them, I will then adopt "principles," or groups of principles ("accounts" or "theories") which are intended to "fit" or "accommodate" as many of these intuitions as possible. These principles will then be compared to alternative principles in terms of *their* fit with these,

[14] There may be other benefits of inquiry into the epistemic condition. E.g., in the philosophy of religion, inquiry into the epistemic condition might have implications for thinking about the conditions under which it is fair for God to deny some people eternal life, given certain theological assumptions about the nature of God.
[15] I treat "considered judgments" and "intuitions" as equivalent and prefer the term "intuition." N. Daniels (2016) notes that this is sometimes done. As such, I should not be interpreted as presupposing *intuitionism*.

and/or others', intuitions, and I will hold that it *counts strongly in favour* of the principle if it best accounts for all of the relevant intuitions.

"Counting strongly in favour of the principle" does not mean "demonstrates it decisively." Best fit with one's intuitions over a range of relevant cases is not a *decisive* factor in establishing the principle, for such a "narrow reflective equilibrium" (Daniels 1979) between theories and intuitions overlooks the possibility that the principle which best fits the relevant range of cases may not cohere with, or gain support from, relevant "background" theories which may themselves have equal (or even greater) intuitive support. These background theories may be found, for example, in inquiries about other matters in ethics, metaethics, metaphysics, epistemology, logic, or even inquiries strictly outside of philosophy, such as in the humanities or empirical sciences. Support for any given principle not only comes from intuitions about the relevant cases, but from the other direction, as it were—from further theory (even if intuitions, of another kind, still play a large part in justifying these background theories).

The above holds true at least on the method that I adopt this thesis. Since the method widens the "reflective equilibrium" to encompass not just considered judgments (intuitions) and ethical principles (theories) but *background theories*, it has been called the method of *wide reflective equilibrium* (Daniels 1979), following John Rawls' famous use of it in *A Theory of Justice* (1999 [1971], 18-19). According to this method, one achieves the desired end-state of wide reflective equilibrium when one works "back and forth" between intuitions, theories, and background theories, tweaking them along the way so as to achieve *maximal coherence* across all three "levels" (not just logical consistency, but consistency-plus-mutual-support). At any point, any one of these beliefs may be rejected if they fail to cohere with beliefs at another level or if they cannot be given much credence.

The method of wide reflective equilibrium lends itself to an investigation of the epistemic condition on blameworthiness for wrongdoing.[16] The three-tiered structure of beliefs involved in a wide reflective equilibrium can easily be mapped onto the debate about the epistemic condition for moral responsibility. *Intuitions* about relevant cases can be found in particular judgments about the blameworthiness of wrongdoers in cases like the Józefów Massacre. *Principles* or *accounts* which attempt to explain these

[16] This methodology has also been brought to bear on theorising about moral responsibility in J. M. Fischer and M. Ravizza (1998, 10–11).

intuitions can be found in the spectrum of competing views between culpability internalism and externalism. And *background theories* can be found in appeals to theories about the other conditions on moral blameworthiness, the concept of moral responsibility, the nature of blame, right and wrong, reasons, the nature of action, and more.

In fact, I suggest that the method of wide reflective equilibrium is illuminating in a special way given certain observations that have been made about current issues in the debate surrounding volitionism's revisionism about blameworthiness. The need for an appeal to the support of background theories for individual accounts is felt acutely, I think, in the debate between quality of will theorists, capacitarians, and weakened internalists—in particular, the need to appeal to theories about the nature and conditions of blame, blameworthiness, and moral responsibility. There is something right about Rudy-Hiller's diagnosis concerning quality of will theorists in particular, that

> the disagreement between quality-of-will theorists and the rest concerning the [epistemic condition] is ultimately grounded on different conceptions of responsibility and blameworthiness. (2018; cf. also Wieland 2017, 6)

We will also see that there is disagreement about how the epistemic condition factors into the (control) condition on blameworthiness concerning what the agent "can" reasonably do or think at the time of acting. What this suggests is the need for accounts of the nature of blame, the concept of responsibility, and the key conditions of responsibility (including how the epistemic condition is related to the other conditions), before saying anything about the epistemic condition in particular. The fact that philosophers see things differently in these cases given acceptance of certain background theories about blameworthiness, suggests that we start first with background theorising. Of course, we should remain open to letting our intuitions about responsibility for unwitting conduct shape these background theories (recall the method's requirement that each belief should at least be hypothetically revisable). But given the observations we can already make about the debate, I think that the priority should be to get our background theories straight before "applying" them to the debate about the epistemic condition.

1.6 Thesis Structure

Three aforementioned factors shape the structure of the thesis: (i) the responsibility revisionism promoted by volitionists; (ii) the thesis' guiding question; and (iii) the thesis' theory-first methodology.

Given that volitionism is itself an answer to the thesis' guiding question, it would be beneficial to get clear on the components of this question first in order to discuss volitionism more fully, and then my proposed alternative. I devote Chapter Two and Three to that. Chapter Two explains what is meant by "the agent" as well as the agent's "moral blameworthiness"—and "moral responsibility"—for their conduct. Chapter Three explains what is meant by the agent's "actions or omissions," the actions' or omissions' "moral significance", and the agent's "beliefs or credences" concerning their conduct's moral significance. Not only will I discuss what I (and others in the literature) *mean* by these concepts, I will motivate a way of understanding these concepts in such a way that opens the door to volitionism and its responsibility revisionist implications. For instance, I will argue that moral responsibility is best construed as *retrospective* rather than prospective; and that the relevant form of moral significance for culpable conduct is *wrongdoing all-things-considered*. Indeed, the thesis' guiding question will be narrowed to the question of whether, and if so in what way, an agent's moral blameworthiness for her *all-things-considered wrongdoing* depends on the individual's beliefs and credences concerning the *wrongfulness* of her conduct, where that is *either its wrongful status or wrong-making features*.

Chapter Four will then set out volitionism and its revisionist implications and motivate both by paying close attention to Zimmerman and Rosen's arguments. Some arguments due to Levy will feature in this Chapter, but his more direct arguments for volitionism are best set out in response to W. FitzPatrick's reply to Rosen which I discuss later (in Chapter Six).

The following Chapter (Chapter Five) is then the *first* Chapter explicitly devoted to setting out my answer to the thesis' guiding question and my case against volitionism. In keeping with my thesis' theory-first methodology, Chapter Five sets out what I think constitutes blame, blameworthiness, and moral responsibility—presupposing certain claims (e.g., the claim that responsibility for wrongdoing is materially equivalent with blameworthiness for wrongdoing) defended in Chapter Two and Three. In Chapter Five, I argue that blameworthiness for wrongdoing is the quality of being morally at fault for

it, of which I give an account in terms of the agent's playing a morally objectionable role in the causal history of the act. To play this role, I argue, is to be morally responsible for it, since moral responsibility for wrongdoing consists in the action or omission's being agent's response to the wrong-making (normative) reasons against it. (Here I align myself with "reasons-responsive" approaches in the moral responsibility literature.) The key upshot of Chapter Six is that a key presupposition of volitionism—as we have seen, culpability internalism—follows from the nature of blameworthiness and moral responsibility.

Chapter Six then defends this presupposition against culpability externalists of the epistemic vice, quality of will, and capacitarian varieties. I also raise novel arguments against these views and develop an account of the conditions under which it is fair to expect someone to avoid wrongdoing as conditions which must also be satisfied for the agent to be blameworthy. By the end of Chapter Five and Six, my assumption of culpability internalism is firmly seated. We have an answer—in the *affirmative*—of whether blameworthiness for conduct depends on beliefs and credences in the moral significance of that conduct.

I then turn to discuss culpability internalist alternatives in Chapter Seven and Eight, out of which emerges my account of the epistemic condition for moral blameworthiness. In Chapter Seven I propose an epistemic condition under which an agent is *directly* blameworthy for wrongdoing, shaped by my background theories as well as by responses to Levy, Robichaud, Sartorio, Guerrero, and the dispositional belief theorists. In Chapter Eight, I then propose an epistemic condition under which an agent is *indirectly* blameworthy for wrongdoing, shaped primarily by a background theory of indirect responsibility and blameworthiness (set out in Chapter Five), and by interaction with certain theories of the epistemic condition on indirect responsibility, due especially to Zimmerman, Timpe, Ginet, and Nelkin and Rickless (among others). By the end of these two Chapters I will have an answer to the question of *how* the agent's blameworthiness for their conduct depends on beliefs and credences concerning its moral significance—the answer summarised above.

The thesis ends with a concluding Chapter in which I summarise my overall answer to the thesis' guiding question and reply to volitionism and its revisionist implications.

Chapter 2
The Rudiments Part I:
Agents and Moral Responsibility

2.1 Introduction

First things first: (1) what are the rudimentary concepts and distinctions employed in this project? And (2) why should they be understood in a way that allows for volitionism's responsibility revisionist implications? Recall that volitionism is the view that responsibility or blameworthiness for any conduct is, or is traceable to, responsibility for acting (or omitting) contrary to the occurrent belief that one's conduct is all-things-considered wrong. And recall that the revisionism arises on account of the fact that it is hard to come by this type of conduct, at least in cases where the responsibility must trace back to some earlier instance of responsibility. The goal of this and next Chapter is to answer questions (1) and (2) in order to clear the ground for the ensuing discussion of volitionism and its responsibility revisionist implications.

What are the rudimentary concepts—"the rudiments"—of this thesis? Let us consult the guiding question of this thesis, that is, whether, and if so, in what way, the agent's moral blameworthiness for her conduct depends on her beliefs or credences about the moral significance of her conduct. Involved in this question are at least five concepts (or sets of related concepts) that need to be unpacked: (i) the *agent*, (ii) the agent's *moral blameworthiness (and responsibility)* for her conduct, (iii) the agent's *conduct* (actions or omissions); (iv) the conduct's *moral significance*; and finally, (v) the agent's *beliefs and credences* about its moral significance. Each set of rudiments can be combined in different ways to produce different "rhythms" or "beats," or (to drop the drumming metaphor) views that answer our guiding question in different ways. In this Chapter I will discuss the agent and the agent's moral blameworthiness and responsibility for her conduct. In the next, I will discuss actions and omissions, their relevant moral significance, and their significance relative to the agent's beliefs and credences. The outline for this Chapter is simple. §2.2 briefly discusses the agent, and §2.3 is discusses moral responsibility and blameworthiness.

An important task for the next two Chapters is to unpack ("learn") the rudimentary concepts. But before we begin, I should make the caveat that there are ways of analysing the relevant concepts which result in taking sides in the debate about the epistemic condition for moral responsibility. Since, at this point, the intention is not to take any sides in the debate, I will simply signal when I touch on a matter of contention when unpacking these ideas.

At the same time, in order for volitionism to get off the ground, certain controversial views on basic questions about the concepts at issue must be embraced. It is a further part of my job in this and next Chapter to motivate those views.

2.2 The Agent

The key object of evaluation for this thesis is the *agent*. In particular, my scope is restricted to the responsibility of the *individual* agent, as opposed to the group agent. By the individual agent, I mean, of course, the person. No particular theory of agents or persons is assumed, except perhaps for the following general features. The persons must have *intentional agency*, or the ability to act for reasons, or on the basis of mental states like beliefs and desires. They must also have *moral* agency, or the ability to act for *moral* reasons. Typically, certain kinds of persons are assumed as well. The persons are typically *human* persons. The persons are typically also *adult* human persons. Finally, persons with "moral competence" are usually assumed, where moral competence is the general ability to respond to moral considerations that count for or against some conduct. We will have occasion to consider those *without* moral competence (e.g., psycopaths) in Chapter Five, but generally when "agent" is left unqualified in this thesis, the agent will be presumed to have moral competence.

2.3 Moral Responsibility & Blameworthiness

The primary evaluative concepts in this thesis are the concepts of *moral responsibility* and *moral blameworthiness*. We need to raise the following two questions: What is it for an agent to be responsible or blameworthy for something? And what are the conditions under which these concepts apply?[17] The concepts of moral responsibility and blameworthiness are, as we

[17] According to A. Eshleman (2014), a "comprehensive" theory of responsibility must answer both.

shall see, closely related. But while we might be able to discuss responsibility without discussing blameworthiness, I think (and defend further below) that we cannot discuss blameworthiness without discussing responsibility, and so I will set up my discussion with a focus on the primary concept of responsibility.

Unfortunately, precise answers to these questions about responsibility and blameworthiness shall have to be saved until Chapter Five, but what we need here is the "gist" of an answer to them. We need, first, to rule out any obviously distinct concepts picked out by the words, "responsible," or "responsibility," or "blame" and "to blame" (§2.3.1). We need to specify the relationship between responsibility and blameworthiness (§2.3.2). We need also to rule out a type of approach to the concept and conditions of moral responsibility and blameworthiness that would prevent the debate surrounding volitionism from getting off the ground in the first place (§2.3.3). Then I will then briefly outline the kinds of conditions that people have taken responsibility and blameworthiness to have, and I will defend the strong link between responsibility and blameworthiness (§2.3.4). Finally, I will close by introducing some other dimensions of responsibility—especially the distinction between direct/original and indirect/derivative responsibility (§2.3.5).

2.3.1 Narrowing Down Responsibility

I begin with the task of distinguishing moral responsibility from other "responsibility" concepts. There are at least *six* responsibility concepts that are not to be confused with the concept at issue. These are: *causal* responsibility, the *virtue* of responsibility, responsibility as a *role or duty*, the *capacity* of responsibility, responsibility as *legal liability*, and (I will argue) responsibility as *accountability* (for these concepts, see, e.g., Zimmerman 1988; Corlett 2008; Anton 2015, 2ff.).

Causal responsibility, to begin with, signals a *causal* relation which can hold between impersonal objects, events, or states of affairs. Given that moral responsibility is strictly a property of *moral agents*, moral responsibility is clearly not equivalent to causal responsibility. The concepts are related, in that a person's being the *cause* of something may be partial grounds for their moral responsibility for it. However, the following observation from Susan Wolf seems to put idea that moral responsibility is a species of causal responsibility to rest:

> [W]hen we hold an agent morally responsible for some event, we are doing more than identifying her particularly crucial role in the causal series… [We] are judging the moral quality of the individual herself in some focused, noninstrumental, and seemingly more serious way. (1990, 41)

Precisely what this judgment consists in is the focus of the debate about the concept of responsibility (to which I turn in a moment and again in Chapter Five). Second, the *virtue* concept of responsibility is distinct from moral responsibility because the former signifies a trait of character (i.e., dutifulness or conscientiousness; cf. Code 1987), or an aretaic evaluation of conduct based on that trait—rather than a merited assessment of the agent *for* performing their conduct. Third, moral responsibility is distinct from the *role* or *obligation* concepts of responsibility, for two reasons. One can be morally responsible for an action without having (had) a role or obligation to do it (e.g., I am morally responsible for donating all of my money to charity). Moreover, one can be "role-responsible" or have "a" responsibility to do something without being morally responsible for it (e.g., my job and duty as pilot is to land this plane in New York even though I am being held hostage and forced to fly elsewhere). Fourth, the concept of responsibility as the *capacity* of moral competence (of being able to respond to moral reasons) is not responsibility *for* anything at all; rather, it is as A. Anton (2015, 3) says, "a baseline or eligibility requirement for all other types of… responsibility." Basic capacities comprising moral competence, such as the general ability to act intentionally, or to appreciate and respond to moral reasons, may still be present in cases where one is not responsible for some conduct. Fifth, responsibility as legal liability is different from the relevant concept, for one can be morally responsible for conduct without being *legally* liable for it (e.g., telling a white lie), and one can be legally liable for something without being morally responsible for it (e.g., strictly liable for a parking offence, even though one did not know nor could reasonably have known about it).

The sixth responsibility concept that I think is irreducible to moral responsibility is "moral accountability." This is the concept of being worthy of praise, blame, or some neutral attitude, which we might call "neutral appraisal" (Peels 2017, 17).[18] Now many take

[18] It is called accountability om, e.g., Watson (1996), Eshleman (2014), Talbert (2019). Some call it simply "appraisability" (Zimmerman 1988, chap. 3). However, I take appraisability *simpliciter* to be the quality of being appraised for something, which may amount to mere moral criticism (see Chapter Five). I also adopt the terminology of "neutral appraisal" from Peels (2017) for the "indifferent" reaction to someone who is morally responsible for something neutrally valenced, neither warranting praise nor blame (e.g., knitting) (cf. Zimmerman 1988, 61–62).

responsibility and accountability to be the same, so why not follow suit? I certainly take it as one of the great insights of Peter F. Strawson's (1993 [1962]) essay, "Freedom and Resentment," that responsibility has more to do with our interpersonal responses to one another than we had once supposed. We need, in other words, to pay more attention to the practices and attitudes involved in *holding one another responsible* in order to formulate a good theory of responsibility.[19] Consider, for example, that "holding someone responsible" usually has the connotation of *blaming* them. In honour of these observations, some have *defined* the concept of moral responsibility as the concept of moral accountability (see, e.g., Strawson 1993 [1962]; Fischer and Ravizza 1993, 1998; Wallace 1996; Wolf 1990; Shoemaker 2011, 2015; McKenna 2012; Fritz 2014; D. Nelkin 2011; Husak 2016; Peels 2017; Rosen 2004; Levy 2011). To be responsible for your conduct *just is* to be susceptible to a range of attitudes and reactions in virtue of your conduct, and paradigmatically, Strawson's "reactive attitudes" (viz., resentment, indignation, and guilt). Although praise and blame are sometimes either omitted (Shoemaker 2015) or identified as particular items from this list of attitudes (M. Fischer 2006, 63), let me characterise the view that we are considering as that:

> **R=A:** An agent's moral responsibility for *x consists in* their being blameworthy, praiseworthy, or neutrally appraisable for *x*.[20]

R=A certainly captures the importance of our interpersonal reactions. On the view that I will defend in Chapter Five, however, moral responsibility consists in whatever it is that *grounds* this accountability; responsibility is not, in other words, to be *identified* with accountability.

But even though I deny R=A, I follow many others who reject R=A (notably Zimmerman 2015) who believe that moral responsibility is still *materially equivalent* with moral accountability.[21] In other words, moral responsibility still entails, and is entailed by, moral accountability. There is, after all, an important difference between something *x*'s consisting in something else *y*, and *x*'s being materially equivalent with *y* such that *x* is necessary and sufficient for *y*, and vice-versa. Having a heart is materially equivalent with having a kidney,

[19] With other theorists (e.g., Talbert 2019), I am using "holding responsible" as a broad "catch-all" term for any responsibility response.
[20] The idea that praise and blame are the paradigmatic responsibility responses stretches as far back as to Aristotle himself, in his discussion of praise, blame and responsibility in *Nicomachean Ethics* (2013, III.1-5).
[21] It is common to characterise moral responsibility neutrally as materially equivalent with accountability (cf. Eshleman 2014; Talbert 2019)

but clearly neither is reducible, nor to be identified with, the other. Thus, I would embrace the following "iffy" analysis:

> **R↔A**: An agent is morally responsible for *x if and only if* they are either blameworthy, praiseworthy, or neutrally appraisable for *x*.

Thus, although accountability is not identical to responsibility, I think that it is still materially equivalent with responsibility, and this seems to preserve the importance of our interpersonal reactions to one another, without going as far as R=A. In fact, it provides us with conditions under which it is proper to apply the concept of responsibility. And as such, accountability can help us distinguish the relevant concept from the other responsibility concepts mentioned above (in a way that would explain, e.g., why the obligation sense of responsibility is different than moral responsibility). I will have more to say about R↔A or at least what I think it entails about the relationship between responsibility and blameworthiness below and especially in Chapter Five.

The challenge for the critic of R=A is to spell out exactly that in which responsibility consists. Critics of R=A divide between theories of the nature or concept of responsibility as *attributability* and theories of the nature of responsibility as *answerability* (Eshleman 2014). Attributability theories come in a variety of forms, but all insist on the claim that moral responsibility for something *x* consists in *x*'s being attributable to the agent, or some important aspect of the agent, in some crucial way. One clear example is the so-called "ledger view," according to which someone's responsibility for *x* consists in *x*'s being part of that person's "moral record" or "ledger of life" (see, e.g., Zimmerman 1988). If you are responsible for something good, you get a "tick" on your moral record; however, if you do something bad, you get a "blotch" on your moral record. Another class of attributability views are those that Susan Wolf (1990) calls "real-self" views, on which moral responsibility for *x* consists in *x*'s being attributable to the "deep" or "real" self, wherever that is identified (e.g., in "second-order volitions," Frankfurt 1971; or in one's "evaluative stance," Watson 1975). Note that attributability theories are able to distinguish moral responsibility from the other responsibility concepts, and they can explain why we might also accept R↔A: when something is attributable to the agent, someone is accountable for it, and vice versa (although the latter depends on the former). In Chapter Five, I will defend a new kind of attributability theory, bearing most resemblance to existing so-called "reasons-responsive" approaches to the *conditions* of responsibility (Talbert 2019).

Answerability theories hold that moral responsibility consists (roughly) in being appropriate or fitting for the agent to give an answer, account, or explanation for why *x* occurred (Oshana 1997; Scanlon 1998, ch. 6; A. Smith 2012). According to M. Oshana who made one of the first statements of this view, "when we say a person is morally responsible for something, we are essentially saying that the person did or caused some act (or exhibits some trait of character) for which it is fitting that she give an account" (1997, 18). Like attributability theories, answerability theories also have the resources to explain why moral responsibility is distinct from the other responsibility concepts, as well as to explain why R↔A may be true.

Another option is, of course, to embrace pluralism about moral responsibility, so that there is more than one "real" form of moral responsibility, apart from the responsibility concepts distinguished above (Watson 1996; Mason 2015). Such a position is usually motivated by the apparent intractability of the debate between accountability theorists, attributability theorists, and answerability theorists. Gary Watson, for instance, distinguishes *two* moral responsibility concepts, and identifies them with attributability and accountability in his important (1996) paper "Two Faces of Responsibility."[22] Going forward, I will presume monism about responsibility, and only consider pluralism about responsibility if I encounter intractability of the kind that motivates responsibility pluralists such as Watson and E. Mason.

2.3.2 Accountability & Blameworthiness

To be accountable for something is to be worthy of either praise, blame, or neutral appraisal for something. Praise is *positive* response, blame is a *negative* response to the agent, and neutral appraisal is a neutral response. Correspondingly, but very roughly, praiseworthiness for conduct requires that the conduct has overall *positive* moral significance, blameworthiness for conduct requires that it has overall *negative* moral significance, and neutral appraisability for conduct requires that it has overall *neutral* moral significance. In the next Chapter (§3.3), I will argue that the overall negative moral significance that an act must have to be blameworthy (viz., the mode of accountability with which we are concerned in this thesis) is *wrongfulness all-things-considered.* (And I will remain agnostic about what praiseworthy, and therefore neutrally appraisable, conduct requires.) With these claims, and given R↔A,

[22] Shoemaker (2015) is often also included in their ranks, but in my mind, he is a pluralist about *responsibility as accountability*, even though he confusingly distinguishes his three faces of responsibility as "attributability," "accountability," and "answerability."

we can say something further about the relationship between responsibility and these three modes of accountability. In particular, I propose that the only difference between the three modes of accountability for a certain action is a difference concerning the *moral significance* of that action. Out of this, we get the following claim concerning the relationship between responsibility and blameworthiness.

(Rw↔Bw): S is responsible for wrongdoing *W* iff S is blameworthy for *W*.

On the one hand, (Rw↔Bw) entails that you cannot be blameworthy for a wrong act without being responsible for it. I know of no philosopher who rejects this claim and I see no reason to reject it either, so I will take it for granted. On the other hand, (Rw↔Bw) entails that that you cannot have responsibility for wrongdoing without blameworthiness for it. This is a well-accepted, but *controversial* claim, which I will defend in §2.3.4 and in §3.3.3 (next Chapter), ahead of its use later in the thesis.

The relationship between responsibility and blameworthiness aside, I would now like to say something about the concept of blameworthiness itself. I will focus first on the concept of blame, before touching on what being "worthy" of it could amount to.

What, then, has been said about the nature of blame? Increasing interest has been given to the nature of blame in the last two decades or so, often independent of interest in the normative issues surrounding responsibility. Important distinctions between *blame* and *punishment* (e.g., fines, confiscation, imprisonment, grounding), and between *inward* and *outward* blame (mental states vs. actions/omissions), have been familiar for longer. However, the last two decades have witnessed a relatively fresh debate about the exact nature of inward blame—or just "blame," hereafter (see especially Coates and Tognazzini 2012; Tognazzini and Coates 2018). Following Tognazzini's and Coates' taxonomy, some take the *cognitive* view that blame is constituted by cognitive judgments or beliefs (e.g., Tim judges Tabatha responsible for the injustice; Zimmerman 1988; Watson 1996; Hieronymi 2004; Smart 1961). Some take the *emotion-based* view that blame is constituted by *emotions* (e.g., Strawson's [1993 (1962)] "reactive attitudes" of resentment, indignation, and guilt; cf. Wallace 1994; Peels 2017, and many other accountability theorists). Some take the *conative* view that blame is (partially) constituted by *desires* (e.g., I *wish* you never hit me; Sher 2005, 112). And others take "functionalist" views on which blame has a certain social function (e.g., protest) which can be realised by any of the above, including outward actions (McKenna 2012; Macnamara 2011). Of course, any hybrid theory of blame is possible too. In Chapter Five, I

argue in favour of a *cognitive* account of blame for action as judging someone morally at fault for wrongdoing.

What does "being worthy of" blame mean? Theoretically thin synonyms include being "susceptible to," "open to," "eligible to," "liable to," or being "an apt candidate for" (Eshleman 2014) blame. Thicker theoretical analyses of "worthiness" include "meriting," being "deserving of," being "the fair recipient of," being "the proper/fitting object of" (Peels 2017) or being "rationally accessible to" (Fischer and Ravizza 1998, 7) blame. Blame can then be described as "apt," "justified," "reasonable," "warranted," "accurate," "fitting," "fair," "appropriate," or "proportionate." Depending on one's theory of blame, some of these terms will be favoured over others. For instance, a strong emotion-based account of blame (e.g., as resentment or indignation) will favour *fittingness*, or *reasonableness*, while a cognitive account may favour *accuracy* or *justification* (Hieronymi 2004; Zimmerman 1988, 38). We will return to these issues again in Chapter Five. We should note, though, that however one analyses the "worthiness" in blameworthiness, to be worthy of blame is not to be worthy of *a particular person*'s blame. This is because a particular person may be *ineligible* for some reason or other to blame the blameworthy agent (e.g., because it would be hypocritical to do so; Tognazzini and Coates 2012, 203ff). Rather, blameworthiness is something like being worthy, *in principle*, of blame.

This normative element of blameworthiness in particular (and accountability in general) sets it apart from any account of blameworthiness, according to which blameworthiness is simply the quality of *being the typical or dispositional object* of blame. P. F. Strawson (1993 [1962]) has been interpreted as a proponent of such a theory, but this interpretation has been debated (see, e.g., Fischer and Ravizza 1993; N. Nottelmann 2007, 41ff.). No one, however, has taken this view seriously, even strong proponents of so-called "response-dependent" theories of responsibility (Shoemaker 2017). Among several other reasons, dispositional response-dependent accounts fail to capture the obvious truth that those who are typically blamed (or praised) may not be worthy of blame, and similarly those who are *not* typically blamed *may well* be worthy of blame. Consider Fischer and Ravizza's (1993, 18) example of a community in which intellectually disabled members are always blamed and no woman is ever blamed. Does that make the disabled members blameworthy and the women blameless? Clearly it would not.

Shoemaker (2017) defends a more plausible variety of response-dependence about responsibility, according to which blameworthiness is equivalent with being the fitting target of blame (or some other appraisal), where what makes this "fitting" is *not* some feature of the

agent that merits the response (e.g., the agent's control over and awareness of wrongdoing), but some feature about the "refined" functioning of the "anger sensibilities" underlying blame. But along with the vast majority of theorists of responsibility, I assume that such a view is mistaken—that we really *can* come up with a response-independent account of responsibility and blameworthiness which holds that the responsibility responses of praise or blame are fitting only when the agent has certain qualities that merit them. At the very least, I assume that a response-independent account is possible from the start, and I believe that I can strongly motivate at least one of the necessary conditions on responsibility and blameworthiness in this thesis—namely, the epistemic condition. (Indeed, I am not moved by Shoemaker's brief attempt to rule out a necessary epistemic condition.) Moreover, as I have already indicated, I think that we have good reasons to reject an accountability concept of responsibility in the first place as well as his notion of blame as anger (see Chapter Five).

2.3.3 Retrospective Responsibility

Above, the point was made that the "worthiness" component of accountability can be cashed out in terms of the agent's *meriting* or *deserving* the responses in question. To take this analysis is almost invariably to adopt what we might call a "retrospectivist" view of responsibility, in contrast to what can be called a "consequentialist" view of responsibility.[23] What exactly are these views? Retrospectivist views—also called "merit-based," or "backward-looking," views (Fischer and Ravizza 1993, 11; Eshleman 2014; Nottelmann 2007, 38)—hold that praising or blaming someone for *x* is appropriate if and only if that person *deserves* or *merits* these responses in virtue of *x*.[24] By contrast, consequentialist views—also called "social-regulation," "forward-looking," or "prospectivist" views—hold that praising or blaming someone for *x* is appropriate if and only if doing so would lead to desirable consequences for them or others in the future.[25] These desirable consequences might be "social regulation" (Schlick 1966, Smart 1961); or "agency cultivation" (Vargas

[23] It seems possible to deprive the terms "merit" and "deserve" of their normal backward-looking connotations (cf. Vargas 2013, 182). I shall take "meriting" and "deserving" to mean exclusively that *what they did* is what merits blame.

[24] As I am using the term "retrospective," I include so called "non-historical" or "a-temporal" views (Fischer and Ravizza 1993; Nottelmann 2007, 38) which hold that responsibility is a function of the synchronic relation between *x* and the agent (or the agent's will) at some time (e.g., real-self views). These views still hold that blame is *merited* due to what is done at the relevant time.

[25] Following A. Eshleman (2014), I express these views as biconditional claims. There is of course a third "hybridist" view requiring that someone is responsible if and only if praise and blame are deserved *and* would lead to desirable consequences (cf. Pereboom 2016).

2013). Responsibility consequentialists allow that that praising or blaming involve backward-looking judgments—J. J. C. Smart (1961) allows that the schoolmaster can blame the schoolboy *for* being lazy and not doing his homework when he "could have done" it—but the point is that holding responsible only gets its *justification* from something like the fact that it is "socially useful in spurring others on to display more drive than they otherwise would" (Smart 1961, 305; cf. also Vargas 2013, 166, 172). Both kinds of theories have long historical pedigrees, stretching back to Aristotle himself (whose own view, expressed in the *Nicomachean Ethics*, is difficult to interpret as falling on either side; see Eshleman 2014). In recent times, however, consequentialist views have, as Manuel Vargas laments, "gone the way of waxed handlebar moustaches" (2013, 166). The vast majority of contemporary theorists of moral responsibility hold retrospectivist views (although the situation looks different in theories of punishment or legal responsibility).[26]

What is important for our purposes is that these views appear to make a difference to how we answer the question of whether, for the agent to be responsible for her conduct, the agent must have any beliefs or credences in the moral significance of her conduct. It is noteworthy that no volitionist has taken the consequentialist view. This is not surprising, given that it seems to be in the interests of the consequentialist to deny that beliefs or credences concerning the act's moral significance are necessary (at least at an individual level, although there may need to some kind of awareness at a *group* level). After all, one of the justifying ends for which we might blame someone is the end of teaching them what is morally what. It would seem, then, that the consequentialist need only require a *capacity* to have this awareness (excluding only those like psychopaths, who are unable to have this awareness and so are not worth teaching). I take it as especially revealing that Vargas, perhaps consequentialism's most prominent contemporary defender, appeals only to a *capacity* for awareness (i.e., "vigilance"; Murray and Vargas 2020)

Consequentialist views capture a strong intuition that we have about the justification of attitudes and practices of holding responsible. When, for example, we feel strongly that verbal condemnation or resentment towards someone will not "help" the situation or the wrongdoer—however much doing so feels *deserved*—then doing so seems ill-advised. This is the kind of concern that is found in a lot of popular psychology. We often have analogous intuitions about why we hold criminals responsible. Prison is not about criminals' "getting

[26] Consequentialism, or at least some hybrid view (see n. 25), seems more plausible as an account of outwardly holding responsible or punishment. See, e.g., J. Hampton (1984), and a review of this literature in H. Bedau and E. Kelly (2015).

their just deserts," but about their *rehabilitation*, as well about deterrence, education, and the like. Those (like myself) who are inclined to reject the moral permissibility of capital punishment are drawn to this essentially forward-looking justification of imprisonment. Consequentialist views also have other virtues. The views capture the intuition that praise so often incentivises good behaviour and blame deters bad behaviour (Vargas 2003, 165). Relatedly, they account for the idea that as Vargas puts it, "we have reason to *care* about moral responsibility inasmuch as we care about getting people to behave in the right ways and getting them to avoid behaving in the wrong ways" (p. 165). Finally, consequentialist views have the well-known virtue of avoiding the problem of the incompatibility of moral responsibility with the thesis of determinism—that for any event, there is a cause that is sufficient for it (Smart 1962; and cf. Strawson's "optimists," in 1993 [1962]). This is because determinism's being true should not affect our decision to hold someone responsible, especially given the *apparent* (i.e., epistemic) openness of the future.

I would like to grant to the responsibility consequentialist the above intuitions and the fact that their views capture them. But I question whether these virtues take them far enough. Moreover, it seems to me that these intuitions are captured well by retrospectivist views which also, as I will endeavour to show in a moment, have greater independent support.

First off, how can retrospectivist views accommodate the above intuitions? In relation to the fact that we sometimes feel that people should not blame even if it is felt deserved because doing so will not "help" the situation or the wrongdoer, it is important that we get clear on what *to be precise* will not help. There are two dimensions here. It could be that we sometimes feel that people should not blame some wrongdoer *in the circumstances* even though the wrongdoer is *blameworthy*, because those blaming do not have the right *standing* to blame (e.g., they are hypocrites; see §2.3.2). Or it could be that we feel that people should not *in principle* blame the wrongdoer, even though it seems deserved. Very often it is the former and not the latter—in which case, this observation does not offer much support to consequentialism. But even when we feel that the wrongdoer should not be blamed *in principle*, I think that we do not have in mind inward blame (e.g., having *beliefs* about someone's being at fault, or having harsh negative *emotions* toward someone) but outward blame (e.g., verbal condemnation or deliberate avoidance), or, more plausibly I think, a certain severe form of sanction (e.g., punishment). But if so, it is doubtful that feeling that people should avoid blame when it will not help, understood as *outward* blame, shows anything about the correct general account of moral responsibility. It seems to me that the accountability (that is materially equivalent with responsibility) is best construed as the

quality of being worthy of *inward* praise and blame (or neutral appraisal);[27] after all, inward blame is, if you like, the *core* of blame. Moreover, one could very well take a consequentialist (or "hybridist"; see n. 25) line about outward blame or certain forms thereof (e.g., punishment), but take a retrospectivist line about inward blame and moral responsibility—the asymmetry owing to the fact that outward blame is much more likely to cause harm than inward blame (and so should be further regulated by a consequentialist condition). We have already noted, after all, that consequentialism about punishment is more widely accepted.

The next two intuitions allegedly in favour of responsibility consequentialism also seem to be accounted for, on retrospectivist theories. It is not surprising on retrospectivist theories that praise and blame often lead to good consequences. Blame hurts; praise feels good; and to the extent that people try to avoid being hurt and try to feel good, blame and praise have a positive effect. But often praise and blame have negative consequences (e.g., in cases of scapegoating or of reinforced bad behaviour). And, of course, both theories can explain why either is inappropriate. But if the basic idea that praise and blame have positive consequences lends (if only slight) credence to consequentialist views, I wonder whether our judgment would be "evened-out" upon consideration of the overall picture, including also *bad* praise and blame.

The other intuition in favour of consequentialism is that we only care about moral responsibility to the extent that we care about getting people to behave properly (in the future), and that we should formulate a theory in line with what we care about. I grant that my own interest in responsibility is largely motivated by this, but I think that it is false that people only care about responsibility for that reason. Another significant motivation for the investigation of responsibility is a motivation of justice or desert. One may be motivated by the desire to determine whether, and the extent to which, someone *deserves* resentment or even punishment, for what they have done. We might also be motivated to study it for purely philosophical reasons. But in the end, not too much weight should be attached to the second premise of this argument. It should not count (seriously) against a normative theory that it is formulated independently of the motivation behind investigating that subject.

[27] In Chapter Five, I defend the view that I blame the mosque shooter *just when* I judge (believe) that he is morally at fault, or in the wrong, for the shooting. But as such, it is plausible that blameworthiness for wrongdoing is the quality of being *worthy of this judgment* and this judgment only, such that being worthy of *outward* blame is at best incidental to blameworthiness, and so also moral responsibility (provided that responsibility is materially equivalent with accountability).

Finally, there has been no shortage of retrospectivist solutions to the problem of how moral responsibility is compatible with determinism or indeterminism (see, e.g., Frankfurt 1971; Fischer and Ravizza 1998; and the review of so-called "Historical Compatibilists" in Levy and McKenna 2009, 108ff.).

Thus, there seem to be good replies to the arguments for consequentialism. But, more positively, retrospectivism also seems able to account for intuitions which consequentialists may find difficult to account for. We have already signalled one of these intuitions, concerning the nature of blame, in our discussion above. But let us look closely at the nature of our characteristic "reactive attitudes." They are called "reactive" for a reason. They are often far from calculated. At the forefront of our minds as blamers is not the thought that it would be beneficial to blame the blameworthy, but thoughts like, "that was horrible!", "how could you!?", "tough luck," and so on. Even when we intend to "teach them a lesson," we are not motivated primarily by teaching the blameworthy anything other than "that's what they get for doing something like that." Blaming the *dead* and *distant*, when we know that our blame will never be communicated to them, also seems perfectly appropriate, even when it leads to no discernibly good consequences for us or for people affected by the blamed. I blame Hitler for what he did, and surely this is justifiable, even though it is unclear how this leads to better consequences.

In response, the consequentialist could account for the justification of these blaming attitudes and practices but reply that this objection only really targets crude forms of consequentialism: the point about the backwards-looking nature of reactive attitudes only really targets a *direct* form of consequentialism on which blame produces the desirable consequences only when there is an (explicit) intention to bring about those desirable consequences. It does nothing to dismantle an *indirect* form of consequentialism, on which blame need not (maybe even *ought not*) involve this intention in order to bring about desirable consequences (Vargas 2013; cf. also McGeer 2014). Furthermore, the point about blaming dead or distant wrongdoers does not dismantle a(n indirect) consequentialism on which blaming them *in general* is justified because, as Vargas puts it, "the system as a whole [including blaming the dead and distant] produces agents that, over time and in a wide range of contexts, are suitably responsive to moral considerations" (2003, 177).

With these more sophisticated kinds of consequentialism, I grant that the consequentialist provides a rational alternative to retrospectivism. But in this thesis, I will rest on my retrospectivist laurels. Moreover, I am not yet convinced that consequentialism is as defensible as restrospectivism. I wish to close my case against consequentialism with the

following objection, designed to target Vargas' indirect consequentialism.

Blame in an Ending World

Suppose that the human race has just been ruthlessly harmed by a race of alien persons, who did so freely and knowing that they ought not to (etc.). Suppose also that in doing so, the aliens set in motion our eventual demise so that we only have one month left to live. We have no way of preventing it or communicating to them. And what is more, we have been separated from each other, so we could not achieve a kind of solidary through bonding over our shared trauma.

In such an ending world scenario, it would seem that the revised—Vargas-inspired—consequentialist should deem displays of outward blame of this now-distant race of aliens unjustified, because there would be "no use" in such blame. Throwing our fists up at the sky would not help us socially, regulate our behaviour, teach us anything, or comfort us. In fact, outward blame would seem a positive waste of time and energy, given only a month left to live. And yet, blame would entirely make sense. Intuitively it would be fair. This reveals to me that even when no good can come from blame (for you, or for anyone else, even at a group level), blame can still be justified—and thus that there is no condition on being worthy of blame that this response brings about desirable consequences. Sometimes blame is appropriate, regardless of the consequences of being blamed.[28]

2.3.4 General Conditions on Responsibility & Blameworthiness

We have already seen that a commonly proposed condition on responsibility and blameworthiness (understood retrospectively) is some kind of awareness or epistemic state. This is after all the key topic of my thesis. But what other conditions have been proposed for responsibility or as conditions that merit blame? An answer to this question will help us tighten our grip on the nature of (retrospective) responsibility.

To a greater or lesser extent we have already introduced the freedom or control condition. It was mentioned briefly in passing last Chapter that most of the work to date on responsibility has centred on the nature of this condition—whether it should be analysed as compatible with determinism or not (the thesis that for every event or action there is a sufficient cause for it). Philosophers have proposed many different types of control or

[28] I have benefited hugely from Stephen Setman's comments on an earlier version of this Subsection.

freedom conditions on responsibility (at a glance, "asymmetrical" freedom, Wolf 1980, Nelkin 2011; "self-determination," G Strawson 1994; compatibilist "reasons-responsiveness," Fischer and Ravizza 1998; "capacitarian control," Rudy-Hiller 2017; and more). Many endorse the traditional view that the control and epistemic conditions are necessary and jointly sufficient conditions for responsibility (see, e.g., Levy 2005; Rudy-Hiller 2018; Fischer and Tognazzini 2009, 531ff.; Nelkin 2011; Clarke 2017, 65f.; Peels 2017; Levy and McKenna 2009, 115; Aristotle 2013, III.1-2). But others propose conditions on responsibility that are either distinct from control or epistemic conditions (e.g., moral luck, Hartman 2016; "origination," Sher 2005) or do not fit neatly into either (e.g., the real-self condition in Frankfurt 1971 and Watson 1975).[29] An important example of the former (for our purposes) is the "quality of will" condition, to which Strawson (1993 [1962], 56) attached much significance. To affirm such a condition is to affirm that responsibility (as accountability) for an action or attitude depends (partly) on whether the action or attitude displays a (good or bad) quality of the agent's will (e.g., her desires, cares, character, or evaluative attitudes). Defenders of the quality of will condition (often called "attributionists," Levy 2005) typically offer their accounts as alternatives to the control condition (see, e.g., Adams 1985; Scanlon 1998; A. Smith 2005; G. Sher 2009; Talbert 2017a; Harman 2011). Faced with the force of intuitions pulling either way, G. Watson (1996) and E. Mason (2015) divide responsibility in two (so that one form of responsibility requires control while the other requires only a display of the self or their quality of will).

As we have already noted, blameworthiness in particular also has a *value* condition which, when we interested in actions and omissions, requires that they are wrong all-things-considered (on the view that I defend next Chapter, §3.3). The value condition on blameworthiness should be distinguished from the "agential" conditions in blameworthiness and responsibility discussed above.

Some (e.g., Fischer 2007, 186; M. McKenna 2012, 19-20) have argued that blameworthiness requires the satisfaction of more agential conditions than responsibility. But notice that such a view would constitute a rejection of our claim above, **(**Rw↔Bw), for the reason that the following claim turns out to be false: that responsibility for wrongdoing entails blameworthiness for it (Rw→B). Fortunately, I do not see that such drastic revisions are required, because I will argue that Fischer and McKenna's objections do not show that, if

[29] These are intended as accounts of responsibility-relevant freedom, however the "freedom of the will" (Frankfurt 1971) that these theorists have in mind is quite different in nature from the other more historical accounts of freedom—from what Frankfurt (1971, 14) calls "freedom of action."

we need to make any revisions, they need to be revisions of any *agential* conditions. Here I shall focus only on Fischer's objections, but I will return to McKenna's case in §3.3.3 next Chapter.

Fischer argues that an agent can be responsible for wrongdoing without being blameworthy on the basis of intuitions generated by reflection on two cases.

Murder After Abuse

Consider a scenario in which there has been substantial and recurrent physical and emotional abuse by a husband of his wife over many years. The wife has tried to leave this toxic and abusive relationship, but she has not been able to summon the strength. Finally, after her husband has begun (yet again) to beat her cruelly and brutally, she shoots him to death. (Fischer 2007, 186)

According to Fischer, the wife is *responsible* for her wrong act of killing her husband, in virtue of meeting certain control, epistemic, and (weak) history conditions, without being blameworthy for it. Only a certain *provenance* condition is not met—namely, the provenance of the agent's "motivational states issuing in [her] behaviour" (p. 186).[30] Since the wife's motivational states issuing in the shot (e.g., her fear, anger, or resentment) were created under the unfortunate circumstances of cruel domestic abuse, the provenance condition is not met and so she is not blameworthy.

A second case seems to involve the failure to satisfy another condition.

Maria Full of Grace

Consider, similarly, a 'drug-runner' of the sort depicted in the film, *Maria Full of Grace*. Suppose, more explicitly, that the individual was born to poverty, and was under considerable pressure to transport illegal drugs to America. Again, we can stipulate that the pressures were considerable, but that they fell short of issuing in compulsion. (2007, 186)

In this case too, Fischer argues that the drug-runner appears to be responsible for the wrong act transporting illegal drugs without being blameworthy for it. This time, however, the agent is off the hook because she was in circumstances such that it was *extremely difficult* to do the right thing (to not transport illegal drugs).[31] The common thread is that both cases concern

[30] See a good account of how this condition might be developed in K. Fritz (2014).
[31] Again, see Fritz (2014).

unfortunate circumstances. So what shall we say about Fischer's case for these conditions on blameworthiness, and not responsibility, in particular?

Like K. Fritz (2014), I find myself unconvinced that the responsible wrongdoers are blameless in these cases. Two considerations are relevant here. The first is one that we have already encountered: a wrongdoer's blameworthiness does not entail that anyone who knows about the wrongdoer's blameworthiness should blame or at least *outwardly* blame the wrongdoer. In *Murder After Abuse*, for example, it may not seem natural or fair to condemn the abused wife (especially after years of ongoing abuse). I know that I probably would not if I were her good friend. But the point is that she can still be blameworthy even if we feel that it would be inappropriate to outwardly blame her. It *would* still be fair, I think, for us to judge that she was morally at fault for it, or to experience mild indignation toward her (e.g., *qua* one of *his* family members), and these indicate blameworthiness. The second consideration is the fact that blameworthiness can come in *degrees*. In both cases, the wrongdoer is clearly not as blameworthy as she would have been if she had not been bedevilled by her unfortunate circumstances—say, if the wife's husband was kind and gentle, or the drug-runner could easily have taken another path in life. But it does not follow that the agent is not blameworthy at all. Rather, I think that in both cases (especially *Murder After Abuse*) the wrongdoer is *minimally* blameworthy. After all, in both cases the agent was in control, knew better, and so on, and yet she committed the wrongdoing anyway. I concede only that the appropriateness of blaming the blameworthy—and that greater *degrees* of blameworthiness (see below)—probably require the satisfaction of provenance and difficulty conditions. Thus, for now **(**Rw↔Bw), because (Rw→B), remains intact.

2.3.5 Other Dimensions of Responsibility: Degrees of Responsibility & Tracing

The appeal to *degrees* of blameworthiness attests to a dimension of responsibility or blameworthiness, among other dimensions, which we have not yet introduced. I will close §2.3 by saying something more about this dimension, and by introducing another—far more important—dimension for this thesis, the phenomenon of *tracing* responsibility (and blameworthiness) at one time back to responsibility at an *earlier* time. Before I do so, however, I would briefly like to acknowledge two dimensions of responsibility which will not be in focus. One is the dimension of *shared* responsibility. Earlier (in §2.2) I introduced the possibility that agents can be groups (and so maybe groups can be responsible), but sometimes there is a question about the extent to which *each individual within* a group is responsible for a shared activity, process, or event. Responsibility or blameworthiness for a

single event can be divided between persons. I will not be concerned with this issue in this thesis, and I will use restrict the language of "partial" responsibility or blameworthiness to the *partial degree* to which an individual is responsible or blameworthy (mildly or severely; see below). Another less important dimension of responsibility is the distinction between *being* and *taking* responsibility. We have so far been talking about *being* responsible (and holding responsible) but *taking* responsibility can mean something different. Taking responsibility for something (e.g., an event) is sometimes about *becoming* responsible for something (e.g., as a new event organiser); other times it is about admitting to, discovering, or "owning up" to, one's *pre-existing* responsibility. I will not discuss taking responsibility unless I take it to be precondition for *being* responsible (as on Fischer and Ravizza's [1998, 211] view).

We have just seen the distinction between degrees of blameworthiness at play in *Murder After Abuse* and *Maria Full of Grace*, but it is a distinction that can also be made in relation to responsibility. In other words, I hold that *responsibility* can come in degrees, and an initial division of different degrees is the division between full responsibility as responsibility to the maximal degree, and partial responsibility as responsibility to some degree. By *minimal* responsibility I mean partial responsibility, but to a low degree. In this thesis, I will not give an account of what makes someone fully as opposed to partially responsible for something, but it is probably a mixture between the degrees to which the conditions of responsibility obtain (e.g., full vs. partial control),[32] and other conditions that are not conditions of responsibility *as such*, but of having responsibility to a higher degree (such as Fischer's provenance and difficulty conditions).

Corresponding to this distinction between full and partial responsibility is the distinction some make between full and partial *excuses* for wrongdoing (see, e.g., Peels 2014). A partial excuse for wrongdoing entails that the agent cannot be fully blameworthy (and responsible) but may still be partially blameworthy (and responsible). A full excuse for wrongdoing entails that the agent cannot be blameworthy or responsible at all for it.

Finally, an extremely important distinction for this thesis (indeed, dividing the subject matter of Chapter Seven from Chapter Eight) is the distinction between *original* and *derivative* responsibility/blameworthiness, or *direct* and *indirect* responsibility/blameworthiness. Moreover, when responsibility for something is indirect or

[32] See Husak (2016, ch. 3) for a good treatment of how degrees of blameworthiness (and criminal liability) are sensitive to the epistemic condition.

derivative, it is said that responsibility for that thing "traces" back to an instance of direct or original responsibility (Vargas 2005; Fischer and Tognazzini 2009). One is directly or originally responsibility for something if and only if, at *t,* one satisfies the conditions on responsibility for that thing (set out above). Typically, the focus is on the satisfaction of the control and epistemic conditions at *t*. Indeed, it is often said that one is directly responsible for something only if one has *direct control* over it—which is probably where we get the "direct/indirect" terminology. For many, this means that one can be directly responsible only over actions (or "basic actions"; see next Chapter), and so direct responsibility is denied for omissions, mental states (beliefs, desires, moods, emotions), character traits, and other consequences of actions.[33] By way of illustration, contrast the action of choosing to binge-drink with the damage to someone's property caused *whilst* heavily intoxicated. Since one's control and awareness of the damage caused whilst heavily intoxicated is heavily compromised, it seems that one cannot be directly responsible for this damage. What this means is either that one is not responsible for the damage *at all* or that one is *indirectly* or *derivatively* responsible for it. If one is indirectly responsible for it, one is responsible for it *via* responsibility for something else at the earlier time *t-1*—for instance, the choice to binge-drink in the first place. This is why the responsibility in question here is described as "indirect." Of course, responsibility for something else could also be indirect responsibility, in which case we must consider responsibility for something further at any even earlier time *t-2*, and so on. It becomes clear that responsibility must "bottom out" in a case of direct or *original* responsibility.

This process of attempting to "derive" responsibility for something from a prior instance of original responsibility has come to be known as tracing. With tracing, it is not as though responsibility from the original act is literally "given over" to the consequence (Wieland and Robichaud 2017, 282). What seems required are at least two conditions (expressed in terms of responsibility for action):

A. *Counterfactual dependence*: One is responsible for the later act *only if* one is responsible for the earlier act—such that without the latter, there would not have been the former.

[33] But see especially Chapter Six for exceptions (especially among so-called "capacitarians" with regard to direct responsibility for omissions and for ignorance).

B. *Explanation*: One is responsible for the later act *because* one is responsible for the earlier act—such that the latter, in some (partial) sense, *explains* the former.

If we were to deny counterfactual dependence, then we could not say that responsibility for the later act was *derivative of* responsibility for the earlier act. But, as Wieland and Robichaud (2017, 282-3) have pointed out, this necessary condition is not quite enough. Suppose that it was necessary for the drunkard to be responsible for the damage to someone's property that he was responsible for binge-drinking in the first place. But suppose also that his choice to binge-drink meant that he forgot to attend his daughter's birthday party (such that he would not have forgotten had he remained sober). In that case, responsibility for binge-drinking explained his responsibility for missing his daughter's birthday party. But if so, then we can say that the drunkard is responsible for the damage caused whilst drunk *only if* he is responsible for missing his daughter's birthday. But the latter clearly does not *explain* the former. Rather, both are explained by a "common cause": *responsibility for getting drunk in the first place*. Thus, tracing responsibility to an earlier act requires that one is responsible for the later act only if *and because* one is responsible for the earlier act.

More could be said to refine the "explanation" condition. Indeed, orthodoxy on responsibility for consequences holds that there must be an *epistemic* condition on tracing—namely, foresight of the consequences (the drunkard must have foreseen the possibility of causing damage to someone's property or of this sort of consequence) (see, e.g., Vargas 2005; Fischer and Tognazzini 2009; Timpe 2011; Nottelmann 2007, 189–201; Nelkin and Rickless 2017). There is also a puzzle (not relevant to this thesis) about how the *degrees* of responsibility for both the earlier act and the later act aggregate, if they aggregate at all (Wieland and Robichaud 2017). But given that I do not offer an account of degrees of responsibility in this thesis, I remain agnostic about that. Finally, it is plausible that in cases of tracing *culpability* back to culpability for an original act, the act is itself bad (or wrong, as I argue culpable conduct must be next Chapter) (H. Smith 1983, 547). If it were morally outstanding, by contrast, how could the act be an original source of *blameworthiness*?

There is one key source of opposition to this overall picture of tracing.[34] Matt King's (2017) view is that every case of tracing can be explained in terms of either the intent, recklessness, or negligence of a temporally-extended complex action, encapsulating both the

[34] A view that Holly Smith (1983) calls the "Liberal View" also challenges this picture. I discuss that view in Chapter Four.

"earlier" and "later" acts that we have been discussing. The drunkard who knows the risks of causing damage to someone's property when getting drunk in the first place commits one temporally-extended action—*risking causing damage*—for which he is responsible (simpliciter) due to its recklessness. Since this view is simpler, for King, it is preferable (2017, 269). But King also insists that tracing explanations are needless: "[there] is no need to carve up the components of each complex action into its constituents and trace responsibility throughout. Rather, the whole can remain" (2017, 270).

I submit that King *does* provide a genuinely defensible alternative to tracing explanations. But I think that Wieland and Robichaud (2017, 297) are right in their reply that King's alternative papers over genuine complexities for which tracing may be needed to account (concerning the puzzle about how to aggregate degrees of blameworthiness for both fine-grained acts). King also concedes that tracing explanations are still *adequate* (2017, 269). In response, I take it that, even if tracing proves unnecessary in the end, I have reason enough to follow convention and employ tracing explanations in this thesis. Moreover, I should submit my own preference for individuating acts into finer grained acts than King prefers to individuate them (see §3.2 next Chapter). For King (2017, 269-70), temporarily putting baking on hold to get supplies before resuming baking constitutes one complex, temporally-extended action. I prefer to see it as at least three distinct fine-grained actions (baking, getting supplies, and baking once again).

2.4 Conclusion

The guiding question of this thesis is whether, and if so, in what way, the agent's moral responsibility or blameworthiness for her conduct depends on her beliefs or credences about the moral significance of her conduct. We have just introduced and defended certain accounts of three of the concepts involved in this question, the concepts of the *agent*, of her *moral responsibility* and her *moral blameworthiness*. When I speak of the agent, I mean the individual person, and typically the normally functioning, fully-formed (human) adult, who possesses moral agency. When I speak of moral responsibility, I mean the moral attributability that grounds moral accountability, of which blameworthiness is one mode. Blameworthiness is merited by the *negative moral significance* of that for which one is accountable. With respect to both moral responsibility and blameworthiness, I am a restrospectivist theorist: neither depend on being the object of blame or being held

responsible when justified by its consequences. Conditions for both moral responsibility and blameworthiness typically notably include, among others, voluntariness, control/freedom, an appropriate history/self-disclosure, quality of will, and a kind of awareness. And I have defended the claim that responsibility and blameworthiness both share the same agential conditions, against the proposal that blameworthiness has further provenance and difficulty conditions. Finally, both come in degrees, and come in direct/indirect varieties, depending on when the responsibility conditions are satisfied; and neither "taking" responsibility nor "sharing" responsibility between individuals is my concern in this thesis. The concepts that we have left to introduce and narrow down are the concepts of the agent's *actions or omissions*, their *moral significance*, and the agent's *beliefs and credences*.

Chapter 3
The Rudiments Part II:
Actions, Wrongdoing, and Epistemic States

3.1 Introduction

Having introduced and motivated a certain understanding of two rudimentary (sets of) concepts in this thesis—the *agent* and their *moral responsibility and blameworthiness*—we turn now to the three others that we must discuss: *actions* and *omissions*, their *negative moral significance*, and the agent's *beliefs and credences* in their conduct's significance (i.e., their "epistemic states"). Recall that our guiding question is whether, and if so, in what way, the agent's blameworthiness for some conduct depends on the agent's having beliefs and credences concerning the conduct's moral significance. Thus, we are to approach the analysis of the key concepts of this Chapter in light of the concepts of moral responsibility and blameworthiness already on the table. We will therefore let considerations of responsibility and blameworthiness from the previous Chapter dictate what should and should not be said about conduct, its moral significance, and its significance relative to the agent's beliefs and credences. *Complete* accounts of these concepts shall not be given, and some aspects of these concepts will not be discussed (e.g., the way that beliefs are related to desires, or the way that actions are related to their consequences).

The structure of the Chapter is as follows. §3.2 discusses actions and omissions. §3.3 discusses the type of moral significance that is required for conduct to be culpable, which I argue is wrongfulness all-things-considered. And §3.4 discusses the agent's beliefs and credences concerning their conduct's wrongfulness all-things-considered.

3.2 Actions (& Omissions)

We first turn our attention to the primary "objects" (Eshleman 2014) of responsibility assessment within view in this project: the agent's *actions and omissions* (constituting the

agent's "conduct"). I begin with actions, which initially should be distinguished from mere "happenings," bodily movements, or instances of physical behaviour. If I wave at you, I may be performing the action of waving at you to get your attention *or* I may be having an uncontrollable spasm. If I am doing the former, I am performing an action, but if I am doing the latter, I am merely behaving physically. Physical behaviour may accompany an action—for instance, as the direct *effect* of it or as a *constituent* of it—otherwise it may not accompany an act at all, such as in the case of a muscle spasm. Actions, on the other hand, need not involve or result in physical behaviour, such as in the performance of *mental* actions like supposing, hypothesising, postulating, accepting a proposition, making decisions, or forming intentions. But what explains how actions can be both purely mental activities and physical behaviour is plausibly the fact that, in D. Davidson's (1980) words, they are "intentional under some description"; they are inextricably linked with an intention to do something (whatever intentions are and their close relation to actions; see Wilson and Shpall 2012). Related to this is the fact that we often have knowledge or awareness of them (at least under a factual description), but also the fact that they are—to a certain extent—under our *direct control*. I can make a direct difference to whether or not I perform an action. It is plausible, however, that I have direct control only over *an aspect* of the action, what we might call the "basic action" or the "volition" (Alston 1988, 260; Zimmerman 2008, 184–85). To wave at someone, I must perform the simple physical movements of raising and moving my hand back and forth, but even to perform these movements, I must *try* (or form a *volition*) to move them. No step guarantees the next step (e.g., trying does not entail succeeding in movement), and the most basic of these elements is the *trying*, and so it seems I have direct control only over the *trying*. But *trying* to wave my hand is surely not the *whole* action here envisioned (as some would like to think), and so I only have direct control over an element of the action, the "basic action" (Wilson and Shpall 2012). That is, at any rate, the notion of action with which I am working; and it is a standard conception. There are other views close to this view (saying, e.g., that each step is a separable, fine-grained action), but not much hangs on the debate between these views. As we saw last Chapter (§2.3.5) I should acknowledge finally that I take a more fine-grained approach to action individuation. For example, temporarily putting baking on hold to go get supplies and then resuming it involves at least three separable actions, not one overall, temporally-extended action.

As we saw when discussing M. King's (2017) view, how the act is specified, or *which* description is used, can make a difference to responsibility assessments. Consider the following. Oxford University is well-known for its quirky traditions, and one such tradition is

that no one has the privilege of walking across the luscious green lawns *other* than the "dons," the Oxford professors. Suppose that a newly arrived foreign graduate student is caught in the act of walking across the lawn. It would make sense for the groundskeeper reviewing video footage of the incident to demand, "who is responsible [blameworthy] for walking across the lawn!?" But suppose that the student was late to a meeting with his tutor and had not yet learned about the conventional restrictions on walking on the lawns. Suppose, also, that he had had no reasonable opportunity to learn about the tradition (e.g., there were no signs up). Would it have made sense for the groundskeeper to insist on the student's responsibility (or blameworthiness) for *violating the lawn rules*? The answer is, plausibly, "no." Yes, the student *violated the lawn rules*, but his ignorance about the lawn rules seems to excuse him for it. Surely, however, he *was* responsible for the acts "by" which he violated the lawn rules—*walking across the lawns* and *taking a shortcut to get to his tutor on time*. What this shows is that even though the agent performs one-and-the-same act, the agent is responsible (blameworthy) only for the act *under one description*, rather than the act under another. (Of course, the tutor was also responsible for *walking across the lawns*, but he could not have been blameworthy for doing so, given that there was nothing in principle wrong with walking across lawns.)

I turn now to omissions which are, on the other hand, "failures" to act. Defining omissions is difficult, for in some sense we are always omitting to do something (e.g., I am not now doing the dishes or watching a Liverpool FC football match). But some features about them should be noted. Omissions, to begin with, can be intentional or unintentional. I can intentionally omit to look at a ghastly image, and when I do it appears that I *decide* to omit. But sometimes omissions can be unintentional, such as in a case of omitting to save a child from drowning in a case where there is no awareness of the incident and therefore no chance for a decision to omit saving them. However, a lack of awareness of an omission is not required for unintentionally omitting to do something. A few moments ago, my wife intended to get changed into her shorts when she went upstairs, but while up there, she got distracted by the need to put the clean laundry away, and so omitted to change into her shorts before she came back downstairs. She did this *without deciding to omit* changing into them (due to the distraction). But she *was* aware of her goal to change into them while upstairs. So omissions can be known without being decided or intended (Nelkin and Rickless 2017). Thus, omissions can be advertent decisions not to act, or non-decisions either despite awareness of them or without awareness of them at all. Finally, as with actions, it matters how the

omission is specified for the purposes of responsibility assessment. And I will work with a basic notion of them in this thesis.

3.3 Blameworthiness for All-Things-Considered Wrongdoing

We turn now to saying something about the overall negative moral significance that actions and omissions must have for agents to be blameworthy for them. Recall the point made last Chapter that there is no "value condition" on *responsibility* as such (i.e., one can be responsible for any kind of act), however, there is a value condition on each mode of accountability which is determined (in part) by the *valence* of that conduct. Praiseworthy conduct is *good* in some overall way (right, virtuous, supererogatory, etc.), neutrally appraisable conduct is neutral in some overall way, and blameworthy conduct is *bad* in some overall way (wrong, vicious, suberogatory, etc.). This initial division is clearly plausible. If some act is terribly wrong, they surely cannot be to praise for it. Likewise, if some act is overall morally outstanding for someone, they surely cannot be to blame for it. Given our focus on blameworthiness, we will focus on the negative valence required for blameworthy conduct. But "bad in some overall way" is not yet precise enough. I endorse the following view:

> **BW:** Blameworthiness for actions (and omissions) requires that they are *all-things-considered wrong*.

BW is the standard view of the relevant negative moral significance for culpable conduct, at least on the face of it. But I take BW to imply some things that others do not. In §3.3.5, I argue that wrongdoing is often in a certain sense "subjective"—that is, relative to what is *factually apparent* to the agent, whether or not that is true. In §3.3.1, §3.3.3, and §3.3.6, I introduce and defend the view that wrongdoing entails a right (apparent) alternative, and thus that blameworthiness requires (apparent) alternatives. But BW, on the face of it, also needs some motivating. I give an initial defence of it in §3.3.1, and then again against a rejection of it on the grounds that permissible but "suberogatory" conduct can be culpable (§3.3.4). Finally, I touch very briefly (in §3.3.2) on two meta-ethical assumptions made in this thesis. For now, let us look more closely at the concept of all-things-considered wrongdoing.

3.3.1 All-Things-Considered Wrongdoing & Normative Reasons

"Wrongfulness" is a negative deontic status that we assign to actions or omissions, and is to be contrasted primarily the positive deontic status of "rightness." As a negative deontic status, the words "forbidden," "prohibited," and "impermissible" are often used synonymously with "wrong." If an act is wrong, it is generally also described as what one "ought not to do" or what one is "obliged" or "required" to avoid. All of these terms have, and have been given, more than one meaning (cf. "ought" as distinct from "obligatory" in: Driver 1992; Harman 2016). But at face value, and in standard deontic logic (cf. McNamara 2010), they all pick out the same kind of act. Accordingly, I will treat them synonymously, at least when "wrong" means "*all-things-considered* wrong," rather than merely "*prima facie* wrong" (Ross 1930, 41). After all, some all-things-considered right or permissible acts are still *prima facie* wrong in the circumstances. My focus will be restricted to wrongdoing all-things-considered (or just "wrongdoing" for short), because one plausibly cannot be blameworthy for conduct that is all-things-considered right or permissible. To illustrate wrongdoing, recall the acts for which the agent is responsible (and blameworthy) in *Murder After Abuse* and *Maria Full of Grace* (from last Chapter): shooting one's husband and drug-smuggling. Consider also the Christchurch mosque shooting, gender pay gaps, and the Józefów Massacre.

In contrast to wrong acts are acts that are "right," "permissible," "obligatory/required," or "supererogatory." When there is only one alternative B to the wrong act A, B is both obligatory and right, but not supererogatory (where supererogatory acts go *above and beyond* obligation). When there is more than one alternative B (C, D, etc.), there are a variety of possibilities.[35]

By "wrongdoing all-things-considered," I also mean wrongdoing having considered all the *types of value* relevant to judgments of wrongdoing (and thereby blameworthiness). What are these types of value? Here we get into very thorny territory. Traditionally, the relevant type of value is *morality*, such that wrongdoing and blameworthiness are always *moral* wrongdoing and blameworthiness. "All-things-considered" wrongdoing would then also, by definition, mean all-things-considered *moral* wrongdoing. But some define "moral" considerations more narrowly to pick out "other-regarding" considerations (e.g., considerations of honesty, justice, respect, and duties; Williams 1985; Scanlon 1998), and

[35] E.g., may be more than one wrong act, there may be a permissible act and a right act, there may be a permissible act and a supererogatory act, and so on.

some hold that moral considerations are one class of considerations *among others* (e.g., epistemic considerations, Nottelmann 2007; prudential considerations, Haji 2010) which factor into all-things-considered wrongdoing and/or blameworthiness.[36] After all, "self-regarding" or "prudential" actions are plausibly sometimes wrong (e.g., harming yourself), and sometimes it seems that "epistemically" bad acts are wrong (e.g., refusing to consider counterevidence; cf. Clifford 1886). Typically, on these views, "all-things-considered moral wrongdoing" means "all-things-considered wrongdoing, for specifically moral reasons," but there can also be all-things-considered wrongdoing, for specifically prudential reasons. However, someone who defines "moral" broadly to cover all of the above types of value (arguing, e.g., that it is *immoral* to be idle or wasteful; Scanlon 1998, 6) does not necessarily disagree that these narrower (first-order) types of value can factor into wrongdoing all-things-considered. They just hold that other-regarding wrongdoing or prudential wrongdoing are always instances of *moral* wrongdoing. In other words, other-regarding, prudential, epistemic, and other considerations are treated as *types of moral* considerations. For my purposes, I shall adopt the term "moral" in this wide sense, unless otherwise specified. I do so in part to leave the scope of my discussion wide enough to apply to (or be *open* to applying to) more than *other-regarding* actions (even though most of my examples will be of the latter kind), but I also do so to keep consistent with my use of "moral" in "moral responsibility" or in "moral significance" which philosophers both inside and outside of the responsibility literature conventionally use to distinguish the relevant concepts of responsibility and significance from other concepts to which the words can refer. (If, however, it turns out that wrongdoing or blameworthiness can only be moral in the narrow sense, implying the violation of *other-regarding* obligations, then "moral" and "all-things-considered" should be read accordingly.)

Now, for each instance of wrongdoing, there are corresponding "wrong-making features," or features of the act, agent, or circumstances, that make the act wrong (see, e.g., Scanlon 1998, 10; Arpaly 2002, 79; Rosen 2008, 593). Any given wrong-making feature may be either individually sufficient, or jointly sufficient with other features for the *all-things-considered* wrongness of the act (which otherwise would only be sufficient for the mere badness or *prima facie* wrongness of the act). Due to my commitment to the normative usefulness, indeed I think indispensability, of the notion of "normative reasons" (shared by

[36] Williams (1985) confines blameworthiness to moral considerations, narrowly construed (apparently not recognising blameworthiness for wrongdoing other than other-regarding wrongdoing).

many other ethicists but which I cannot defend at length in this thesis), I shall prefer to call these features *reasons why the act is wrong* or wrong-making reasons, among the wider class of *normative reasons* against the act (e.g., normative reasons why the act is imprudent) (following, e.g., Arpaly 2002, 72; Rosen 2008, 593; Harman 2016, 374). Thus, the fact that the murdered husband had a right to life was a reason why it was wrong to shoot him to death. The fact that women are equal to men is a reason why it is wrong to give men more money for equal work. Both of these "facts" are not themselves *moral* facts, but facts with important moral implications; they are "morally significant." They may have thick evaluative content (e.g., "that is just," involving both evaluative and descriptive content) or no evaluative content (e.g., "he is human"); but neither presuppose *moral wrongness* or *rightness* (like, for example, "that is morally untoward"), for they are supposed to *determine* those properties, not presuppose them. (Confusingly, philosophers sometimes call them "moral reasons" but only given that they are what *make* the acts in question morally right or wrong; Arpaly 2002, 72. I will try to avoid this terminology when I can.) This distinction between moral status and reasons seems to apply across the board to all forms of deontic status (but perhaps not to evaluative statuses like "good" or "bad," which plausibly *generate* reasons [Raz 2002, 1; Tappolet 2004, 399]). Precisely *which* reasons against the act are the wrong-making reasons in a given case is clearly not within the scope of this project: that is the question for first-order normative ethical theories (e.g., deontological, consequentialist, and virtue ethical theories; cf. Arpaly 2002, 72). Since these theories tend to agree on the vast majority of cases of alleged wrongdoing, though, we should not be at a loss for clear examples of wrongdoing.

The type of reasons to which I refer here are *normative* reasons, or the reasons "that there are" in favour of, or against, some act (or omission). Some think that they are always "considerations that count in favour" of an action (Scanlon 1998, 17); others think that they come in both "reasons for" and "reasons against" varieties (Snedegar 2018). Importantly, normative reasons should be distinguished from what have come to be called "motivating reasons," or the reasons that the agent *takes herself to have* (see §3.4.2), as well as from the reasons known as "explanatory reasons" that actually *explain* the agent's conduct (Alvarez 2016). The distinction between normative and motivating reasons is an extremely important distinction for this thesis. Someone might have a normative reason to do something without a motivating reason to do it (e.g., a normative reason to avoid drinking a liquid whose toxicity you have no awareness of) and someone might have a motivating reason to do something without a normative reason to do it (e.g., a motivating reason to commit mass murder).

On some views, I have just made an error in associating wrong-making features with normative reasons (of a certain kind, against the act). Some might reject the common-sense thesis of "moral rationalism"—the thesis that for any given wrong act, there is a normative reason not to do it (Foot 1972). Such a rejection might be motivated by the conjunction of the claim that some acts we are definitely wrong, with a certain "internalist" theory of normative reasons, on which someone has a normative reason to act only if they have a *desire* or a *partial motivation* to act (Finlay and Schroeder 2017). Shooting innocent worshippers is probably necessarily wrong, but if the shooter has neither the desire nor the partial motivation to avoid shooting them, then, according to this view, the shooter does not have a *normative* reason to avoid shooting them. I find this view extremely troubling, however, for surely the shooter has a normative reason to avoid shooting them (in the fact, e.g., that they are innocent human beings). Thus, I think that we should reject such an internalist theory of reasons.[37] Accordingly, I presume the thesis of moral rationalism in this thesis.

Someone could also challenge this picture by arguing that the moral status of an act can itself be a normative reason for or against the act—that is, without reference to any right- or wrong-making feature. There are very good reasons for rejecting "rightness," "wrongness," or "oughts" as normative reasons, however. For P. Stratton-Lake (2001, 11–28), it is because normative reasons are "symmetrical" with the good-willed agent's *motivating* reasons, which would not be reasons of the rightness of the act, since the *fact that it is right* is not the reason *why* it is right. To insist otherwise, as Stratton-Lake plausibly argues, would be to dissolve the connection between morality and rationality. For J. Dancy, it is because *thin* normative concepts like "good" or "right" cannot be normative reasons, for:

> to say that it is good or right is merely to express a judgement about the way in which other considerations go to determine how we should act… The action's being good or right merely passes on whatever normative pressure is coming from below, without increasing that pressure. (2004, 16–17)

These look to me to be good reasons for denying that "right," "wrong," or "ought" can generate normative reasons, independently of right- and wrong-making features.[38]

[37] With John McDowell (1995), however, I am open to a version of counterfactual reasons internalism (if it is still a form of reasons internalism) on which to have a reason requires that you would be so motivated if you were a good willed or virtuous agent.

[38] Note that this is not to deny that they can constitute the reasons that the agent takes themselves to have—namely, *motivating* reasons (discussed below).

Thus, so far, we have the view that blameworthiness requires that the agent does something wrong, "all up," or having factored in all of the relevant normative reasons, where the relevant normative reasons are "moral" in the broad sense and represent the (other-regarding, self-regarding, etc.) types of value that determine wrongfulness.

3.3.2 Meta-Ethical Assumptions: Cognitivism & Moral Truth

It should be clear by now that I am assuming the meta-ethical position that there is such a thing as "moral truth" and that moral truth is a property sometimes enjoyed by moral propositions. It is, for example, *true* to say that the mosque shootings were wrong, or that I morally ought not to lie. Hence, I also embrace cognitivism, the view that moral sentences express moral beliefs which have a truth value (as opposed to emotions or "likes" and "dislikes"); and so, with the affirmation of moral truth, I deny error theory. These are both significant assumptions that I am afraid will have to left undefended in this thesis. The reason is that it is an interesting and challenging question in itself, what the epistemic condition for moral blameworthiness is, on the assumption of cognitivism and the denial of error theory.

I do, however, try to remain neutral on (i) whether some non-individualist form of *moral relativism* is true—according to which moral truths are indexed to groups, cultures, practices, or contexts ("true for us, and not for them")—and neutral also on (ii) whether *moral objectivism* or *realism* is true—according to which moral truths are *mind-independent* (e.g., as based on non-natural properties, or God's commands, rather based on the agent's beliefs, desires, intentions, hypothetical consent, etc.).[39] On both of these issues I have personal views, which may come out in what reasons I take agents to have in the examples I use, but I consider this project on the nature of responsibility and the epistemic condition for blameworthiness to be largely orthogonal to these issues. In §3.3.5, however, I will argue against a certain kind of non-objectivist, individualist, moral relativism.

3.3.3 Responsibility, Moral Dilemmas, & Alternatives

Earlier, I stated that when there is only one alternative to some wrong act, the alternative is either (or some combination of) right, permissible, obligatory, or supererogatory. Notice that I omitted "wrong" from this list. I did so for a reason. It is my view that whenever there is a wrong option, there is always a *right* alternative. Accordingly, so-called "tragic irresolvable dilemmas" (Hursthouse 1999, 72) are not dilemmas between wrong acts, but horrific or

[39] I am relying on R. Joyce's (2015) distinction between moral relativism and non-objectivism.

"terrible" acts. Not only is this claim relevant for later discussion, it is relevant for defending the claim introduced last Chapter, that responsibility for wrongdoing is materially equivalent with blameworthiness for wrongdoing **(**Rw↔Bw), and in particular, the claim that responsibility for wrongdoing entails blameworthiness (Rw→Bw).

Michael McKenna (2012, 20) has recently challenged (Rw→Bw) on the basis of the well-worn example of *Sophie's Choice* (Styron 1979). Sophie and her two children are prisoners in a Nazi concentration camp and Sophie is told by an evil prison guard that only one of her children is allowed to live. She is told that she must decide who lives by choosing which of her children must die. If she does not choose, the guard will kill both. In the end she chooses one of her children to die to let the other live. The relevant questions for our purposes are the following. Does she do what is wrong? Is she morally responsible for her choice? Is she blameworthy for it?

McKenna argues that whatever she chooses, she will do something *wrong* in doing so. This is because there are "nonoverriding moral reasons" against each option. But McKenna argues that she is still responsible, if only minimally, for her choice. She has *some* freedom over her choice, and she is aware of the moral significance of her conduct. On (Rw→Bw), this makes Sophie *blameworthy*. However, McKenna argues that she is surely not blameworthy. And I think that we should agree. Under the circumstances, it would have been excessively harsh to blame her for her choice. She deserves our empathy and pity, not our blame—and this holds true on a number of theories.[40]

McKenna's solution is to reject (Rw→Bw) and posit that blameworthiness in particular requires the satisfaction of an additional quality of will condition. In support of this, McKenna asks us to imagine:

> "Sophie's evil twin, Cruella, who is placed in the same situation, and who is absolutely delighted to be in this moral dilemma" (2012, 12).

If Cruella then delightfully chooses her daughter to die, McKenna has the intuition that she would be blameworthy for this choice. This, he thinks, demonstrates that blameworthiness for wrongdoing requires responsibility *plus* the display of ill will.

[40] Assuming an emotion-based theory of blame, e.g., the emotion of *indignation* (possibly experienced by an observer of the case) would seem inappropriate. Assuming the blame-as-faulting view that I defend in Chapter Five, Sophie cannot be morally at fault for her choice.

K. Fritz (2014) has argued that as a matter of fact, Sophie is not responsible, and so (Rw→Bw) is saved. For Fritz, certain "Strawsonian" excuses apply—for example "'I had to do it', 'It was the only way', [and] 'They left me no alternative'" (2014, 585). I am not persuaded of this move, however. Sophie has three options. She is not forced to make any specific choice (although she is certainly forced to choose from three horrific options). Her contribution is necessary for what happens. And I take it as especially revealing that it would make sense, long after what happened, for the saved child (the boy) to wonder whether there was anything "to" the fact that *he*, and not his sister, was saved. To me, this strongly suggests (minimal) responsibility.

What then is the solution? In my mind, it is to hold that she does not commit *wrongdoing* in the first place, and so (Rw→Bw) is saved, and **(**Rw↔Bw) is left untouched. Consider after all that it would have been appropriate for a friend to console Sophie, filled with deep sadness or regret afterwards, with the sentiment that she did "nothing wrong." With Rosalind Hursthouse (1999, 74-5), I deny that wrongdoing is committed in such a tragic irresolvable dilemma, even though every option is "terrible." But I do not deny it for Hursthouse's reason that wrongdoing requires blameworthiness (the inverse of the view defended in this Section),[41] but that irresolvable dilemmas involve no right alternatives and that wrong acts imply right alternatives. Once it is understands what I (and many others) mean by "wrong act"—as "the wrong thing to do" in the circumstances—I hope that this concern fades: there is no "wrong thing to do" in Sophie's dilemma, at least between choosing her son or daughter to die to save the other. Note also that it is not denied that such an act is "normally," "generally," or *prima facie* wrong.

3.3.4 The Suberogatory

Some acts are not wrong but are still, to some extent, morally bad or vicious. These are sometimes known as "suberogatory" acts, for the fact that they are supposed to be the inverse of the "supererogatory," standing in relation to the forbidden how supererogatory conduct stands in relation to the obligatory. Since some defenders of the existence of suberogatory acts argue that these acts can be *culpable* (Driver 1992; McKenna 2012, 187), this view therefore threatens BW. Consider J. Driver's case of the inconsiderate train passenger.

[41] I know of no one in the *responsibility* literature who argue that wrongdoing depends on blameworthiness. Plausibly the conditions on wrongdoing are more easily satisfied anyway: they do not include robust freedom, control, awareness or even self-disclosure conditions.

> [Suppose] that the train is almost full, and a couple wish to sit together, and there is only one place where there are two seats together. If the person ahead of them takes one of those seats, when he could have taken another less convenient seat, and knowing that the two behind him wanted to sit together, he does something [suberogatory]… [But t]he people who want to sit together have no claim against the person ahead of them in line. Thus, he has no obligation to pass up the more convenient seat. (1992, 286-7).

But the question is whether these acts are not mildly *wrong*. Wrongdoing can come in degrees. Acts can be *slightly* wrong, or horrifically wrong. Moral obligations can sometimes be quite demanding; and in relation to this case, we might think that the couple do not have a claim on him relative to our *social norms*, but that the story might be different relative to our *moral norms*.

Hallie Liberto rather poignantly criticises the category of the suberogatory by arguing that "it takes serious work in applied ethics, combined with some missing empirical data, to determine whether he acts wrongly, or permissibly, or if a new category is needed" (2012, 399-400). It turns out, for Liberto, that no new category *is* needed. Concerning the case above:

> if it turns out that the couple announced, before boarding the train, that this ride constitutes their final hour together before one member of the couple is shipped off to war then the train-rider's action is certainly impermissible. (p. 400)

On the other hand, if the couple has made it clear to the passenger that they are alright with him taking the seat, then it would seem permissible. This strikes me as a reasonable reply.

There are also theoretically conservative reasons for denying that suberogatory actions can be blameworthy, even if they exist. Consider that is virtually analytic to say that *permissible* acts are blameless, a claim that Driver and others must deny. In my view, then, I think that we have good reason to hold onto BW.

3.3.5 Wrong-Making Features & What is Factually (Not Morally) Apparent

It is time to consider an important debate about whether or not the agent's *moral* beliefs or credences—what is *morally apparent* to them[42]—makes a difference to the wrongfulness of their conduct (see §3.4 for a discussion of beliefs and credences). This question, and my thesis' guiding question, are not so unrelated. If, for example, culpable conduct is necessarily wrong and moral beliefs and credences can themselves constitute normative reasons why the agent should not do it, then there is an *indirect* sense in which blameworthiness depends on beliefs and credences concerning the act's wrongfulness, and so we would have a partial "yes" answer to the question of whether blameworthiness so depends. I argue, however, that what is *morally* apparent to the agent cannot (directly) determine what is wrong for them.

Suppose that Jo, a Christian, is (psychologically) certain that pre-marital sex is wrong but engages in it anyway. Even if it is denied that pre-marital sex is objectively wrong, some philosophers would hold that Jo still does something wrong (cf. B. Weatherson's [2014, 2019] "normative internalists" (about moral norms); and E. Harman's [2015] "Uncertaintists"). These philosophers—call them "moral internalists"—hold that the agent's beliefs or credences in the wrongfulness of the act can determine its wrongfulness. Disagreeing with these philosophers are perhaps the majority of philosophers who hold that there is no form of wrongness that depends (significantly)[43] on what is morally apparent to the agent (Weatherson's "normative externalists" about moral norms, or Harman's "actualists"). These theorists—call them "moral externalists"—would hold that Jo does something wrong only if pre-marital sex is actually wrong (as many conservative Christians believe), regardless of what she believes.[44]

There are two particularly poignant motivations for moral internalism (among others). The first is that moral internalism allegedly captures the need for action-guiding norms, given *moral uncertainty* about what to do—uncertainty based exclusively upon one's indeterminate moral beliefs or credences. Moral internalists give us are theories which reach into our uncertainty and offer action-guidance: "do what you believe is right," "do you what your

[42] With the language of "what's morally apparent," I do not mean to imply anything about the possibility of moral *perception*. I mean only what's morally apparent relative to one's moral *beliefs or credences.*
[43] E. Harman (2015), for instance, qualifies that moral beliefs or credences can play a distant, or indirect, role through influencing what factual beliefs or credences one comes to have.
[44] There is also the possibility of holding that there is a ("subjective") *sense* in which Jo does something wrong, regardless of whether there is *another* ("objective") sense in which it is right or wrong, independently of what she believes. The trouble with this view is that we want a way of arriving at the *one* thing that the agent *ought* to do (Zimmerman 2008, 7).

conscience tells you," or "do the least morally risky thing" (Guerrero 2007; Moller 2011; cf. Weatherson 2019, 1–30). A second motivation for moral internalism is that it captures the intuitive wrongfulness of *moral recklessness* (that is, acting the way that is most morally risky relative to one's moral uncertainty). It is objectionable because moral recklessness appears to gamble morality needlessly or display disregard for the avoidance of wrongdoing (Moller 2011, 435-6).[45]

Notwithstanding its advantages, I regard moral internalism as mistaken. First, the view has embarrassing consequence in cases where the moral internalist's prescriptions are to do what is morally heinous (as even other kinds of normative internalists have acknowledged: Sepielli 2017; Geyer 2018; Bykvist 2014). Consider K. Bykvist's (2014) case of the *egoist*, who believes that egoism requires of him to torture babies, and acts accordingly. Suppose also that this belief is subjectively rational, in the sense that it is supported by his beliefs and credences, and that the egoist believes with *certainty*. Any form of moral internalism would then entail that the egoist *morally obliged* to torture babies. But that seems outrageous.[46]

A second reason against moral internalism—or at least against "hedging" internalism which recommends moral caution in the fact of uncertainty (Geyer 2018)—takes its cue from William James (1896, VII). Plausibly, if everyone were to guide themselves in accordance with hedging principles, then genuine progress in moral understanding (and thus moral life) would be objectionably rare. Often moral recklessness (doing the most risky thing, relative to the weight of one's moral credences) is required for genuine progress in moral understanding, and I would wager that much of our progress in moral thinking has come about only because people were willing to take moral risks even when appreciative of the less morally risky nature of conservative alternatives.

I am inclined also to think that moral internalism actually fails to provide cherished action-guidance, given that we do not instantly or even upon reflection find it obvious what these principles are *when* morally uncertain. This is not surprising, given ongoing debate about what the correct principle is.[47] Finally I do not believe that we need to account for moral recklessness as *wrong* in order to do justice to its moral badness. We can describe it, for example, as *morally unconscientious*, *inconsistent*, *hypocritical* or even (as I will argue)

[45] These two advantages to moral internalism are outlined in Weatherson (2019, 1-30).
[46] See also E. Harman's (2015, 59ff.) more subtle case of the sexist Bob who refuses to teach his daughter to drive.
[47] See for a good review: K. Bykvist (2017).

potentially inculpatory when actual obligations are violated. The cost of deeming it wrong is just too great, on my view.

It does not follow that what is *factually apparent* to the agent—that is, what would be wrong if the agent's *factual* beliefs or credences were true—is not wrong-making. If you were (psychologically) certain that that the sugar jar had arsenic in it and you were to put sugar into my tea, you would have done something wrong, even if in reality you only put sugar into my tea. Of course, much of the time, what is factually apparent *is* correct, but sometimes it is not, and when it is not, plausibly your obligations are to do what your factual beliefs or credences, suggest that you ought to do. Consider a thought-experiment in favour of this view (adapted from M. Zimmerman, originally from F. Jackson 1991, 462–63):

> Jill, a physician, has a patient, John, who is suffering from a minor but not trivial skin complaint. In order to treat him, she has three drugs from which to choose: A, B, and C. Drug A would in fact be best for John. However, Jill [is psychologically certain] that B would be best for him, whereas the available [factual] evidence indicates that C would be best for him. (Zimmerman 2008, 6)

What drug is Jill overall required to prescribe? It seems obvious that it is Drug C. *For all her factual evidence* (i.e., her factual beliefs and credences), Drug C is best, and so she should prescribe Drug C. To fail to give John Drug C is to deliberately act contrary to her best factual evidence, which seems highly objectionable. What is different about this case, though, is that Jill is psychologically certain that Drug B is best. So what should win out in a duel between *factual evidence* on the one hand, and *psychological moral certainty* on the other? Factual evidence is surely the answer, for surely the doctor should proportion her degree of belief to the evidence. And since moral internalism would recommend that Jill prescribe Drug B (given her moral certainty), this further demonstrates the counterintuitive nature of moral internalism. Drug A is of course "objectively" correct, but Jill does not have access to why that is the case. For reasons of what is actually action-guiding, then, she ought overall to prescribe Drug C.

What, then, regarding our claim BW about what is required for blameworthiness? BW, combined with my argument in this Subsection, entails that one can be blameworthy for those acts that are wrong relative to what is factually apparent. But one cannot be blameworthy for acting contrary to what one's strictly moral beliefs or credences prescribe. However, some (e.g., Zimmerman 1997a; Haji 1997, 526; Capes 2012; Guerrero 2007, 80) would want to put

pressure on the way that I have developed BW, by arguing that *even if* there is no wrongdoing relative to what is morally apparent, the agent can still be *blameworthy* for acting contrary to what is morally apparent to her (e.g., to what she *believes* to be wrong). In reply, however, it seems to me that many of the cases offered in support of this view (e.g., in Capes 2012) involve wrongdoing relative to what is *factually* apparent to the agent *as well* as what is morally apparent, and so much of its intuitive support is diminished. But for the few cases in which the agent's moral beliefs/credences do not align with their *factual* beliefs/credences and yet they do the thing prohibited by their moral beliefs/credences (e.g., a case in which Jill chooses Drug A while mistakenly certain that Drug B is best and Drug A is worst), I am not moved by these theorists' intuition of blameworthiness. Invoking comments that I made above, I think that these theorists confuse the concept of blameworthiness with something akin to moral unconscientiousness, inconsistency, recklessness, or hypocrisy. Indeed, as I argued above, sometimes moral recklessness is instrumental to genuine moral progress.

3.3.6 Blameworthiness, Wrongdoing, & Frankfurt-Style Cases

I have been defending the view that:

> **BW:** Blameworthiness for actions (and omissions) requires that they are *all-things-considered wrong*.

In §3.3.3 I argued that a wrong option implies a right option, and that in tragic irresolvable moral dilemmas, terrible deeds are not wrong, precisely because there is no right alternative. Notice, however, that if this is the case, and if BW is true, then blameworthiness for an act requires that there was a right alternative. There is, however, a very strong intuition that one can be directly blameworthy for an act, despite the lack of alternatives.

This was, of course, H. Frankfurt's (1969) key insight—where "the lack of alternatives" is understood as a lack different outcomes or courses of action. To take an adapted version of his central example, suppose that Jones deliberates about whether or not to shoot Smith, knowing that it would be wrong to do so. Suppose, however, that a mad neuroscientist, Black, is lurking in the background and would intervene to ensure that Jones kills Smith if Jones were to show any sign of faltering. Black would intervene only if Jones falters because Black would rather Jones did it himself. It turns out that Black does not have to intervene, because

Jones shoots Smith on his own, for his own reasons. On Frankfurt's view, Jones is fully responsible and blameworthy for shooting Smith.

My take on this case is that when deliberating, Jones does not have the "robust" alternative of refraining from killing Smith—or at least, the ability to prevent the *alternative outcome* of not killing Smith—however, I do not think that the case rules out what we might call a "micro-alternative" (e.g., *trying* to do otherwise).[48] But suppose that we could tell a story in which Jones does not even have a micro-alternative. My intuitions still align with Frankfurt's original intuitions: Jones can still be blameworthy for killing Smith. And yet because Jones has no alternative (not even a micro-alternative), it follows from the claim that I defended above—that for every wrong act there is a right alternative—that either wrongdoing is not necessary for blameworthiness or that wrongdoing does not require right alternatives.

My reply is that this argument neglects the distinction between *actual* and *apparent* alternatives. For Jones to deliberate about whether or not to shoot Smith, it has to *appear* to Jones that the alternative of not shooting Smith is genuinely available to him (i.e., he has to have a belief or a non-negligible credence with the content that he has an alternative, or in propositions that strongly suggest that he has an alternative—e.g., "there is no one controlling me"). As I argue in Chapter Seven, deliberation requires apparent alternatives. But consider that whether or not one has an alternative is strictly a *factual* matter, rather than a normative matter. Although mistaken, Jones' factual take on the situation is that he has an alternative (i.e., not shooting Smith). But *if* it is apparent to Jones that he has an alternative, then the door is opened to the verdict that shooting Smith is genuinely *wrong*, while refraining from doing so is *right*, but right or wrong *relative to Jones' factual beliefs or credences*. Thus, when Jones shoots Smith, he does what is wrong, even though it is not wrong from an external perspective (due to lacking actual alternatives). Thus, BW is vindicated.

I have argued in this Section that blameworthiness for conduct requires that the conduct is all-things-considered wrong, that conduct is wrong by virtue of wrong-making normative reasons, that wrongdoing requires right (apparent) alternatives, and that wrongdoing is sensitive to what is factually apparent to the wrongdoer.

[48] See A. Mele and D. Robb's (2003) concession that there are still what I call micro-alternatives (such as the alternative of Jones' not shooting Smith on his own) in Frankfurt-style cases.

3.4 Wrong-Sensitive Beliefs, Credences, & Motivating Reasons

The guiding question is the question of whether blameworthiness for actions or omissions requires that agent has beliefs or credences concerning the act's moral significance. Given that I have just argued that one can be blameworthy for actions and omissions only if they are *all-things-considered wrong*, we can now reformulate the question as the question of whether blameworthiness for actions or omissions requires that the agent has beliefs or credences concerning the act's all-things-considered wrongfulness. Accordingly, let us call these beliefs/credences concerning wrongdoing, "*wrong-sensitive* beliefs/credences" (where the "-sensitive" implies that the wrong-related content of these beliefs or credences is actually *correct*). Now I hope that some sense can already be made of this notion wrong-sensitive beliefs/credences, but it is time now that we address them more directly.

3.4.1 Wrong-Sensitive Beliefs & Credences

I propose that an agent has wrong-sensitive beliefs/credences concerning some act (or omission) if, and only if, (i) the act is wrong, and (ii) the agent has beliefs/credences whose content either includes the wrongfulness as such, of the act, or the features of the act that make it all-things-considered wrong (regardless of whether the content includes *that* they are wrong-making). Clearly condition (i) tracks the distinction between the status of wrongfulness and normative reasons *why* an act has that status (the wrong-making features; see §3.3.1). Importantly, beliefs/credences in wrong-making features need not include the fact that they make the act wrong for them to be wrong-sensitive; if the agent denies that a feature is wrong-making but believes that the act has that feature, the belief is still wrong-sensitive.

So what beliefs/credences satisfy condition (ii)? Given the distinction between the act's wrongful status and its wrong-making features, the answer should be clear. Beliefs/credences exclusively concerning the wrongful status of the act may include any of the statuses that I am taking to be equivalent to "all-things-considered wrongdoing" (see §3.3.1). They include beliefs that the act is morally "forbidden," "prohibited," or "impermissible," or the beliefs that it would be "right," "required," or "obligatory" to do something else. Since these beliefs pass a moral verdict on the act, they constitute what are sometimes called "verdictive judgments" (Stratton-Lake 2001, 14; Dancy 2004, 16). Beliefs/credences whose content involves only wrong-making features may include any of the features that are in fact wrong-

making. Consider, for example, "he has the right to life," "that would be inconsiderate," "it would adversely affect the environment," or "she is telling me something important." These are sometimes called "evidential judgments." Although these examples include only *evaluative* reasons (evaluative—in the sense that they have thick evaluative content), these beliefs/credences can include purely *factual* or non-evaluative reasons, such as the reasons "she has not consented," "it would cause them to feel pain," "we do not have time," and so on. Finally, beliefs/credences whose content includes both the wrongful status *and* the reasons why, include examples like "his right to live makes it wrong to kill him," "I morally ought not to be inconsiderate," "when she is telling me something important it is right to listen to her," and so on. All three types of beliefs/credences can obviously be held together, constituting a two-step practical syllogism with the verdictive judgment as the conclusion.

Some terminology used in the literature applies to these different forms of wrong-sensitive beliefs/credences. A belief/credence is (morally)[49] *de dicto* just if it includes a moral status in its content. It is *morally de re*, however, only if it excludes this content. Thus, beliefs/credences in the act's wrongful status and beliefs/credences that include both the act's wrongful status and its wrong-making features, are *morally de dicto*. Beliefs/credences which include only the wrong-making features are *morally de re*. The distinction originally comes from Michael Smith (1994, 71-76) and is used by various others (notably, Arpaly 2002), in distinguishing "*de re* concern about morality" from "*de dicto* concern about morality." The former is concern about what is in fact morally significant (e.g., telling the truth, giving to charity), whether or not the agent considers it as such, while the latter is concern about morality as such (e.g., doing the "right" thing). Accordingly, *de dicto* beliefs/credences about wrongdoing are about wrongdoing "*so conceived*" (Rudy-Hiller 2018), while *de re* beliefs/credences about wrongdoing are about the features that, in fact, make the act wrong, whether or not the agent realises that they do.

3.4.2 Beliefs, Credences, & Motivating Reasons

It is now time to turn our attention directly to the nature of beliefs/credences. In doing so, I will also discuss the important concept of *motivating reasons*.

I begin with the claim that beliefs/credences in question are *propositional*, rather than, for instance, *objectual*. This is the claim that they have *propositions* as their representative

[49] This use of de re/de dicto is distinct from other uses of the de re/de dicto distinction applied to beliefs/credences—e.g., the distinction between *de re* believing (that *p)* and *de dicto* believing (that *p is true*) (Gallois 1998).

content, which include subjects, copulas, and predicates, such as "eating fries each day [subject] is [copula] unhealthy [object]." This is opposed to non-propositional representative content such as the perceptual impression of "the blueness of the sky" or "what it feels like to be blamed." In the case of beliefs, propositional beliefs are also to be distinguished from beliefs *in* things (e.g., people)—from having faith, trust, or practical commitment in the object—however sometimes I will use the term "belief in" simply as a quickhand for "belief that."

In simple terms, to *believe* that *p* is to have a strongly held attitude that *p* or that *p* is "true" or that *p* "the case," and to have a *credence* in *p* is to have a "degree of confidence" in *p*, however, strong or weak. Beliefs and credences are clearly very similar, and both play similar (and sometimes competing) roles in epistemology and in ethics. What more can we say?

First, it is well accepted that beliefs/credences "come in degrees." Precisely how this should be analysed is subject to much debate. I adopt the controversial (though, for our purposes, immaterial) notion that beliefs themselves are "flat-out," "binary," or "all-or-nothing" (cf. G. Harman 1986, 22f.), and thus that beliefs only come in degrees *secondarily* in the sense that they can be held with a greater or lesser corresponding *degree of confidence*—or credences. Thus, I take credences to be the notion that is *primarily* scalar, rather than beliefs. Credences are sometimes identified as "degrees of belief," but by "degrees of belief" I shall mean "beliefs held with certain degrees of confidence." Thus, I do not use "degree of belief" to refer to credences that fall below what is required for flat-out belief (e.g., 0.2 credences); plausibly, beliefs require a certain minimum, indeed relatively high, credence threshold (probably above a confidence of 0.5).[50] The threshold degree of confidence for belief is not, however, *sufficient* for actual belief, for I can have 0.999 confidence that I will lose the lottery and yet still not *believe* that I will lose in the hope that I will win (Schwitzgebel 2019). A 100% confidence that I will lose seems sufficient, however, for the belief that I will lose.

A 100% confidence is, of course, the highest degree of confidence, and it is equivalent to *complete certainty* that *p*—or complete "psychological" or "subjective" certainty, as opposed to objective, certainty which requires the truth of *p* (Reed 2008). *Un*certainty whether *p* is technically speaking the *lack of certainty* that *p* (and so is compatible with belief to a high confidence that *p*), but I should note that it is common in the literature on the epistemic

[50] I am open, however, to the view that this threshold may well be context-sensitive (cf. Hawthorne and Bovens 1999, n. 7).

condition for moral responsibility to find that uncertainty whether *p* means a "middle-of-the-park" confidence that *p*, somewhere from about 0.3 confidence to about 0.7 confidence that *p* (and corresponding confidences in *not-p*).

There is some debate about what statements are best used to *express* credences, rather than *report* them. Consider beliefs. We *report* the belief that Covid-19 kills by saying "I *believe* that Covid-19 kills" while we *express* this belief by saying "Covid-19 kills." Following A. Sepielli (2012, 2017; cf. also Yalcin 2007), I think it is plausible to say that credences work in a similar way. A 0.6 credence in the fact that Covid-19 kills is *reported* by a statement like "I am more confident than not that Covid-19 kills," but, lest credal reports and expressions are collapsed, the credence is *expressed* by another kind of statement—"Covid-19 probably kills" or "there is a minimal probability of 0.6 that Covid-19 kills." Sepielli argues that these are not statements of objective probabilities (based on facts out in the world), and neither are they statements of subjective probabilities (probabilities according to one's credence values) but statements of "minimal" or "epistemic" probabilities (probabilities based on one's evidence). Yalcin defines an epistemic probability as the "the objective degree of confirmation a body of propositions confers on a given proposition" (2007, 1021). Credences are ideally rational insofar as they accurately conform to the epistemic probabilities (cf. Sepielli 2012, 49-50). They are the credences "that the agent epistemically ought to have, given their evidence" (Geyer 2018, 403).

I hope it is now clear why I have put the thesis' guiding in terms of beliefs *and* credences. There are good reasons not to conflate them.

A few further features about beliefs/credences are worth mentioning. Beliefs/credences involve complex or "multi-tracked" dispositions to think, feel, assent to, or "accept" their contents (Schwitzgebel 2002; Cohen 1992, 4). The reason is straightforward: if someone was alleged, for example, to have a belief that *p* but was not disposed *in any way* to think or act on *p* (even under circumstances in which *p* had been raised for reflection), we would naturally question whether they believed that *p*. Dispositionalists about belief elevate this feature as the defining feature of beliefs, but even representationalists (who hold that beliefs are "stored in the mind") involve dispositions to do or think things (Schwitzgebel 2019).

Second, epistemological orthodoxy says that beliefs/credences are not under direct voluntary control (Alston 1988). This thesis about beliefs is dubbed "doxastic involuntarism" (Steup 1988, 73), and it entails that we cannot form a belief "at will" or as the outcome of an immediate decision or intention to believe. We can only have "indirect" control over beliefs/credences, control over them *via* inquiry or the activity of belief/credence

management (reading a book to inform yourself, checking the facts, evaluating possible answers, trying to remember what happened, etc.). Indirect control over beliefs/credences is commonly taken to require *foresight* of their formation or retention, akin to our indirect control over the formation of our character or health (Alston 1988, 275ff.; Peels 2017, 66–67). But more commonly we *influence* our beliefs/credences by doing something necessary for their formation or retention (Alston 1988, 276ff.) without foresight of them (Peels 2017, 66-67).

Third, beliefs/credences plausibly have conditions under which they are *justified*. This may mean justified, relative to the attempt to attain the *epistemic* good of acquiring *true* beliefs or accurate credences (and maybe other epistemic goods like "knowledge"). Or it may mean justified relative to some non-epistemic *prudential* or *moral* value (e.g., getting through cancer or being loyal to a friend).[51] We have already suggested a natural account on which credences are justified—namely, when they conform to the epistemic probabilities. A similar account of the conditions under which a *belief* is epistemically justified makes it a function of the belief's fit with, or basis in, the *evidence* (viz., evidentialism; Conee and Feldman 2004; cf. Clifford 1886). Its main opponents are those who hold that epistemic justification is a matter of the belief's *being the product of a reliable process*, regardless of one's evidence (e.g., "reliabilists" [Sosa 1980], or "proper functionalists" [Plantinga 1988]). Suffice to say that there is some property (or more than one) enjoyed by beliefs/credences that makes them either *sufficiently likely to be true*, or at least *permissible* with respect to the value of acquiring true beliefs or accurate credences (Steup and Neta 2020). More relevantly for our discussion, those who take the latter, more "deontological," conception of epistemic justification sometimes conceive of it as epistemic *blamelessness* (cf. "weak justification" in Goldman 1988, 52) or treat the notion of "blameless belief" independently of the notion of justification, suggesting that *blame* can be merited for a failure to meet this standard (Nottelmann 2013; *pace*, e.g., Kauppinen 2018).

An important feature about both beliefs/credences—related to their dispositional nature—is that their *contents* can be either "occurrent" or "dispositional" (Rudy-Hiller 2018; Schwitzgebel 2019). An occurrent belief/credence that *p* is a belief that at a given time manifests in conscious thoughts or feelings that *p* (or that *p* "is true" or "is the case") at that time. Occurrent beliefs/credences are such that the self-conscious or self-aware agent can

[51] For the example of friendship warranting a (potentially epistemically unjustified) belief in support of your friend, see Keller (2004).

report them with utterances like "I'm thinking that *p*" or "it feels that *p*." By contrast, a dispositional belief/credence that *p* is a belief/credence that, at a given time, does *not* manifest in thoughts or feelings that *p* but is disposed to so manifest at other times. Before I brought it to reflection just now, I surely believed "my wife is beautiful," even though I was not then thinking it. The belief/credence that *p* may also be "tacit," such that it has never been thought before (Peels 2017, 32–33). Up until now the fact that "Kazakhstan is north of New Zealand" never crossed my mind, but I am sure that I *believed* it. Notice that these are not different *kinds* of belief/credence, but different *ways* in which the propositional object of one's belief/credence presents itself to the agent.

A related distinction between different dispositional beliefs/credences has been made between *accessible* and *inaccessible* beliefs/credences. Accessible beliefs/credences are I think best understood as beliefs/credences of which the agent can *come to be aware by reflection* (Pappas 2014). Sometimes this can happen instantaneously (e.g., consider the answer to 4+8); other times, it can take time (e.g., through unearthing suppressed memories). The kind of accessibility that (I believe) is relevant for blameworthiness will be discussed in Chapter Seven.

We have reached the end of our sketch of wrong-sensitive beliefs or credences. Before we draw this section to a close, however, we should discuss one more related concept—indeed, an extremely important concept for my later argument. This is the concept of *motivating reasons*. Motivating reasons find their home in this part of the conceptual terrain, for motivating reasons are widely defined as reasons that *the agent takes herself to have* for or against some action, and in light of which they act or omit (Dancy 2000, 98-120; Stratton-Lake 2001; Alvarez 2016; Levy 2009; Robichaud 2014). What is it for an agent to "take" herself to have a reason? It is commonly regarded to be about *believing* something (Dancy 2000; Alvarez 2016; cf. Levy 2009, 735). But plausibly one may take oneself to have a reason if one would express a *non-negligible credence* as a reason. (Consider two examples: 1. If the expression of a 0.6 credence in "Covid-19 kills" is "it is more probable than not that Covid-19 kills", and the latter does not express a belief, then it appears that a credence (or its expression "Covid-19 probably kills") may constitute a motivating reason not to (say) expose yourself to someone else if you have the virus. 2. As we noted above, a 0.999 confidence that I will not win the prize need not entail a *belief* that I will not win, but surely the expression of that credence—"it is vastly improbable that I would win"—could constitute a motivating reason not to enter the draw.) Thus, I propose the following working account of motivating reasons:

M: An agent has a *motivating reason* for or against an act at some time *t*, *if and only if*, at *t*, (i) they believe or have a non-negligible credence in *p*, (ii) they would, in principle, express their belief or credence in *p* in answer to the question of why they should or should not act (or an equivalent question), and (iii) this belief/credence is capable of featuring in the explanation for why they act (or omit).

If the agent then acts, they would then express that belief or credence as the reason "in the light of which" they acted (Alvarez 2016). Several features of motivating reasons are worth mentioning. There need not be a belief/credence in the fact that the reason counts *as* a reason; it is enough that the agent would express the belief/credence when asked why they should do or not do something (or an equivalent question—e.g., when asked for their reasons). The belief/credence need not be something that the agent takes to *justify* their actions, for one may still have a motivating reason to do something one knows to be wrong all-things-considered. The belief/credence can be dispositional, but it must be accessible, for otherwise one could not express it in answer to the relevant why-question. (In this thesis, when the beliefs/credences involved are occurrent, I say that the motivating reason is "explicit"; when they are dispositional, I say that the motivating reason is "implicit.")[52] A further feature: the belief/credence that *p* need not be true; and if it is belief, it need not amount to knowledge that *p*. I follow Dancy (2011) here, and others in the epistemic condition literature (Levy 2009; Robichaud 2014), against those who require knowledge (e.g., Hyman 2011) for the possession of a motivating feature. This feature of possible falsehood poses a challenge for defenders of "non-psychologism" about motivating reasons, according to which motivating reasons are the "facts" represented in the *expression* of the relevant beliefs/credences, rather than the beliefs/credences themselves—as on "psychologism" (Alvarez 2016). (On this debate I remain neutral, which is why I express M in terms of material equivalence, not identity.) Also, I add the qualification, "in principle," in order to dismiss the relevance of circumstantial factors that prevent the expression of one's beliefs/credences in answer to the relevant why-question (e.g., misunderstanding, mental fatigue, or cognitive disability).

[52] Some treat the explicit/implicit distinction as distinct from the occurrent/dispositional distinction (at least as applied to beliefs), on the grounds that the former concerns whether or not there is representational content in the agent's mind (G. Harman 1986, 13–14; Schwitzgebel 2019). There are two reasons I treat these distinctions as synonymous: (1) it does not matter for blameworthiness whether or not someone's belief has representational content in their mind, at least independently of their accessibility; (2) in the epistemic condition literature (Levy 2009; Robichaud 2014) the implicit/explicit distinction has been applied more readily to *motivating reasons*, than the occurrent/dispositional distinction.

Concerning the content of motivating reasons, an action's wrongful status can be a motivating reason for the agent, but it can never be a normative reason (as we have seen), because normative reasons are the facts that *determine* the act's wrongful status in the first place.

Motivating reasons should be distinguished from *explanatory* reasons, or reasons that actually explain or cause one's behaviour (e.g., one's desires, intentions, emotions or character) (Alvarez 2016). For example, Jennifer might take as a motivating reason to go to church that she wants to deepen her relationship with God, when what she really wants to do is to see friends. However, motivating reasons may be cited as part of explanatory reasons (e.g., Jennifer may also cite her need to socialise as a motivating reason). And it is plausible that a motivating reason must be capable of constituting (if only partially) an explanatory reason, just as beliefs/credences must be capable of disposing people to act.

3.4.3 Awareness & Ignorance

We now have a clear picture of the wrong-sensitive beliefs/credences concerning which I ask the guiding question. Before wrapping up the Chapter, however, it would be worth briefly commenting on the concepts of *awareness* and *ignorance*, for they are common terms in the epistemic condition literature. Fortunately, these states are best analysed in terms of beliefs (and sometimes their epistemic justification), but they crucially also involve the concept of *truth*—namely, that which (according to the correspondence theory of truth assumed in this thesis) corresponds to reality. Beliefs can be *true*, and credences *accurate*. A belief that *p* is true, or my credence in *p* is accurate, just if *p* corresponds to some fact in reality. With the addition of the concept of truth, what can we say about awareness and ignorance?

It is clear that to *know* that *p* is to be *aware* that *p*. We often use these terms interchangeably. It is also clear that awareness that *p* requires that *p* is *true*. But we might ask whether something weaker than knowledge can count as awareness. Could *epistemically justified true belief* constitute awareness?[53] Could even *true belief* constitute awareness? Could even an *accurate credence* constitute minimal awareness? It would be worth noting that philosophers of the epistemic condition actually take awareness—or at least awareness of the relevant "kind" for blameworthiness—to be mere *true belief* (Rosen 2008: 596f.; Rudy-Hiller 2017, 400; cf. Peels 2010, 60). As a matter of fact, I think that we can go further and say that minimal awareness can be constituted merely by an *accurate credence*. However,

[53] Knowledge used to be thought of as epistemically justified true belief, but nowadays, following E. Gettier (1963), the vast majority think that epistemically justified true belief is not sufficient for knowledge.

nothing for this thesis actually hangs on this issue. If it turns out that acting contrary to a mere true belief in wrongdoing may be blameworthy, but that true belief in wrongdoing does not count as *awareness* of wrongdoing, then blameworthiness sometimes does not depend on awareness but true belief. Indeed, there is a plausible argument for thinking that lack of *knowledge* of some wrong-making feature is no excuse for performing wrongdoing if one still *truly believed* in that feature (cf. Rosen 2008, 596-7). Suppose that Dylan stole Daisy's smartphone after being told by Dasher that the smartphone on the table was Daisy's. Since it really was Daisy's, Dylan acted on the true belief that it was hers. But suppose that Dylan's true belief did not amount to knowledge, because Dasher had a reputation for "bullshit"[54] and was also, on this occasion, bullshitting Dylan by telling him it was hers in order to see what Dylan would do if he thought it was Daisy's. Is Dylan excused because he lacked *knowledge* (even *justified* true belief, *pace* Ginet 2001, 275; Timpe 2011, 20) that the phone was Daisy's, even though he truly believed that it was? It hardly seems so. If Daisy caught him in the act and demanded an explanation, Dylan's retort, "but I didn't *know* it was yours; I just believed it was," would not cause her to revise her blame. Now with Rosen and Rudy-Hiller, I think that true beliefs *do* constitute awareness, and so it makes sense to me to say that Dylan *was* aware that the phone was Daisy's. Accordingly, I will use the term awareness for true belief (or credence). But in the event that I am mistaken—that is, that awareness requires *more* than true belief—then every such reference to "awareness" should just be translated accordingly (to true belief or credence).

The fact that the conditions on awareness are contested suggests to me that it is not the fundamental epistemic state with respect to which we should describe views on the epistemic condition (*pace*, e.g., Rudy-Hiller 2018). Relatedly, it is also not the epistemic state with respect to which we should characterise the culpability internalist/externalist distinction, for some theorists (as we saw in §3.3.6) make beliefs/credences concerning wrongdoing necessary for blameworthiness, but do not take *true* beliefs/credences concerning wrongdoing necessary. Yet surely, in virtue of these theorists' requirement of otherwise identical epistemic states, they belong in the culpability *internalist* camp along with those who require true beliefs/credences.

Regardless of how awareness is defined, *ignorance* is plausibly the lack of awareness. Thus, in this thesis, I will take it to mean *the lack of true belief* (cf. also Rosen 2008, 596f.; Guerrero 2007, 62-3; Peels 2010, 60; Rudy-Hiller 2017, 400). Since ignorance involves the

[54] Bullshitting by asserting something involves carelessness about whether is true (Frankfurt 2005).

lack of true belief, there are two important kinds of ignorance:

1. *Disbelieving ignorance*: the agent lacks the true belief that *p* and believes that *p* is false, or that *not-p*.
2. *Suspending ignorance*: the agent lacks the true belief that *p* but also lacks the belief that *p* is false or that *not-p*.

Disbelieving ignorance is the "strongest" form of ignorance: if the agent has considered whether *p*, it involves the agent's "taking a stand" on whether *p* by believing it false or by believing its opposite, *not-p*. Suspending ignorance is weaker: if the agent has considered whether *p*, it is the kind of epistemic position occupied by those who are "on the fence."[55] We will have occasion to return to this distinction.

3.5 Conclusion

The guiding question of this thesis is whether, and if so how, the agent's blameworthiness for conduct depends on the agent's having beliefs/credences concerning their conduct's moral significance. After this Chapter, we can sharpen this question. Culpable conduct is always wrong all-things-considered, and so the relevant question is whether, and if so how, blameworthiness for (all-things-considered wrong) actions or omissions depends on the agent's having beliefs/credences that are sensitive to the all-things-considered wrongfulness of their actions or omissions. In short, it is the question of whether, and if so how, culpable conduct depends on wrong-sensitive beliefs/credences. In the next Chapter, we will see investigate one prominent answer about the way that it so depends—*volitionism*, according to which blameworthiness depends on the agent's occurrently believing that the act is all-things-considered wrong, and for the normative reasons why it is in fact wrong, either at the time of the act or at some point in its causal history.

[55] Note that neither entails that one has considered whether *p* before. This is plausibly because one can believe that *p* without ever having considered whether *p* (see above). Accordingly, R. Peels (2010, 62) has divided ignorance into at least *four* different kinds, involving forms of ignorance involving the conjunction of 1 or 2 *and* the further condition on each that one has considered whether *p* before.

Chapter 4
The Regress Argument and Responsibility Revisionism

4.1 Introduction

In the last two Chapters, I set out the rudiments of the key question to which this thesis attempts an answer. And last Chapter, we saw that the question is best interpreted as the question of whether, and if so in what way, an individual's moral blameworthiness for her all-things-considered wrongdoing depends on the individual's beliefs and credences concerning the wrongfulness of her conduct (whether its wrongful status or its wrong-making features). In this Chapter, the widely discussed answer to this question that we introduced in Chapter One—called "volitionism"—is raised and is given *prima facie* support. When combined with some plausible premises about how easy (or hard) it is for their conditions to be met, volitionism strongly suggests that we ought to revise our ordinary ascriptions of blameworthiness and responsibility. This is, of course, the key problem that motivates our wider investigation. To be fair, the problem is not necessarily seen as a "problem" by its volitionist defenders but rather as an *argument* for revisionism about responsibility. Accordingly, and given that volitionism requires a kind of *regress* when tracing culpable ignorance, I will call the core argument for volitionism the "Regress Argument" (following Rudy-Hiller 2018 and Wieland 2017).

Chapter One introduced the Regress Argument and its revisionist implications in a rough-and-ready form with the help of the illustration of the Józefów Massacre. In this Chapter, I outline the argument more formally, together with the two kinds of revisionism that have been associated with it; I explain each of the steps in greater depth; and I tease out the most plausible reasons given for each premise. I will also assume the understanding of the argument's key concepts (e.g., blameworthiness, wrongdoing, and belief) as developed in Chapter Two and Three.

The structure of the Chapter is as follows. In §4.2, I provide an initial formulation of the Regress Argument and the two revisionist options taken by its defenders. In §4.3, I explain and motivate the first premise in the argument. In §4.4, I explain and motivate the second and third premises in the Regress Argument, and then the inference to the main volitionist conclusion. In §4.5, I outline and motivate the main two revisionist options that volitionists have taken. And I then conclude the Chapter with a Section (§4.6) on the significance of responsibility revisionism.

4.2 A Formulation of the Regress Argument & Its Revisionist Implications

The Regress Argument is defended by M. Zimmerman (1997b, 2008), G. Rosen (2003, 2004, 2008), N. Levy (2011), and probably C. Ginet (2000).[56] Between them, Zimmerman, Rosen, and Levy draw two kinds of revisionist conclusions from the argument, forming overall the *rarity* argument (Zimmerman and Levy), and the *epistemic* argument (Rosen).[57] The conclusion of the former is that blameworthiness is *rarer* than many think, while the conclusion of the latter is that blameworthiness is more difficult *to ascertain* than many think.

Let me formulate both the Regress Argument and the revisionist options as follows:

The Regress Argument

(1) An agent S is blameworthy for performing all-things-considered wrongdoing *A* only if S does *A* contrary to S's true occurrent belief that *A* is wrong all-things-considered, based on true occurrent beliefs that *A* has the features that make it wrong (i.e., *A* is "fully advertent" wrongdoing), or S does *A* in/from culpable ignorance of *A*'s all-things-considered wrongfulness (i.e., *A* is "culpably unwitting" wrongdoing).

[56] Note two complications: (i) Later Zimmerman (2017) defends a slightly different version of the Regress Argument from the one that I will present, narrowing its scope to a particular kind of culpability, and altering the premises slightly by allowing direct culpability for "willing" wrongdoing "in" ignorance. When setting out and motivating the Regress Argument as I present it, I will be drawing mainly upon Zimmerman's earlier (1997b; 2008) work, unless otherwise stated. Accordingly, "Zimmerman" by itself will refer to Zimmerman in that earlier work. (ii) Ginet does not explicitly require an occurrent belief in wrongdoing (as the others do) but a belief that some act/omission will cause *harm* (which is, to me, a *wrong-making feature*; following Rosen 2008, 593). However, in every other respect Ginet appears to fit the description of volitionism, he cites Zimmerman in agreement that directly culpable conduct involves "no ignorance" (2000, 275), and he has been interpreted as a volitionist (in Rudy-Hiller 2018).

[57] Ginet (2000) does not draw out any revisionist implications.

(2) S is culpable for the ignorance in/from which S does *A* only if S's ignorance is the foreseen upshot of performing a "benighting" act or omission *B* for which S is blameworthy.

(3) S is blameworthy for performing a benighting act *B* only if *B* is either fully advertent or culpably unwitting wrongdoing.

Therefore,

(4) An agent S is blameworthy for performing all-things-considered wrongdoing *A* only if *A* is fully advertent wrongdoing, or the ignorance in/from which S does *A* is the foreseeable upshot, ultimately, of fully advertent wrongdoing (*B*, or *C*, or *D*, etc.).

Revisionism Option #1: The Rarity of Culpable Ignorance

(5) Culpable ignorance is *rarer* than many think. [partially from (2) & (3)]

Therefore,

(6) Blameworthiness for acts is rarer than many think. [from (4) & (5)]

Revisionism Option #2: The Inaccessibility of Fully Advertent Wrongdoing

(7) It is extremely difficult to ascertain whether someone has committed fully advertent wrongdoing.

Therefore,

(8) Blameworthiness for action/omissions is extremely difficult to ascertain. [from (4) & (7)]

(9) Many think that blameworthiness for actions/omissions is relatively easy to ascertain.

Therefore,

(10) Blameworthiness for actions/omissions is harder to ascertain than many think. [from (8) & (9)]

Concerning the Regress Argument, Conclusion (4) follows deductively from Premises (1)-(3), at least given a couple of suppressed analytic premises about the meaning of being the "ultimate" and "foreseeable" (rather than "foreseen") upshot of an act as described. Conclusions (6) and (10) are the revisionist conclusions of the revisionist options, and both deductively follow from the conjunction of Conclusion (4) with Premises (5) and (7)-(9), respectively. Taking Option #2 obviously does not require taking Option #1, nor vice versa, but they may conceivably be combined for an even more radical form of revisionism.

4.3 Fully Advertent or Culpably Unwitting Wrongdoing

4.3.1 Explaining Premise (1)

Premise (1) states that an agent is blameworthy for wrongdoing only if either they act contrary to their occurrent belief that it is wrong all-things-considered, or act in/from a culpable ignorance of its wrongfulness. If an act is culpable, the Premise states these ways of acting *exhaust* the possibilities. Wrongdoing is culpable only if it is *either* fully advertent *or* culpably unwitting. The act cannot be culpable if it is non-culpably or blamelessly unwitting.

Fully advertent wrongdoing entails wrongdoing despite "full" or "clear-eyed" awareness of its wrongfulness—that is, despite the true occurrent belief that the conduct is wrong,[58] based on true occurrent beliefs in the act's wrong-making features. Some also require that these true beliefs are *justified* (Ginet 2000, 270) or constitutive of *knowledge* (Rosen 2003, 2004; however later Rosen [2008] requires only true belief). Finally, Rosen (2003, 75-83) and Levy (2011, 138) also explicitly use the term "all-things-considered" to describe the relevant belief in wrongdoing as not *just* the belief that they *morally* ought not to act—where "morally" is construed narrowly (likely to mean "other-regarding")—but the belief that they ought not to act *all-things-considered* as well. After all, there seem to be cases in which one makes a correct moral assessment of wrongdoing but where one believes that one should *all-things-considered* act in that way anyway (e.g., because it would satisfy one's desires, and one judges all-things-considered that satisfying desires here takes priority over moral considerations).[59]

Now, many, including volitionists themselves, have taken acting contrary to one's all-things-considered judgment just to amount to *akrasia* (Rosen 2004; Levy 2009; Peels 2011; E. Harman 2011; Robichaud 2014; Wieland 2017; Rudy-Hiller 2018). But this is arguable.[60] D. Davidson (2001 [1970]), and others following him, have insisted on the possibility of acting contrary to a judgment about what to do all-things-considered yet *in accordance* with one's "all-out" or "unconditional" judgment, *to do* the thing that all the relevant considerations tell against doing. Davidson thinks that this counts as akrasia, but some

[58] Alternatively, it could be the belief that the act violates a moral *obligation*, or *ought* to be avoided. This conforms to the link between wrongdoing, obligations, and "oughts" (at least in the relevant sense) drawn last Chapter.

[59] In the wide sense of "moral" used in this thesis, my preferred way of putting this would be that "there seem to be cases in which one makes a correct *other-regarding* assessment of wrongdoing but where one believes that one should *all-things-considered* act in that way anyway."

[60] I owe this worry about calling it "akrasia" to my supervisor, John Bishop.

disagree on the grounds that akrasia in the circumstances would be acting contrary to one's *all-out* judgment (since an all-out judgment would be one's "final" judgment in the circumstances) (see, e.g., Bratman 1979; Stroud and Svirsky 2019).[61] Instead of picking out the relevant state as "akrasia," then, I pick it out with the term "advertent wrongdoing" (assuming, of course, that one's moral judgments correspond to the moral reasons not to so act).[62] Examples of advertent wrongdoing thus include cases of *deliberately* acting contrary to what one knows one ought, morally-all-things-considered, to do, as well as (and far more probable) the familiar cases of akrasia exemplified by morally reluctant addicts (e.g., to gambling or pornography), unfaithful marital partners, lazy workers, and so on.

But even more is required for blameworthiness according to the volitionist: the advertent misconduct must also be "fully" advertent, by which I mean that the relevant beliefs (in wrongdoing) must be *occurrent*—that is, conscious in the agent's thoughts or feelings—at the time of acting (Zimmerman 1997b, 421-2; Rosen 2004, 309; Levy 2009, 726, n. 16; Ginet 2000, 270). Thus, I employ "fully advertent misconduct" instead of what it is sometimes called in the literature: "clear-eyed akrasia" (FitzPatrick 2008; E. Harman 2011; Robichaud 2014; Peels 2011). That "I should not cheat on my wife," must be phenomenally conscious to the unfaithful husband at the time of the act (e.g., in thoughts or feelings that he should not do it). This stands in contrast to "bleary-eyed" wrongdoing (Mason 2015, 4), where the agent acts contrary only to a *non-occurrent* or *dispositional* belief in wrongdoing. The entrenched addict knows "deep-down" that it is wrong to keep giving in, but she has become so numb to the act that her conscience often fails to alert her to its wrongfulness. For the volitionist, this addict cannot be directly blameworthy for their behaviour, because their deep-down knowledge is not occurrent to them at the time of acting.

If the wrongdoer lacks this "full awareness"—this true occurrent belief in wrongdoing based on true occurrent beliefs in its wrong-making features—then they are what volitionists describe as "ignorant" and they perform wrongdoing "in" this ignorance. As we noted last Chapter, this ignorance could be characterised by false beliefs or simply by the lack of true beliefs. To act *in* ignorance is just to act *while* ignorant (Rosen 2008, 598n; Peels 2014, 479),

[61] S. Stroud and L. Svirsky (2019) observe that "critics think the cases permitted by [Davidson's] analysis simply do not exhaust the range of actual cases of weakness of will", and so note that "those writing after Davidson have tended to focus, then, on the question of the possibility and rational status of action contrary to one's *unconditional* better judgment."

[62] Since Zimmerman (1997a) thinks that freely acting contrary to *any* belief in wrongdoing is culpable, it is possible on Zimmerman's view that the agent is culpable for acting contrary to a mistaken belief about the act's being wrong for reason R, when it is actually wrong for reason S (given some acts can be wrong in two different ways; Rosen 2008, 593-4). I will ignore this complication.

but sometimes the volitionist demands that the wrongdoer acts *from* this ignorance, such that the wrongdoing can be "attributed" to the ignorance (Zimmerman 2017, 81) or the wrongdoer would not have acted were they not ignorant (Rosen 2003, 62; Zimmerman 1997b, 424). Whatever one's preferred view (and we will not discuss this issue until Chapter Five), the result is that the ignorant agent acts "unwittingly" (H. Smith 1983). If the agent acts unwittingly, Premise (1) holds that there is still an important possibility that the agent is blameworthy for doing so (discussed further in §4.4).

Volitionists hold one can be ignorant of wrongdoing for either of two key reasons: factual or moral reasons (e.g., Rosen 2003; Zimmerman 1997b, 422-3). Arthur Coningham was *factually* ignorant, because his ignorance of the wrongfulness of bombing the ships in Lübeck was based exclusively upon ignorance of the prisoners of war and camp survivors on board (facts). The same holds for doctors who prescribe the wrong thing after failing to keep up-to-date with new research (Smith 1983), pilots who initiate take-off without knowing that the gust-lock is engaged,[63] and Zimmerman's (1997b) "Perry," who unwittingly paralyses someone while saving them from the wreckage of a car accident. Although the agents in these cases are ignorant of the act's wrongfulness, this is only because of a "factual error" (Rosen 2004, 304)—that is, ignorance of a non-moral *fact* (i.e., the new research or the gust-lock's being engaged). In contrast, ignorance may be *moral*, where this amounts either to *other-regarding* (narrowly moral) ignorance or to what M. Peterson (2017) has helpfully called "radical evaluative ignorance."[64] In cases of this kind, the agent is aware of all the relevant non-normative facts, and yet they are still morally ignorant. The Battalion 101 shooters who were not "truly conscious" of what they had to do already give us an example of *moral* ignorance. Rosen's "ancient slaveholder," who "buys and sells human beings, forces labour without compensation, and separates families to suit his purposes" (2003, 64) because he does not believe slavery in general is morally wrong, provides another. The relevant cases of radical evaluative ignorance are cases where the agent is aware that they should not act as they do (with regard to morality narrowly considered), but where they have nevertheless judged all-things-considered to act in that way anyway. Rosen appeals to a memorable example of Bonnie, who "rushes over from across the street, elbows you into the gutter" (2003, 77) and jumps into the taxi cab before you, even though she saw that you were there

[63] This was the cause of the Gulfstream IV crash killing seven people in Bedford, Massachusets, 2014 (Jansen 2015).
[64] Rosen (2003, 75) calls the former moral ignorance and the latter normative ignorance, or ignorance "about the reason-giving force of moral considerations."

first, drenched in the rain, and having to put up with whining children. A week later, you learn that she had suffered a virus which "'rewired' the neural circuits underlying her normative sensibility: her view of what matters, and in particular her view of what counts as a reason for action" (p. 78). As a consequence, she "no longer attaches much importance to morality" (p. 78). Bonnie knew it was wrong to act as she did but her virus-produced radical evaluative ignorance of the fundamental normative (I think ultimately *moral*) principle that one ought all-things-considered to do what is morally required left her without any reason to infer that she should not all-things-considered have elbowed you out of her way.

Crucially, of course, the volitionist holds that the agent's wrongdoing may be blameworthy even though it is unwitting. They are blameworthy for the unwitting act only if the ignorance in or from which they act is itself culpable, where this ignorance may be either factual or moral (Rosen 2003, 64; Zimmerman 1997b, 422-3). Thus, only blameless ignorance fully excuses wrongdoing in or from it. Volitionists are open to holding that culpable ignorance provides at best a *partial* excuse for wrongdoing, in the sense that the culpably unwitting act is *less* culpable than a fully advertent act but is still culpable to some degree (Ginet 2000, 271). Otherwise they may hold that culpable ignorance offers no excuse at all (Zimmerman 2008, 175), in which case the unwitting act is just as culpable as the fully advertent act, though (evidently) not exactly in the same way. What volitionists deny is what H. Smith calls the "Liberal View" (1983, 555; cf. Ross 1939, 163–64) that the unwitting act is not culpable at all if it is done in or from culpable ignorance.

4.3.2 Justifying Premise (1)

What is the justification for Premise (1)? Premise (1) (shortened) is the claim that an agent is blameworthy for all-things-considered wrongdoing only if the wrongdoing is fully advertent or culpably unwitting. N. Levy, in my mind, produces the most promising defense of volitionism and Premise (1) in particular. Nevertheless, I will reserve discussion of his defence until Chapter Six, since it is best described in response to FitzPatrick's objection to volitionism. Still, Zimmerman's and especially Rosen's arguments for Premise (1) lay the foundation for Levy's defense (after all FitzPatrick directly challenges Rosen), and so in this Section, I will discuss their arguments. Along the way I will appeal to related work for additional support (especially due to H. Smith 1983) and indicate those arguments that I take to have the most plausible premises.

It is not controversial that fully advertent conduct can be culpable; no one denies that. But the view that *culpably unwitting* wrongdoing can be culpable is more controversial, for defenders of the Liberal View hold that culpable ignorance fully excuses unwitting wrongdoing. What is culpable is the advertent action that led to the state of ignorance and the ignorance itself. Still, whether or not you are a volitionist, you will probably be persuaded of the following:

> Suppose I walk down a crowded sidewalk with my nose in a book. When I knock you over it does me no good to say, 'But I didn't know you were there!' This may be true, and in another context it might signal an excuse. (Suppose I was looking out for obstacles, but you were cleverly camouflaged.) In this case, however, while [I] do act from ignorance, in the sense that I would have acted differently had I known better, my ignorance is obviously no excuse whatsoever. (Rosen 2003, 62)

Now the "Liberal" would probably reply that you do not technically blame me for *walking into you* exactly, but for the fact that I must have decided recklessly to walk down a crowded sidewalk with my nose in a book, and also for my ignorance that you were not there (H. M. Smith 1983, 566). But is that true? I confess to having Rosen's intuitions on this one: surely it would be natural for you to blame me for walking into you (as well as for the ignorance and the reckless choice). This is analogous to the intuition that we can be culpable for the bad *consequences* of our actions (as H. Smith [1983, 567-8] is surely right to acknowledge), even though we do not intentionally cause the event at the time of its occurrence.

Why does culpability for the unwitting act have to depend on *culpability* for the ignorance, as opposed to, say, its epistemic unjustifiability? The reason is simple: it is not enough that the ignorance is merely epistemically *unjustified*, for on virtually every theory of epistemic justification (cf. those mentioned in Chapter Three §3.4.2), it is quite possible that one's unjustified ignorance is not something for which one is morally responsible or at fault.[65] The ancient slaveholder's morally ignorant beliefs may be unjustified (due to a lack of evidence or due to being unreliably formed) but it may not be the slaveholder's fault at all, because he has been brought up to think, indeed has only ever "known," that a perfectly legitimate social

[65] An exception would be views like A. Goldman's (1988 view of weak justification) on which justification is analysed in terms of blamelessness. But even then, what Goldman's view entails about doxastic blameworthiness entails that the conception of blameworthiness/blamelessness he is using is rather thin—e.g., the control condition is not required.

practice is to enslave conquered groups, and make them work for you through whatever means you can.

Now implicit in Premise (1) are two controversial claims: that (i) culpable wrongdoing must be *fully* advertent, and (ii) that when the act is not fully advertent, culpability for the act depends on culpability for the ignorance in/from which one acts. Claim (ii) implies that blameless ignorance exculpates (i.e., fully excuses wrongdoing in or from it). What reasons are there for (i) and (ii)?

Concerning (ii), there is, of course, a long tradition of thinking that blameless ignorance exculpates, especially blameless *factual* ignorance. Aristotle (2013, III.1) thought that "[everything] that is done by reason of ignorance is not voluntary" and that "praise and blame attach to voluntary actions."[66] The ignorance that he had in mind, however, was ignorance "of the circumstances of the action and the objects with which it is concerned" (i.e., factual ignorance), not ignorance "of what he ought to do and what he ought to abstain from," or "ignorance of the universal (for that men are blamed)." Blameless *factual* ignorance may exculpate, but does blameless *moral* ignorance exculpate? Aristotle appears not to think so; but the volitionist disagrees. We will return to reasons for this disagreement when discussing Rosen's arguments for Premise (1). While discussing the tradition of thinking that ignorance is an excuse, it is also worth noting that traditional modes of criminal liability (or *mens rea*) are distinguished in large part along the lines of the degree and content of the agent's (typically factual) ignorance. Criminal "intent" or "knowledge" are the strictest forms of criminal liability; "recklessness," which involves *awareness* of "a substantial and unjustifiable *risk*" of harm (Husak 2011, 200), is less strict; and negligence, which involves the ignorance of that risk where one *should have* been aware of it, is the most mild form of criminal liability. Some argue that we should reject negligence as a form of liability altogether (Alexander and Ferzan 2009). Once again, however, it appears to be an open question whether this framework suggests that blameless *moral* ignorance exculpates, and whether all of the volitionist's conditions must be met. Finally, one must also cite the McNaughten Rule, according to which criminal responsibility requires that one understands the nature of one's actions and generally knows right from wrong, that one is not *criminally insane* due to some affliction of the mind (Wolf 1987, 57). But here too, the rule

[66] Thomas Aquinas (who, of course, drew heavily upon Aristotle) was another important figure in the history of philosophy who argued that ignorance is exculpatory. See P. Furlong's (2017) treatment of his view. The tradition of Judaeo-Christian thought on which Aquinas also drew recognises the exculpatory significance of ignorance (of fact). See, e.g., Jesus of Nazareth's saying in John 15:22, and St. Paul of Tarsus in Romans 1:20.

underdetermines volitionism. The rule only excuses those who are blamelessly ignorant due to some mental affliction (cf. Rosen 2004, 71). And many say that the rule is satisfied by those who *generally* know right from wrong, even if for whatever reason they believed that the specific act was morally right in the circumstances (Vinocour 2020; cf. Wolf's [1987] appeal to it in the context of moral *competence*, rather than moral awareness). Still, despite the inconclusiveness of tradition, Premise (1) seems to get some support from these considerations. How do Zimmerman and Rosen argue for Premise (1)?

Zimmerman's Defence. In explicit defence of the claim (ii) that culpability for unwitting wrongdoing traces back to culpability for the ignorance, Zimmerman (1997b, 2008) takes the *lack* of ignorance as to the moral status of one's conduct to be a "root requirement for responsibility" (2008, 177). The agent must be *aware* of the relevant facts (and norms) to be directly responsible. Zimmerman rightly rejects one proposal for justifying deriving culpability for the unwitting act to culpability for the ignorance, namely, the proposal that if one acts from *any* mental state, then any culpability for the act traces back to culpability for that mental state; anger, as he points out, does not work like that. But in place of this failed justification he finds no other, except to say that it is "deeply embedded in our everyday practice of blaming people for their ignorant behaviour" (pp. 177-8).

In earlier (1986) work, however, Zimmerman appeals to the way in which a relevant level of *control* is crucially compromised when the agent is ignorant, at least when *factually* ignorant. Suppose (to use his example) that Sam could have won a stunning prize as a department store's millionth customer if only he had entered the store when he walked past it earlier in the day (p. 205). Alas, Sam had no clue that he could and would have won it. Zimmerman argues that although Sam had "standard control" over whether he would win the prize—that is, there was nothing physically preventing Sam from getting the prize—Sam nevertheless lacked the "enhanced control" necessary to be directly blameworthy for failing to walk in (p. 205). For Sam to have enhanced control, Sam had to "advert to" (i.e., be aware of) the possibility of the prize, and had he adverted to the possibility and recognised its significance but failed to walk in, he would have been directly blameworthy for doing so. His wife's later berating him would not have been out of place. So not just any kind of control is relevant, for Zimmerman; *enhanced* control in particular is relevant to direct culpability. Otherwise, of course, Sam was blameworthy only indirectly, through being culpable for the ignorance in the first place. And we will see below how it would be natural for Zimmerman to extend the notion of enhanced control to ground this culpable ignorance.

Although Zimmerman (1997b, 422-3) applies Premise (1) to cases of moral ignorance, he does not (explicitly) appeal to these kinds of considerations to support it. There is good reason to do so, however. Consider, for instance, E. Harman's (2015) "Gail," a gangster who believes that she has a moral obligation to kill someone from a rival gang in revenge for a killing of hers. Gail is morally ignorant; her ignorance is rooted in her moral belief that it would be disloyal to her gang not to avenge the death by killing someone from the rival gang. Applying Zimmerman's conception of control, perhaps Gail has *standard* control over whether she carries out the revenge killing, without having *enhanced* control over it. After all, she believes that she is *obliged* to kill the rival gangster. Loyalty requires it, for Gail; she believes that failing to do so would, all things considered, be morally wrong. Thus, she cannot be directly blameworthy for acting from her moral ignorance; she can only be indirectly blameworthy for it if her ignorance itself is culpable.

I do not think that we should be surprised by an appeal to control. On the one hand, control is an intuitive condition on moral responsibility (as we have seen). On the other hand, an analysis of responsibility-relevant control plausibly includes reference to the agent's beliefs or credences. We have already touched on this idea (last Chapter §3.4.2) with the distinction between doxastic *control* and doxastic *influence*, where only the former entails *foresight* (awareness of a doxastic consequence). Consider it also in the words of Levy:

> we need to know what we do in order to be able to control it. But in order to have control over what we do, we need to know how our concepts apply to the case at hand. (2005, 14)

Thus, responsibility-relevant control appears to require the satisfaction of epistemic conditions.[67]

The appeal to control does not go all the way, however, in justifying claim (ii) that culpability for the unwitting act depends on culpability for the ignorance. In his latest (2017) work, Zimmerman—rightly I think—argues that to insist that "enhanced control" is necessary for responsibility is to *beg the question* against those who would argue that control is necessary for responsibility but deny that the relevant control is enhanced control (p. 81). In

[67] See also Rik Peels' (2017, 57-59) account of "intentional" vs. "causal" control. And note how libertarian accounts of freedom or control often have epistemic conditions, since they often envisage choosing from conflicting reasons to act (cf. Kane 2007, 26).

other words, the question remains why *enhanced* control is necessary for responsibility. Thankfully, Zimmerman (2017) *does* say something more in defense of (ii).[68] Zimmerman argues that (ii) is "almost universally accepted" (p. 85) and rebuts one objection raised against it in the literature due to Randolph Clarke (2014). Although the full justification for Clarke's objection will have to be saved until later (Chapter Six), Clarke argues that wrongdoing can sometimes be culpable when it is done from ignorance that merely falls short of some cognitive standard; thus, the ignorance need not be culpable. But Zimmerman replies that this "non-moral" faultiness could never explain the "moral" fault involved in moral blameworthiness. In particular, Zimmerman argues that these cases of cognitively substandard ignorance crucially do not involve a poor moral quality of will, which he thinks is necessary for blameworthiness. This appeal to quality of will puts him on shaky grounds, however, given that quality of will theorists have typically rejected Premise (1) (as we shall see in Rosen's defense below, and in Chapter Six). Indeed, it is partially *because* of this appeal that Zimmerman narrows his (2017) version of the Regress Argument to a claim about a "narrower" form of culpability, not culpability in general. This narrower kind of culpability becomes central to his argument for (ii), however. Zimmerman argues that:

> Given the obvious, and dramatic, difference in quality of will that exists between cases in which one unwillingly does wrong and cases in which one willingly does so, it seems to me plain that, whatever kind of culpability may be associated with the former kinds of case, there is indeed a *new* kind that attaches to the latter. (p. 91)

However, because this new kind of culpability cannot arise in a case of *unwitting* wrongdoing (for one cannot willingly do wrong while ignorant of doing wrong), if such an act is going to be culpable in the same way, culpability for it is going to have to arise at some earlier point "by way of attaching to the ignorance of which the wrongdoing is a consequence" (p. 91). In other words, the ignorance itself must be cashed out as in some sense "willing" (which as we shall see, he cashes out in terms of being the upshot of a willing wrongdoing for which the agent is directly culpable). At any rate, that is his main 2017 argument for (ii) (or a version of it; see n. 68).

[68] Or at least a variant of (ii) which applies only to cases of acting *from* ignorance, not acting *in* ignorance (which he now thinks can be directly culpable). Consequently, Zimmerman would reject Premise (1) on the grounds that it includes a false dichotomy: willing wrongdoing *in* ignorance can be culpable and need not be wrongdoing in *culpable* ignorance. Still, since the considerations that he adduces support (ii), his 2017 argument is relevant here.

It might still be pressed what this new form of culpability could be which attaches particularly to willing wrongdoing, but Zimmerman reassures the reader with an answer: this new form of culpability is the kind of culpability "that is correlated with the appropriateness of the particular reaction of *punishment*" (my emphasis; 2017, 92); indeed it is the kind that he "always had in mind" in previous (1997b, 2008) iterations of the Regress Argument. Thankfully, we find further support for (ii) in this appeal to culpability as liability to punishment:

> Perhaps it is… perfectly appropriate—not at all unfair—to punish someone who has freely and willingly engaged in wrongdoing, but it is *certainly inappropriate*—grossly unfair—to punish someone who acts either unfreely or in keeping with his conscience—unless, of course, he is to blame either for his lack of freedom or for his erroneous conscience. (my emphasis, 2017, 92)

Norms of fairness governing punishment demand tracing culpability back to the ignorance when the act is unwitting or "in keeping with conscience." The question becomes whether it would be fair to punish the agent for her ignorance in the first place.

That is as far as Zimmerman goes in defense of (ii); so what about (i), that directly culpable conduct requires that the agent's judgment of wrongdoing is *occurrent*? Since in his 2017 paper, Zimmerman retreats from the appeal to *occurrent* belief in wrongdoing, we must consider his argument for (i) in earlier work. In the continuing spirit of control or of what one "can" do in cases of fully advertent wrongdoing, Zimmerman argues in the following way:

> [If] a belief is not occurrent, then one cannot act either with the intention to heed the belief or with the intention not to heed it; if one has no such intention, then one cannot act either deliberately on or deliberately despite the belief; if this is so, then the belief plays no role in the reason for which one performs one's action; and one incurs culpability for one's action only if one's belief concerning wrongdoing plays a role in the reason for which one performs the action. (1997b, 421–22)

Imagine that last week you came to believe that referencing a minority's ethnicity in a context to which their ethnicity is irrelevant is wrong, but that *this* week, although you still believed this, you forgot about your moral revelation, and you referenced a minority's ethnicity in a conversation where their ethnicity was irrelevant to the topic. Can you be

directly blameworthy for doing so? Since you could not act with the intention of heeding this forgotten insight at the time, you cannot be directly blameworthy, on Zimmerman's (1997b) view. It was not occurrent in your thoughts, and so there was nothing to intentionally heed when choosing to act. Thus, at best, you can be to blame only indirectly, provided that you are to blame for this belief's failure to be occurrent at the time. We will return to this argument in detail later.

Rosen's Defence. We can summarise Zimmerman's case for Premise (1) as hinging on considerations of control, on awareness being a "root requirement" on responsibility and being a requirement on a certain particularly severe kind of culpability *qua* basis for punishment. Let us turn now to Rosen's case for Premise (1). Since Rosen takes the emotion-based view of blame that "you blame X for doing A when you resent him or feel indignant towards him for having done it" (2004, 296–97), the relevant question for Rosen is when it is appropriate to display emotions like resentment or indignation toward an agent who does not act in full awareness. (This is different from Zimmerman, who holds to a ledger view of responsibility and cognitive view of blame; see Zimmerman 1988.) Now many philosophers have said that the blaming emotions of resentment or indignation are appropriate only when agents display ill will in their conduct. At least on certain versions of what is meant by "quality of will," such a view would seem to account neatly for our intuitions that *factually* ignorant wrongdoers are blameless for their wrongdoing too, unless they are culpable for their ignorance. The newly arrived foreign graduate student who walks across the Oxford lawns entirely oblivious to the fact that only the dons can walk across them, does not seem to deserve resentment or indignation (or even guilt), unless he could reasonably have known about lawn rules. The quality of will view explains his blamelessness by pointing out that he does not display ill will in so doing (e.g., an intention to violate the tradition). The student's factual ignorance blocks inferences to ill will. If his ignorance itself displayed ill will, however (e.g., indifference towards the rules), then he could have been to blame.

But in defence of claim (ii), Rosen objects to this account of why blameless factual ignorance exculpates. His reason is that it yields a mistaken verdict about cases of *moral* ignorance, which often still involve displays of ill will (and so blameworthiness, for quality of will views). Rosen is duly sensitive to the fact that excusing agents for morally unwitting wrongdoing is more controversial than excusing them for factually unwitting wrongdoing, but he thinks that if we take blameless *factual* ignorance to exculpate, then we should think take blameless *moral* ignorance to exculpate. There is a strong enough analogy between them, he thinks, to suggest what he calls the "parity thesis":

> whenever an agent acts from ignorance, whether factual or moral, he is culpable for the act only if he is culpable for the ignorance from which he acts. (2003, 64)

In response to the quality of will theorist, Rosen invites us to consider the case of the ancient slaveholder who mistreats slaves but does not know any better due to being raised in a culture where the "institution of chattel slavery was *simply taken for granted*." Let us:

> [bear] it fully in mind *both* that [the slaveholder] fails to show adequate regard for his slave, *and* that he is altogether blameless in failing to know what sort of consideration the slave is due. When we focus on the act and the objectionable attitude that underlies it, we are no doubt powerfully inclined to blame—so long as we ignore the stipulated fact that he is blameless for not knowing that his slave deserves much more. But when we bear this further fact in mind—when we 'zoom out', as it were—then (I claim) our sense of his culpability evaporates… (2003, 73).

When the ignorance from which the slaveholder acts is blameless, Rosen's claim is that it is unfair to resent him for mistreating the slave, even if the slaveholder manifests an objectionable disregard for the slave in his mistreatment. If part of what is involved is an attempt to "see things from his perspective," I have a similar intuition. Once we do that, we recognise that *from the slaveholder's perspective* he does nothing wrong, and so we lose our sense of his blameworthiness. In other words, attempting to *empathise* with the slaveholder causes us to doubt the appropriateness of resentment, and plausibly resentment without the attempt to empathise is unjustified. But the job is not done at this point, for now we may think him likely to be culpable for failing to have the right perspective in the first place. However, when we observe that it would have taken "a miracle of moral vision" (2003, 66) for him to have formed that perspective (given his background), it seems that we are powerfully disinclined to feel any resentment for the slaveholder, however inclined we may be to resent the institution of slaveholding.

Attempting to harmonise his work, the lesson that Rosen draws from these considerations is something like the view that it is unfair to resent or be indignant toward (and so blame) someone for wrongdoing if for every feature F that makes the act wrong all-things-considered, they blamelessly lack the belief that the act possesses F as well as the belief that

F makes it wrong.[69] Although the slaveholder may believe that mistreating the slave *harms* them, the slaveholder cannot be fairly blamed (resented) for his mistreatment if he blamelessly lacks the belief *that this makes it wrong*. In the case of the newly arrived graduate student, the student is not the fair target of indignation for walking across the Oxford lawns if the student does not recognise the one and only feature that makes the walking across the lawn wrong—that it would breach the lawn etiquette.

Rosen appears to endorse two further principles in support of his conclusion:

It is *unreasonable to expect* someone to avoid a wrong act if, for every feature F which makes the act wrong all-things-considered, they blamelessly lack the true (occurrent) belief that *A* possesses F and that F makes *A* wrong all-things-considered.

"[It] is unreasonable to subject people to sanctions [resentment, etc.] when it would be unreasonable to expect them to have acted differently." (2003, 74-5)

Add this, finally, to an implicit (and plausible) premise that unreasonableness here determines unfairness. Thus, we have a principled way of determining the appropriateness of the blaming emotions (of resentment, indignation, and guilt). This appeal to the constraint of reasonable expectations strengthens Rosen's case for Premise (1) beyond his basic intuition about when resentment would be "properly blocked."[70] We say that the slaveholder cannot be a fair object of resentment, indignation, or anger, because it would have been unreasonable to expect him to have acted differently, given his blameless lack of belief that the features which make it wrong are wrong-making, and because fair resentment, or indignation require that it would have been reasonable to expect him to act differently. The latter principle may in turn be founded on the observation that resentment and the like have a certain force or sting, together with the premise that harmful interpersonal reactions of this sort should be directed only towards those who could have avoided wrongdoing (and thereby avoided incurring these reactions) (cf. Wallace 1996, ch. 7; Levy 2009). (Note that this sort of consideration may also

[69] This principle is the combination of a 2008 principle and a principle from 2003. The 2008 principle is: "X does A from ignorance iff at the time of action, X is unaware that A possesses each of the wrong-making features that it in fact possesses" (2008, 598). For Rosen possessing a wrong-making feature also means recognising (de dicto) *that* it is wrong-making. The 2003 principle is that: "It is unfair to blame someone for doing something if he blamelessly believes that there is no compelling moral reason not to do it" (2003, 74).
[70] Or rather, he sees this appeal as a way of explaining what "underlies" this intuition: 2003, 75.

underlie Zimmerman's [2017] appeal to the *unfairness* of punishment—another harmful reaction—for unwilling wrongdoing.)

The "all-things-considered" in brackets is important for Rosen. Recall the case of Bonnie, who violently elbows you into the curb and steals the cab, even though she believes her doing so to be morally wrong. She does not act fully advertently, however, because after suffering a virus which upset her normative sensibilities, she wrongly takes self-interested considerations to trump moral (or other-regarding) considerations. Given that her radical evaluative ignorance was blameless due to a virus, Bonnie is therefore blameless for taking the fulfilment of her desires to be more important than your rights or well-being. Rosen therefore extends his conclusion to apply to cases of radical evaluative ignorance too. Supposing, however, that Bonnie *was* to blame for her normative ignorance (she intentionally injected herself with the virus, knowing how it would affect her normative sensibilities), then Bonnie would not be let off the hook.

She is clearly not ignorant of every feature of the act that makes it *morally wrong*, though, and so someone might contest Rosen on the extension of his thesis to radical evaluative ignorance. Rosen imagines his objector insisting on the following, by analogy to cases of ignorance of law:

> Bonnie knowingly violates the moral law. She is a competent adult who knows what that law requires, and she knows that competent, witting transgressions are typically met with resentment. So she should not think that it's unfair for others to blame her. Instead she should think, "Blame is what you get when you break the moral rules. I knew that in advance, so blame is perfectly appropriate. In this case, however, it's a price I'm willing to pay." (2003, 81)

Is she not therefore directly blameworthy for her transgression, rendering irrelevant her radical evaluative ignorance? Rosen thinks that even in these circumstances, she is not to blame:

> When Bonnie acted she was deluded through no fault of her own. If you like, she was half blind; her moral vision was distorted. Despite every reasonable reflective effort, she failed to see a sufficient reason to treat you better. And given all that, it's unfair for you to blame her. To do so would be to expect more of a person than it is reasonable to expect. (p. 82)

This is Rosen's final (albeit "inconclusive") word on the matter and so he concludes with Bonnie's blamelessness. I too find this plausible. Recall my claim, above, that in order for resentment to be appropriate, the blamer must have attempted to *empathise* with the wrongdoer, to try to see things from their perspective, or feel what the wrongdoer would have felt at the time. In the case of Bonnie, I think that an exercise of empathy would help us to see that Bonnie could not have felt the weight of the reasons why elbowing us into the gutter and stealing the cab were wrong. Bonnie's moral judgment would not have stood in the special relation that evaluative judgments usually stand with our practical reasoning. Bonnie's judgment that it would be wrong to steal the cab would have been no more linked to her practical reasoning as M. Bratman's (1979, 158) "it would be chic" would be linked to practical reasoning for someone who does not care about the chic-ness of their conduct. Thus, it seems unfair to blame Bonnie for doing what she did.

Concerning claim (i) that *full* advertence is required for direct blameworthiness, Rosen does not say much at all. Levy will pick up on this absence with an argument in favour of this requirement, from reasonable expectations, which I will discuss in Chapter Seven.

That concludes my presentation of Rosen's defence. Despite no argument for (i), I take it that Rosen goes further than Zimmerman in defence of Premise (1). Zimmerman is right, I think, to appeal to considerations of control or what the agent *can* plausibly do, but he is also right to recognise that this does not go all the way toward supporting Premise (1) (claim II)—and unfortunately his remedy for this, in 2017, is really an argument for a slightly different premise (allowing direct blameworthiness for any "willing" wrongdoing, which encompasses acting *in* ignorance) and thus continues to offer inconclusive support for claim (ii). Moreover, Zimmerman's 2017 appeal to a pluralistic quality of will theory of blameworthiness fails to account for Rosen's plausible point that we lose *all* sense for the morally ignorant wrongdoer's blameworthiness, even while recognising their ill will, when we realise that it would have taken a "moral genius" to do the right thing. We do not *just* lose a sense for their liability to punishment while retaining a sense of their culpability in a weaker sense, as Zimmerman would have it. (I should add, too, that for all that I have discussed in this thesis so far, there has been no need to *pluralise* blameworthiness as Zimmerman does here.) Finally, I find that a strength of Rosen's argument is his appeal to what it would have been *reasonable to expect* of the ignorant wrongdoer at the time, for reasonable expectations seem inextricably linked to responsibility, they seem to be sensitive to capacities to act differently, and one's capacities to act differently appear to be inhibited in cases of ignorance. As we will

see in Chapter Six and Seven, this line of reasoning is further developed and refined by Levy (2009).

4.4 Culpable Ignorance

4.4.1 Explaining Steps (2)-(4)

We have seen that the ignorance must be culpable according to the volitionist if the agent is to be blameworthy for wrongdoing in or from it. But *how* can the ignorance be culpable? Premise (2) holds that ignorance is culpable only if it is the foreseen upshot of some earlier benighting act for which they are blameworthy. And Premise (3) holds that the benighting act is culpable only if the benighting act is either fully advertent or culpably unwitting. Conclusion (4) then takes Premise (1)-(3) together (along with a couple of suppressed premises concerning the meaning of an act's being the "ultimate" and "foreseeable" upshot of a benighting act) and concludes that the agent is blameworthy for an act only if the act is either fully advertent, or the ultimate and foreseeable upshot of fully advertent (benighting) wrongdoing. I will explain the Premises first, before discussing the arguments for them.

Two key features of Premise (2) merit explanation: the earlier benighting act, and the fact that the ignorance is the "foreseen upshot" of this act.

Volitionists follow H. Smith (1983) in using the term "benighting act" to refer to the earlier act (or omission) which, in some important sense, *causes* one's ignorance. In her words, a benighting act is "an initial act, in which the agent fails to improve (or positively impairs) his cognitive position" (1983, 547). The pilot deliberately omits to run through the full pre-flight checklist before initiating take-off, thereby causing ignorance about the status of the airplane. An ancient slaveholder deliberately chooses not to listen to critics of the institution of chattel slavery and therefore sustains his moral ignorance. Benighting conduct must itself be wrong, for Smith, because it violates obligations to check, be attentive, ensure that one will remember, or take precautions. This is also Rosen's view, who holds that a benighting act must violate a "procedural obligation" to take a "precaution against a certain bad outcome—ignorance" (2008, 603)—whether that obligation is epistemic or more narrowly moral (2008, 601-3).[71]

[71] See his discussion of this distinction in relation to the case of "Goldberg" and "Himmelfarb." This is not of course, Zimmerman's view, who denies that conduct must be wrong to be culpable (see §3.3.5).

Following Smith, the volitionists also affirm that the benighting act must itself *culpable* to be the grounds of culpable ignorance (just as they think that the *ignorance* must be culpable for any acts issuing in/from it to be culpable). And here the volitionists apply their epistemic condition on culpable conduct all over again (Rosen 2004, 303): if the benighting act is itself culpably unwitting,[72] then we must trace culpability for it back to culpability for *another* benighting act, where we might find that *this* benighting act is *itself* culpably unwitting and so we must trace once again, until we find a fully advertent benighting act in which culpability "bottoms out" (Levy 2011, 116). Of course, if there is no original locus of blameworthiness, then the agent is off the hook for their eventual unwitting wrongdoing. This is why their argument is called the "*Regress* Argument." The possibility of culpability passing through multiple "chains" (Zimmerman 2008, 176) or "sequences of blameworthy benighting acts" (Ginet 2000, 274) is what explains the concept of tracing to an "ultimate" source invoked in Conclusion (4).

Following orthodoxy on responsibility for consequences (see §2.3.5), volitionists also add a *foresight* condition on tracing. More precisely, for the eventual unwitting act to be culpable, there must be (occurrent; Zimmerman 1986, 206; Ginet 2000) foresight of the ignorance (and the act from it)[73] as a probable or possible consequence of one's benighting behaviour, or else there must be the culpable *lack* of that foresight. The benighting act is fully advertent only if the agent *foresees* the ignorance resulting from it, and so a lack of foresight would also make the benighting act *unwitting*, and so would demand further tracing (Ginet 2000, 273). Now one will notice that Conclusion (4) uses the word "foreseeable" rather than "foreseen" (as Premise (2) does). The reason is that in cases where there are multiple links in the culpability chain from the original benighting act—say, *C*—to the ignorance from which one commits the eventual unwitting wrongdoing *A*, the volitionist need not believe that the foresight at the time of *C* must be of *A* or the ignorance from which one commits *A*—the ignorance at the *end* of the culpability chain. Rather, the volitionist, following Ginet (2000, 276), need only say that the foresight at the time of *C* must be of the *next* benighting act in the culpability chain—namely, *B* or of the ignorance from which one does *B*—and that at the time of *B*, there must

[72] Levy (2017, 252) *defines* benighting acts as fully advertent, but Smith (1983) does not appear to do so. Terminologically, I follow Ginet (2000, 274-5) in allowing benighting acts to be unwitting.
[73] This foresight of the *ignorance* seems to be sufficient on Rosen's view for the *act from* ignorance to be culpable too (2008, 603–4). But if so, there is some disagreement here. Zimmerman (1997b, 421) and Ginet (2000, 276) argue that culpability for the act from ignorance would require foresight of the act itself. Since Premise (2) is a necessary condition for culpably unwitting wrongdoing according to Rosen, Zimmerman and Ginet, I have chosen not to make much of this detail.

be foresight of *A* or of the ignorance from which one commits *A*, for *A* to be indirectly culpable. That is why I have chosen "foreseeable" rather than "foreseen" for Conclusion (4).

In contrast with Zimmerman and Ginet, Rosen says relatively less about this foresight component but insightfully links foreseeability and the notion of obligation inherent in taking *precautions against ignorance* (2008, 604).

4.4.2 Justifying Premises (2) and (3)

What reasons are there for Premises (2) and (3)? Beginning with Premise (2) (that the ignorance is culpable only if their ignorance is the foreseen upshot of a benighting act for which the agent is blameworthy), we should observe that there are two controversial propositions contained in this Premise. The first is that ignorance is only culpable indirectly through culpability for an earlier benighting act. The second is that the ignorance must be *foreseen* for the agent when performing the benighting act.

The basic argument for the first claim is the thesis of *doxastic involuntarism* (cf. §3.4.2) combined with the premise that, as Rosen puts it, "I am responsible for [any passive] occurrence only if it is the (foreseeable) upshot of prior culpable activity on my part " (2004, 302). Focusing on the thesis of doxastic involuntarism, culpability for the ignorance is *indirect* because we do not have *direct* voluntary control over our ignorance or other belief states. Volitionists do not defend the thesis of doxastic involuntarism at length, for it is well-accepted in ethics and epistemology and by most critics of volitionism.[74] It is also assumed in this thesis. In favour of the claim that culpability for any passive occurrence can only be explained indirectly, Rosen contrasts his view with a character-based quality of will account of the conditions on responsibility on which responsibility attributions are grounded in attributions of good or bad character (see, e.g., Owens 2000, 123). On this sort of view, passive occurrences (e.g., feelings, thoughts, involuntary reactions) are occurrences for which the agent is responsible when they manifest the agent's character; no tracing to voluntary activity is required. Although this view makes responsibility attributions simpler, Rosen takes there to be a critical rejoinder to it—that it is "obtuse" to blame the following agent:

[74] Volitionists may be able to accommodate doxastic voluntarism (if it turned out to be true) provided they were able to argue that direct belief-formations (-retentions) or suspensions could themselves be fully advertent misdeeds (e.g., believing that one ought not to form that belief, but forming it anyway). Note that the volitionist, C. Ginet (2001) has in another context defended the possibility of deciding to believe.

> Suppose Fred finds himself momentarily wracked with schadenfreude when he discovers that his rival's book has been remaindered. Since he knows that it's wrong to take pleasure in someone else's misfortune, he instantly resists and (let us say) stifles the emotion… [Now] stipulate that hitherto Fred has been neither reckless nor negligent in the management of his character; he has done whatever it is that a person is supposed to do (which may be nothing at all) in order to prevent this minor vice from taking up residence in his personality. (2004, 302)

The character quality of will view (and likely other quality of will views; see Chapter Six) would hold that Fred is blameworthy for the manifestation of ignoble character. But this seems false, since Fred could not have done anything about it. The reaction is not really Fred's *fault*. Thus, it appears that we do need to trace culpability for the ignorance involved in an unwitting act back to an earlier benighting act or omission.

The second proposition contained in Premise (2) is that the ignorance must be foreseen. As we have already noted, the requirement of foresight gets widespread endorsement in the responsibility literature and so the volitionist has no special burden of justification (see §2.3.5). Even still, in favour of the foresight condition, Zimmerman argues from a more general premise that culpability for *any* consequence of an action is determined by the extent to which the consequence is "cognitively connected" to the act (1997b, 418–19). If Daisy did not have the foresight that a mad man would set her house on fire if she did not complete her puzzle today, then she cannot be held to account for it actually happening. Rosen (2008, 603-4) appeals to the fact that standard legal practice requires foreseeability, and also to a compelling thought experiment, highlighting a strong intuition in favour of the possibility of blameless ignorance as a consequence of culpable benighting wrongdoing.

We turn now to the *prima facie* justification of Premise (3), that a benighting act is culpable only if it is either an instance of fully advertent or culpably unwitting misconduct. The issue here is whether the same considerations in favour of this disjunction being true of normal non-benighting acts are true of benighting acts as well. I shall not rehearse those details here. Volitionists at this point tend to argue that the same considerations that apply to ordinary acts apply to benighting acts as well, on the grounds that benighting acts are acts after all, and that the principles already on the table apply to acts in general (Rosen 2004, 303). One might, however, doubt whether benighting acts ought to be treated the same way as non-benighting acts when assigning culpability. Focusing on the case of Perry, Zimmerman admits that an "obvious" answer to the question of what grounds culpability for his ignorance

(of the fact that rescuing Doris could paralyse her) is Perry's "carelessness," "inconsiderateness," or "inattentiveness" in deliberation.[75]

> Surely, we might say, any careful or considerate person would at least have entertained the possibility of doing more harm than good by means of a precipitate rescue. (1997b, 417)

So why do we not we ground culpability for ignorance in simply a manifestation of carelessness or inconsiderateness? Zimmerman (1997b, 417) argues that Perry must have been aware of the wrongfulness of his carelessness or inconsiderateness in the first place to be directly responsible for his careless conduct. Otherwise he would have had to be indirectly responsible for this conduct, by way of being ignorant of the wrongfulness of it in the circumstances. "Thus the argument would apply all over again" (p. 416). Benighting and non-benighting acts should be treated the same as far as culpability is concerned.

4.5 The Revisionist Options

From Premises (1), (2) and (3), we get the Conclusion of the Regress Argument: (4), otherwise put: blameworthiness for an act or omission always is, or bottoms out in, blameworthiness for fully advertent wrongdoing. But notice now that this makes culpable wrongdoing only ever as easy to come by as fully advertent or culpably unwitting wrongdoing. It turns out, however, that this makes blameworthiness *harder* to come by than many think, because *revisionist* volitionists (excluding Carl Ginet) take it that either fully advertent misconduct or culpable ignorance is harder to come by than many think.[76] As we saw above there are two options for the revisionist here, concluding either that (6), blameworthiness for acts is rarer than many think, or that (10), blameworthiness for acts is much harder to ascertain than many think. I turn first to rarity revisionism.

[75] Rosen's go-to answer for explaining culpable factual ignorance is roughly the same: "epistemic irresponsibility—e.g., negligence or recklessness in the management of one's opinion" (2003, 65).
[76] "Hard to come by" is sufficiently ambiguous between the claims of Options #1 and #2 (below), and so I use it intentionally to pick out the general revisionism of both.

4.5.1 Option #1: The Rarity of Culpable Ignorance

The first revisionist option is the one that Zimmerman (1997b, 425-6; 2008, 175) and Levy (2011, 110-32) take to conclude that blameworthiness is *rarer* than many think. The key premise in this option is Premise (5), that culpable ignorance is rarer than many think. The explanation for this Premise should be obvious: culpable ignorance is uncommon according to volitionism, but many people think that it is quite common. This supports Conclusion (6), given that if many take culpable ignorance to be common while volitionists take it to be rare, then the total amount blameworthiness will be rarer on volitionism than for "many" people. This holds true, even if fully advertent *non-benighting* wrongdoing is easy to come by (which the volitionist is inclined anyway to deny).

But why is culpable ignorance rarer than many think? The justification for Premise (5) depends partly on what people think, and partly on the significance that the volitionist attaches to Premises (2) and (3). Let me focus first on what people typically think.

Note, to begin with, the aforementioned point that it is natural to account for culpable ignorance in terms of the wrongdoer's "carelessness" or "inattentiveness" in the circumstances, but carelessness or inattentiveness are surely quite *common* intellectual vices. There is also the commonsense link between judgments of blameworthiness for ignorance and judgments of what someone "*should have* known." As Zimmerman puts it:

> our common practice indicates that we think that such culpability is frequently incurred; for we often blame people for performing actions that were wrong (or that we take to have been wrong) on the grounds that, even if they *didn't* know that what they were doing was wrong, they *should have* known this. (2008, 175)

Although we often have this reaction to cases of factual ignorance, Levy is at pains to show how we have this reaction *a fortiori* to cases of *moral* ignorance. It seems, after all, that it is "*easy* to know moral facts" (2011, 118), given our innate dispositions (p. 121). Furthermore, morality's "distinctive importance is almost universally appreciated" (p. 122), and so Levy reasons that:

> in the absence of compelling grounds to find that someone is non-culpably ignorant with regard to morality, [many think that] we should assume that they either know full

> well what they are doing, or are wilfully ignorant of morality's demands. (p. 119)

As we shall see, many philosophers are themselves committed to the commonness of culpable ignorance. On Moody-Adams' (1994, 300-301) view, for example, ignorance is often culpable when it is "affected"—or the product of a *desire* to remain ignorant. The responsibility externalist theories that we will discuss in Chapter Six also entail the relative commonness of culpable ignorance, and collectively, these theories probably amount to the *majority* of theories on the epistemic condition in the recent literature. We have good reason, then, to think that culpable ignorance is believed to be common by many.

Why, *on volitionism*, is culpable ignorance relatively rare? Well, seldom ever is ignorance the foreseen upshot of a benighting act which is also fully advertent; and yet these are required for culpable ignorance according to the volitionist. Begin with the fact that the benighting act must be fully advertent. How often are we in a position to take a precaution against ignorance but decide in full awareness to forgo that precaution (and thereby commit a benighting act)? It is difficult to say with much confidence, but these cases are probably quite rare. There is to start with, already the *a priori* observation that such acts, to be directly culpable, must meet several "demanding" epistemic conditions (Levy 2011, 131). Not only must they be like ordinary non-benighting acts that are performed fully advertently, but they must also be performed while (occurrently) foreseeing the risk of causing or sustaining ignorance (and even of subsequent unwitting conduct, for Zimmerman and Ginet). Obviously the greater the number of conditions that must be satisfied, the lesser the number of directly culpable benighting acts there will be. But on top of these *a priori* remarks, there is also *empirical* evidence in favour of the rarity of fully advertent *benighting* wrongdoing. Levy argues that moral ignorance is ubiquitous, on account of what we might call moral "line drawing" (to borrow Rosen's [2004, 305] terminology). There are significant differences between cultures, for instance, on the appropriate application of basic moral concepts like justice and temperance, leading to different, incompatible, and hence (for some) *ignorant* views about their proper application. Appealing to Marc Hauser's (2006) empirical work, Levy argues that:

> [we] have an innate sense of fairness, but what counts as fair is subject to cultural modulation. Similarly, we have an innate prohibition against killing, but each culture defines a class of circumstances in which the prohibition is lifted, either failing to

> apply to particular groups of people, or failing to apply to particular groups of people in particular circumstances. (2011, 122)

Not only do we "extend the scope of [our] norms rather unevenly," but we regard exceptions to those norms as principled, says Levy (p. 121). "Racists" are sad case in point (2011, n. 4). The significance of these considerations is that seldom do we ever perform benighting acts while in full awareness of the way in which they continue to nourish our moral ignorance. Moreover, this ignorance affects both our judgments of the moral significance of our benighting wrongdoing and our judgments constitutive of *foresight*:[77] even if the ancient slaveholder had a chance to revise his objectionable beliefs about slaves, it is unlikely that he ever *foresaw* his ignorance about how to properly treat slaves. We should add that our deeply held false moral beliefs are often the products of behaviour during adolescence, when we did not know better, and could not have foreseen the consequences of that behaviour in terms of our later *ignorance*. Youth are notorious for their underdevelopment when it comes to appreciating the *consequences* of their actions. And finally, there are those who have been diagnosed with *psychopathy*, a condition which is now well-regarded as involving moral blindness, the inability to empathise, or the inability to tell *moral* transgressions apart from social transgressions (Levy 2011, 119-20; Talbert 2008, 518).

I have been focussing on the case of moral ignorance, partly because it is easier to find cases of fully advertent misconduct leading to factual ignorance. Nevertheless, we can also think of many cases in which one's factual ignorance failed to be the foreseen upshot of a fully advertent benighting misdeed. My (rather daft) failure to be aware of the need to get my car serviced for over three years was not traceable to a moment when I ignored the judgment that I ought all-things-considered to check when I last had a service, because (and this is the daft part) there was no such judgment. The comments about youth also apply to foresight of factual ignorance (e.g., not entertaining the thought that saying something will hurt someone's feelings).

There is good reason, then, to think that we rarely ever perform benighting acts (or omissions) in full advertence and with foresight of resulting ignorance. But as such, our ignorance is rarely ever culpable, and so it less culpable than commonsense would have it.

[77] These arguments also strongly suggest that we are seldom ever blameworthy for our morally unwitting *non-benighting* wrongdoing as well, but Levy (2009, 2011) focuses mainly on the case of benighting wrongdoing.

4.5.2 Option #2: The Inaccessibility of Fully Advertent Wrongdoing

The second revisionist option is the one that Rosen (2004) takes to conclude that (10) culpable conduct is more difficult to *ascertain* than many think. Rosen argues that this is because (7), any given case of fully advertent wrongdoing is almost impossible to ascertain, and because (9), many take blameworthiness to be relatively accessible.

Start with Sub-conclusion (8) in the words of Rosen: "*it would be unreasonable to repose much confidence in any particular positive judgment of responsibility*" (2004, 308, his emphasis). For Rosen, this is due to the inaccessibility of fully advertent misconduct (Premise (7)). And Premise (7) is justified by the following reasoning:

> given the opacity of the mind—of other minds and even of one's own mind—it is almost always unreasonable to place significant confidence in… a judgment [that the act 'is, or derives from, an episode of genuine, full-strength akrasia']. (2004, 308)

The point is that we have, if anything, only a very blurry view of someone's mind and so we should not have confidence in a judgment that there has been fully advertent wrongdoing (cf. Levy 2017, 259). As a matter of fact, fully advertent wrongdoing may be hard to notice in either ourselves or others, for as Rosen notes:

> it is not readily distinguishable from an impostor: ordinary weakness of will. The akratic agent judges that A is the thing to do, and then does something else, retaining his original judgment undiminished. The ordinary moral weakling, by contrast, may initially judge that A is the thing to do, but when the time comes to act, loses confidence in this judgment and ultimately persuades himself (or finds himself persuaded) that the preferred alternative is at least as reasonable. Moreover, in between these two pure cases there is a continuum of cases; cases in which the agent *suspends* his original judgment without quite rejecting it; or cases in which it is simply indeterminate as the agent acts whether he in fact believes that all things considered he should do A. (2004, 309)[78]

Although Rosen regards his argument as the outcome only of "reflection," I think that the line

[78] From what I can gather, writers on akrasia would probably identify "ordinary weakness of will" with akrasia. See, e.g., how the terms "weakness of will" and "akrasia" are used interchangeably in Stroud and Svirsky (2019).

of reasoning is good. Think of the many occasions on which you have done something that you later regretted or felt guilty about. How many times did you decide to act despite the persisting judgement that all-things-considered, you should not do it? Perhaps there were times when the thought entered your mind during deliberation, but it is more likely that you ended your deliberation with a thought to the effect that "she'll be right" (as we say in colloquial New Zealand English). In such a case, you would not have acted contrary to, but *in accordance with*, your (new) all-things-considered better judgment. In other such times, you probably never judged that, all-things-considered, you should not perform the act at all. Although you judged that there was a reason against acting, you probably *also* judged that there were reasons *for* it, and so you did not judge that all-things-considered you should not act.

If, as we have just been arguing, it is difficult to determine whether we performed a fully advertent misdeed through introspection or memory, it will be even harder to determine such a deed on the part of others. This is no doubt due to what some call the "privacy of the mental." We do not have access to others' minds or to what they are thinking or feeling, beyond what we can infer from their words or from the complexity of their behaviour. This is not to say that we can *never* be sure of fully advertent wrongdoing on the part of others (although I may be disagreeing with comments by Rosen here; cf. 2004, 309). After all, there is often good reason to trust the testimony of those who confess to doing wrong from the (often akratic) weakness involved in fully advertent wrongdoing. I would, for instance, trust Paul of Tarsus' testimony: "For I do not do the good that I want to do, but the evil I do not want to do—this I keep on doing" (Romans 7:19). And if we have been journeying with a close friend or client through recovery from an addiction or an unhealthy lifestyle, we may also be in the position to be confident of their fully advertent wrongdoing in their occasional relapse. But in the absence of these close relationships, and of the sincere testimony of others, it appears extremely difficult to be confident that others were fully aware of their wrongdoing. I do not know, for instance, whether my old teacher who, despite having a loving family, turned out to be leading a double-life and would compulsively lie to them, was really acting fully advertently. In fact, I have good grounds to think that he was not acting in this way by the time that he was caught, due to evidence of how he justified his lifestyle even after it all came out. Since he was someone I deeply respected, I would like to have thought that he kicked off his duplicitous lifestyle in full awareness of its significance, but I cannot really know for sure. This reveals that the problem of identifying fully advertent misbehaviour is compounded once we are convinced that the wrongdoer has acted in or from

ignorance, because then we have to determine whether there was a fully advertent misdeed *in the past*, on top of the difficulty of determining whether there was a fully advertent misdeed in the (near) present.

We have good reasons, then, to endorse Premise (7). But given Premise (7) and the conclusion of the Regress Argument, (4), it follows that (8) blameworthiness is difficult to ascertain. The next independent Premise is (9) that many think that blameworthiness is easy to ascertain. There are reasons for this claim which mirror those offered in favour of the common acceptance of the commonness of culpable ignorance. Just as many are inclined to hold people blameworthy for their ignorance as soon as they have judged that the wrongdoer "should have known" and "could have known" the relevant fact or norm, it is natural for people to hold people blameworthy as soon as they judge that the wrongdoer "should have" done something else, and "could have done" so. (This is why I think that the classic focus on whether alternative possibilities are necessary for blameworthiness is not misplaced.) Akin to the point that I made before, we should also note that many theories of the epistemic condition entail that blameworthiness for actions is easy to ascertain, or at least easier to ascertain than volitionists make out (see Chapter Six). A character-based quality of will view, for example, will hold that blameworthiness judgments are warranted as soon as the act reflects the agent's bad character. But plausibly, we have good access to whether or not someone has the bad character that issued in the act; we can tell that a lie issued from that person's dishonesty—that Cruella's choosing her daughter to die issued from maleficence or filial indifference. So it seems that we are well positioned to accept Premise (9) as well. But then from (8) and (9) it follows that blameworthiness is harder to ascertain than many think.

4.6 Conclusion: The Significance of Revisionism

The upshot of either form of revisionism (or indeed some combination of both) is that blameworthiness is harder to come by than many think. This conclusion is significant, for a number of reasons. To begin with, as Rudy-Hiller (2018) has observed, most philosophers feel that volitionism somehow goes "awry," that surely there are more instances of blameworthiness out in the world or that we can access. This gives us *one* reason, I think, if only a minor reason, to resist volitionism about the epistemic. The *reflective equilibrium* methodology adopted in this thesis—according to which a theory is *pro tanto* better for

harmonising with our intuitions than if it does not—seems to suggest as much. And I endorse the general strategy in the epistemic condition literature (and ethics more widely) to be "methodologically conservative" (*pace* Levy 2017), at least to the extent that this strategy entails that *one* advantage of a theory *among many others* is that it harmonises with commonsense intuitions. But even if a theory has no such advantage, we might think that the revisionist conclusion is still significant *precisely because* of the way that it strikes us as somehow "off" and thereby open to scrutiny. Of course, the conclusion is also significant for its practical import. It entails that we should blame less than we do (in conjunction with the obvious claim that we should only blame the blameworthy). Is this in-and-of-itself *bad*? Probably not, for it may well be "music to our ears." We might find delight in the knowledge that we are off the hook as often as the volitionist thinks. Many of those who write and read popular psychology would likely also praise this revisionist upshot, given their tendency to argue that we should rid ourselves of blame and replace it with something like "taking responsibility" or "accountability" (meaning something quite different than in the philosophical literature).[79]

However, one reason for worrying about a prescription to blame less is that it may introduce a certain level of artificiality to our natural blaming practices. If, for example, we think that certain emotions like resentment or indignation towards the wrongdoer are only justified if the wrongdoer is worthy of blame, we might often find ourselves having to suspend natural impulses towards having or displaying those emotions, because we cannot be certain of responsibility for fully advertent or culpably unwitting, or because we cannot really prove that the agent could have avoided causing their own ignorance. We might in turn feel frustrated that our emotions are not getting an outlet. There may, moreover, be good reason for thinking that our natural blaming attitudes and practices underlie important social institutions (e.g., the law and education), and that if this foundation were shaken, we would incur the significant cost of having to revise these institutions accordingly (van Woudenberg 2009). Indeed, such a cost seems to me to be the reason why we should accept *some* form of methodological conservatism.

Whether or not this revisionism is primarily a "good" or a "bad" thing, I think it is clear that it is worthy of philosophical attention. At least I feel myself compelled to give a response.

[79] I have personally encountered this when describing my work on blame. For examples from the web, see Murphy (2015), Timms (2017), Marilyn (2018). One should note, however, that most of the time these writers condemn *blame-shifting*, as opposed to blame as such—i.e., unjustifiably blaming others when one is just as much to blame—or they condemn *blaming* (outwardly) without giving consideration to inward blame.

Chapter 5
A Reasons-Respondence Theory of Responsibility

5.1 Introduction

In the last Chapter, my review of the Regress Argument suggested that it does a decent job of supporting the core view of volitionism, that blameworthiness for any conduct always is, or bottoms out in, blameworthiness for fully advertent wrongdoing. As a matter of fact, it is the goal of this and next Chapter to argue that the Argument is even *better* than on first impression. In this Chapter, I will claim that the Argument's critical presupposition of "culpability internalism" follows from the nature of blameworthiness as being morally at fault, and responsibility as consisting in responses to reasons. Culpability internalism, recall, is the view that moral blameworthiness for some conduct depends on the agent's having beliefs or credences about the moral significance of that conduct. In the next Chapter, I argue that some of its greatest foes—those who fall into the "culpability externalist" camp—fail to dismantle culpability internalism. The upshot of both Chapters is that the need to offer a response to the Regress Argument is felt ever more acutely. In the end, however, there is a plot twist: the present Chapter's case for culpability internalism is, in Chapter Seven and Eight, turned against the Regress Argument and its volitionist proponents. For now, however, let us concern ourselves with the case *for* culpability internalism.

The basic thesis of this Chapter is that volitionism, and thereby the Regress Argument, gets support from the nature of blameworthiness for wrongdoing as being morally at fault for wrongdoing and moral responsibility as consisting responses to normative reasons. This view of responsibility stands in contrast to answerability views, extant attributability views, and accountability views—in relation to which I argue that it comes out on top. Although broadly reasons-responsive approaches to responsibility can be found in the literature, what is unique about the account that I am proposing is that it is an analysis of *the very nature* or *concept* of responsibility and turns on an *actual* relation between agent, reasons, and action. But if this

account of the concept is correct, I argue, culpability internalism follows. This is so, due to the (previously defended) claim that responsibility for wrongdoing entails blameworthiness, a claim that I continue to defend in this Chapter.

The structure of the Chapter is as follows. In §5.2, I problematise answerability views and extant attributability views of responsibility. In §5.3 I put forward my view of blame for wrongdoing as moral faulting for wrongdoing, which emerges out of my discussion of extant attributability views, and which is in turn used to support my account of the nature of responsibility. In §5.4, I then discuss the more respectable accountability view of the concept of responsibility, but which I find wanting. I then conclude the Chapter with §5.5 and §5.6 which set out and defend a reasons-respondence view of moral responsibility.

5.2 Against Attributability & Answerability Views

It will be recalled (from Chapter Two) that there are three broad families of views on the nature or concept of responsibility: attributability views, accountability views, and answerability views. In this Section, I would like to problematise extant attributability views and answerability views, even though (as I will explain below) I think that their agent-centredness is a mark in their favour.

Recall that attributability views hold that responsibility for *x* consists in *x* being attributable to the agent in some deep and particular way. Recall that one type of view of this kind is the "ledger view," according to which someone's responsibility for *x* consists in *x*'s being part of that person's "moral record" or "ledger of life." But notice that this theory is laden with *metaphor*. Apparently, the only way of maintaining a literal description of the view is to hold that there *really is* a moral record of everyone's moral achievements and failures, but this seems to make the concept of responsibility uncomfortably beholden to the existence of God, Karma, or an afterlife (whether or not we believe in these things).

The second family of attributability views are "real-self" (Wolf 1990) or "deep-self" (Wolf 1987) views which hold than an agent is responsible for her act just when she acts from her real-self (e.g., located in her second-order desires, Frankfurt 1971; or in her values, Watson 1975). Apart from the objection that this account seems to threaten a vicious regress—given that one can surely be *responsible for* one's deep self (Wolf 1987)—these views do not seem to identify a property that is *sufficient* for responsibility (i.e., in the property of "self-disclosure"; Watson 1996). To see why, consider Wolf's famous thought experiment:

JoJo the Insane

JoJo is the favorite son of Jo the First, an evil and sadistic dictator of a small, undeveloped country. Because of his father's special feelings for the boy, JoJo is given a special education and is allowed to accompany his father and observe his daily routine. In light of this treatment, it is not surprising that tittle JoJo takes his father as a role model and develops values very much like Dad's. As an adult, he does many of the same sorts of things his father did, including sending people to prison or to death or to torture chambers on the basis of whim. He is not *coerced* to do these things, he acts according to his own desires. Moreover, these are desires he wholly *wants* to have. When he steps back and asks, 'Do I really want to be this sort of person?' his answer is resoundingly 'Yes,' for this way of life expresses a crazy sort of power that forms part of his deepest ideal. (1987, 54)

Jojo, it seems, acts from his real self. Yet is Jojo morally responsible for whimsically inflicting these horrors on subjects? His deeds are horrific, sure, but like Wolf, I struggle to see how if he really is that "insane" or morally incompetent, he can be blameworthy for his horrific deeds. He is causally responsible, but he does not seem *morally* responsible for his actions, since he cannot in principle appreciate what is morally what. It is not in his power to do the right thing, at least in the way that would be fair to expect of him. Consider the similar judgments we might make of serial murderers like Robert A. Harris (Fischer and Ravizza 1993) and other offending psychopaths.[80] Consider the aforementioned McNaughten Rule, used to exculpate the criminally insane (Wolf 1987, 57).[81] It seems, moreover, that any uncertainty about Jojo's blamelessness would rest either on uncertainty about whether he *ever* had the power to do the right thing in the first place (e.g., by questioning his father's authority at the age of reason) or uncertainty about whether he ever really appreciated the reasons against his conduct. We should try to fix in our minds that Jojo has had no such opportunity whatsoever. He really is no better than a zombie in terms of his capacity for moral awareness and could not have been any different. The intuition should then be clearer. The intuition is strong enough that it drove Watson (a former real-self theorist about responsibility

[80] Actual psychopaths, to which I am referring here, should be distinguished from "philosophical" psychopaths, on the grounds that the former *tend* to lack moral competence, whereas the latter lack it *by definition* (Talbert 2008, 2018).

[81] Concerning the link between insanity and responsibility, consider the New Zealand Crimes Act 1961: "Insanity before or after the time when he or she did or omitted the act… may be evidence that the offender was, at the time when he or she did or omitted the act, in such a condition of mind as to render him or her irresponsible for the act or omission."

simpliciter) to concede that there is at least one form of moral responsibility that is missing in cases like *Jojo the Insane* (Watson 1996).

This objection is related to another objection against real-self views, also originally from Wolf (1990, 40–45). As G. Watson helpfully summarises it:

> Holding people responsible is not just a matter of the relation of an individual to her behavior; it also involves a social setting in which we demand (require) certain conduct from one another and respond adversely to one another's failures to comply with these demands. [Real-self] views have largely ignored this context. (1996, 229)

And I think that Wolf's response to this objection is plausible: it is right to demand certain conduct and respond adversely for failures to comply with these demands only if the agent has *normative competence*, or the ability to recognise and respond to "the True and the Good" (Wolf 1990, 124-6).[82] This seems precisely to be what is lacking in *Jojo the Insane*. We have good reason, then, to reject real-self attributability views.

Answerability views hold that to be responsible for some act or attitude is for it to be "connected to [the agent's] capacity for evaluative judgment" (Eshleman 2014) such that it is appropriate or fitting for the agent to justify or "answer for" it (Scanlon 1998, 20, ch. 6; A. Smith 2005, 2012; Oshana 1997). The view has something to be said for it, especially for how it combines elements of attributability and accountability (i.e., the act or attitude is attributable to the agent such that it make sense for us to hold the agent to account) (Eshleman 2014). Alas, answerability views go down with real-self attributability views on the grounds that they pronounce morally incompetent agents responsible for wrongdoing (Scanlon 1998, 283-4). Since Jojo's actions are "open, in principle, to demands for justification" (A. Smith 2012)—in part, because his actions are the kinds of things for which it is fitting to give an account and reflect the (motivating) reasons that he takes to justify his actions (A. Smith 2012, 578, n. 6; Scanlon 1998, 22)—answerability theorists hold that Jojo is responsible. But Jojo is not responsible, as I have argued. Answerability views invite two further worries. One is that they put a strain on the intuitive link between responsibility and *praiseworthiness* by tying responsibility too closely to blameworthiness. Consider that when we say that the moral hero is "responsible" (in the sense implying "praiseworthy") for risking

[82] Wolf (1987) calls this "sanity," but I think that there are good reasons to prefer the term normative or moral competence.

her life to save the child from drowning, we do not, first and foremost, mean *answerable*. After all, why should there be a demand to justify a right or morally outstanding act?[83] The second further concern is that there are some acts or attitudes for which one is responsible but for which one does not take there to be justifying reasons (e.g., in cases of akrasia). For these reasons, I think that we do best to avoid answerability views as well.

5.3 Blame as Moral Faulting: A Cognitive Account

Return to Watson's claim that what is characteristic of holding one another responsible are "demands" for certain conduct and responses of "adverse treatment" to wrongdoers. Adverse treatment—consisting of "censure," "remonstration," "legal sanction" (punishment) as well as expressions of "negative attitudes" (e.g., Strawson's [1993 (1962)] "reactive attitudes" of resentment or indignation)—constitutes "accountability blame" for Watson (1996, 230).[84] Watson agrees with Wolf that fair accountability blame requires moral competence. And this seems acceptable: censure, sanctions, and punishment bring harm to the wrongdoer, but, as such, seem fair only when the wrongdoer could have avoided them, as the morally incompetent could not.

But Watson disagrees with Wolf that this all there is to blame. What real-self views such as Watson's are concerned with is not, as Wolf objects, "superficial blame" or "mere grading" (e.g., judging someone to be "ugly," or "a bad writer"), but a deep form of "aretaic appraisal" which Watson (1996) calls "aretaic blame"—and A. Smith (2012) and Scanlon (1998, 267–94) call "moral criticism." We see this sort of appraisal in judgments of (or beliefs in) people's being "cruel," "reprehensible," "shoddy" or "selfish" (cf. Adams 1985), and these judgments are appropriate just when they are properly *attributable* to the agent. Since morally incompetent agents can still be correctly evaluated by these judgments in light of bad behaviour, the significance of these judgments is that morally incompetent agents may still merit aretaic blame, even if they do not merit accountability blame. As I have hinted already, Watson (1996) ends up arguing that corresponding to these two forms of blame are two "faces" of responsibility: responsibility-as-*attributability*, whose (minimal) conditions are met when the judgments constituting aretaic blame are accurate; and responsibility-as-

[83] Consider, e.g., the way that A. Smith (2012, 577-8) talks of responsibility and blameworthiness in the same breath in the summary of her view.
[84] I agree with A. Smith's (2008, 379) interpretation that these exhaust accountability blame for Watson.

accountability, whose conditions are met when it is fair to subject wrongdoers to accountability blame.

Although I deem Watson's pluralism ingenious, I think that there are good reasons to resist pluralism about responsibility. I think Watson is right to insist that there is something deeper to blame than "accountability" blame. Blame can be private or "inward" (§2.3.2). But should we follow Watson in thinking that inward blame is a form of negative *aretaic* appraisal of the sort appropriate for morally incompetent agents? I think that the answer lies in the negative.

In my view, there is only one "face" of responsibility—the face that denies responsibility for the morally incompetent agent. I reject Watson's view that there are some forms of blame that are merited by the morally incompetent agent. My contention is that there are inward judgments (beliefs) that are not aretaic judgments, which pick out the distinctive core of inward blame, and which account for the intuition that we are not to blame the morally incompetent agent. These judgments also account for the intuition not to judge the morally incompetent agent *responsible,* given the assumption that I have defended that moral responsibility for wrongdoing is materially equivalent with blameworthiness for wrongdoing, and given my argument (set out in §5.6) that the content of one of the blaming judgments implies a judgment of responsibility. Finally, these judgments promise to explain why we get the intuition that the morally incompetent agent should not be subject to "accountability" blame or outward blame in the first place: there is a link between the correctness of these judgments and the fairness of outward practices of holding responsible.

What then is my proposal? It is the following:

> **B:** An agent S (inwardly) blames someone T for an action/omission if, and only if, S believes (propositions that, for S, entail)[85] that (a) T has committed *wrongdoing all-things-considered*, and that (b) T is *morally at fault* for it.

Judging (a) by itself can clearly be appropriate for entirely non-responsible agents (e.g., young children or adults with excuses). Judging (b) by itself can be appropriate for agents of bad actions that are not wrong (and so do not merit blame, given my argument in Chapter

[85] Instead judgment (a), it could be a judgment of doing what they morally should not do, or of violating a moral obligation—or of some judgment that S believes "obviously entails" (a) (see Peels [2017, 33-34] for the notion of personal or "obvious entailment" here). Instead of a judgment (b)—and in light of my argument about the implications of being morally at fault (below)—it could be a judgment of being morally responsible, blameworthy, guilty, having no excuse, for wrongdoing, or even of having full control, or of knowing what they were doing. Here too, (b) needs to be obviously entailed by the agent.

Three §3.3). Together, however, judgments (a) and (b) constitute (inward) blame of someone for an action (or omission). In simple terms, to blame someone for doing something is to pronounce them *morally at fault* for doing what is believed to be wrong, or to judge that their wrongdoing is *their (moral) fault*. Importantly, this is not merely aretaic appraisal. According to Watson, in the aretaic sense, "to blame (morally) is to attribute something to a (moral) fault in the agent" (1996, 230–31). But when I *morally fault* someone for something, I do not necessarily imply that she is generally morally faulty or *has* a moral fault (e.g., a vice of character, or a bad intention). Neither do I merely regard the *act* as morally faulty. Let me suggest that when I *morally fault* someone *for* wrongdoing—or hold that they are *in the wrong* for it—I judge that they played an important causal role in performing the act, and that this *role* was "morally faulty," or else it was conducted "morally faultily." Of course, I agree with Wolf that:

> [W]hen we hold an agent morally responsible for some event, we are doing more than identifying her particularly crucial role in the causal series… [W]e are not merely judging the moral quality of the event with which the individual is so intimately associated; we are judging the moral quality of the individual herself in some focused, noninstrumental, and seemingly more serious way. (1990, 41)

But to account for this judgment of the moral quality of the individual, I think that all we need to say is that the *role* that she played was morally faulty or conducted objectionably. The act is not necessarily attributable to a moral fault *in the agent*, but to a moral fault in the *role* that she played in the reasoning process that led to the act. This does not imply that her role as agent *in general* is morally faulty, only that in the causal genesis of *this* action, she played a morally faulty role.

There is more going for this identification of blame with holding morally at fault. It is quite ordinary and natural to run together the judgments, "I blame her," and "it is her fault," or "he is excused for that" and "that was through no fault of his own." Moreover, this view of blame explains why there has been so much concern over whether there can be freedom and moral responsibility in a deterministic universe. To hold someone like Jojo *morally at fault* for his behaviour connotes that he—and not someone or something else—was the *cause* of the act. But this raises the question of how Jojo can be said to be "the cause," given determinism. Indeed, just as there is a sense of something's being "to blame" which simply means being "the cause" of some effect, so there is a sense of something's being "at fault" which means

the same. It is no surprise that the process of trying to find the *cause* or explanation for something bad is often called "fault-finding." Finally, this link between blame and holding "at fault" is also found in the literature (e.g., Rosen 2003, 63; FitzPatrick 2017, 35).

What is the consequence of this view of blame for the blameworthiness of morally incompetent agents like Jojo? Jojo does merit certain aretaic appraisals like "cruel" or "reprehensible", but if his mistreatment of his subjects issues from his moral incompetence, how does he merit the judgment of being *morally at fault* for his behaviour? It looks more like the "fault" lies in his moral incompetence. Indeed, if, as stipulated in the description of the case, he does not have the capacity for awareness of the reasons against his conduct, it is very difficult to see how there is a moral fault in the role that he plays in the production of wrongdoing. He does not intentionally ignore moral reasons or his moral conscience. He does not act irrationally, given that he does not act contrary to his "deepest ideals." But even if he does, his irrationality is not a *moral* fault in his reasoning, given that it is part of the description of the case that his immoral behaviour is expressive of his deepest ideals. Thus, it seems to me that it would be *incorrect* to judge Jojo as morally at fault, and so *blameworthy*, for his behaviour.

Now, as discussed in Chapter Two (§2.3.2), this *cognitive* view of blame stands in contrast to *emotion-based* views and *conative* (desire-based) views of inward blame.[86] Why then take this cognitive view, as opposed to any of the others? There are, in my mind, decisive reasons against emotion-based and conative views, revolving around the possibility of *blaming dispassionately*. In order to blame someone, I need not feel any "affect" whatsoever—whether anger, indignation, resentment, or the desire that the wrongdoer have done differently. I blame the Christchurch mosque shooter just when I hold him to be morally at fault for the shootings. Although I sometimes feel indignation toward him for it, it would be incorrect to say that I blame him *only because* I feel indignation or wish that he had done otherwise. But neither does my blaming someone require that I have the *disposition* to feel a certain emotion or desire. In fact, I think that most of the time when it occurs to me that the shooter is in the wrong, I do not have any negative emotions toward him or desires about what he should have done at all. That being said, some significance should be attached to the observation (made in Chapter Two) that beliefs consist of dispositions to *feel* certain things to

[86] It also stands in contrast to *functional* views (e.g., where blame is protest or the expression of moral disapproval). For blame as protest, see Talbert (2012); for blame as the communication of dissaproval, see McKenna (2012). I do not discuss these views below, for they seem to fit best as views of *outward* blame. After all, protesting and showing moral disapproval are forms of *behaviour*.

be true, or to have certain *affective* experiences, in certain situations. Beliefs are not just information states. This is significant because it may explain why it is easy to get blame mixed up with an affective state.

Some emotion-based views (unlike those in Rosen 2004, or Levy 2011) wind up also with the wrong verdict about the blameworthiness of agents in cases like *Jojo the Insane.* On M. Talbert's view,

> agents who are constitutionally unreceptive to moral reasons sometimes guide their behavior in accordance with judgments and attitudes that make appropriate the emotional reactions that constitute blame. (2008, 518)

But I have already argued that this clashes against the strong intuition that these kinds of agents cannot be blameworthy. (See my discussion in Chapter Seven, §7.2, however, for a more detailed rejection of Talbert's argument for this position.) Or consider how Shoemaker (2017, 494) treats blame as *anger*—but I can feel anger toward someone who I know could not have done anything to help the situation and so someone who I do not deliberately blame. I think, for example, of the hardened racist whose upbringing and cultural isolation have done no favours to their cultural and moral awareness. I would not be surprised if I heard the following from one of his (more forgiving) victims: "I'm just angry at the man for calling me that. I don't necessarily blame him for it, though—I don't how he came to be so racist."

Other cognitive views also leave something to be desired. I have already dismissed "mere grading" and "aretaic appraisal" accounts. Ledger views hold that blame is about a judgment of discredit, but this view of blame inherits the same objectionable metaphoricity (or metaphysics) of ledger views. P. Hieronymi's (2004, 129ff.) cognitive account of blame equates it to a *judgment of ill will*, but like the aretaic appraisal accounts, this too is developed to pronounce the morally incompetent blameworthy, given that these agents still *display ill will*. Of all the cognitive views, my view is a sibling of the view that blame for wrongdoing is the judgment that the agent is morally responsible, culpable, or has no excuse, for wrongdoing (see, e.g., the one discussed in Pickard 2013; cf. Bryant 2016, 1ff). But B only entails the correctness of these judgments as the outcome of an *argument* that being morally at fault for wrongdoing entails blameworthiness for it, and therefore, given that blameworthiness for wrongdoing is materially equivalent with responsibility for wrongdoing (defended in Chapters Two and Three), responsibility for wrongdoing.

I would like to close my case with replies to three objections. The first is an objection to the implication, just mentioned, that blame for wrongdoing is materially equivalent with judging someone morally responsible or inexcusable for wrongdoing. Is it not possible to judge someone morally responsible, even *blameworthy,* for wrongdoing without blaming them for it? Insofar as one has strictly *inward* blame in mind, I am not really sure that this is possible. However, we should consider D. Justin Coates and Neal Tognazzini's argument:

> [If] you are a co-conspirator in a crime, your partner might be perfectly justified in judging that you acted viciously or wrongly [and that you were blameworthy for it], while simultaneously congratulating you for these things rather than blaming you for them. (2012, 200)

Your partner may not blame you because, for Coates and Tognazzini, she admires your "skilful execution" of wrongdoing (Tognazzini and Coates 2018). But it seems to me that your co-conspirator does not blame you, for their judgment is only a judgment of blameworthiness for wrongdoing *relative to standard morality*. Relative to what would be wrong, and culpable, according to the norms that you accept are governing your joint criminal project, your partner would not be judging you blameworthy for wrongdoing. In other words, your partner would not be judging you *all-things-considered* blameworthy for *all-things-considered* wrongdoing. But this is required on my account.

The second is an objection to the consequence of cognitive views for culpability. Blame*worthiness,* on cognitive views, entails being *worthy* of judgments or beliefs. But how can one be *worthy* of a belief? Being "worthy of" seems to have heavy connotations of being "deserving of" or being "the fair target of"; but how can someone be deserving of, or the fair target of, a *belief*? Indeed, if, as I have been suggesting, holding responsible (and so blaming) is governed by norms of *fairness* which are sensitive to whether or not the target of blame is morally competent, then this is an especially pressing problem. In reply, I think that sense can be made of the language of the desert or fairness of a judgment or belief. "Regard" or "esteem" can be deserved and yet both are plausibly given an analysis in terms of judgments. Moreover, Hieronymi has observed, insightfully, that judgments or beliefs can be regarded as "unfair" when *inaccurate*.

> Suppose, for example, that you innocently and justifiably yet inaccurately conclude that I will default on my loan. I might then claim that your innocent and justified

> belief depicts me in an inaccurately bad light and so might call your judgment 'unfair.' I would not be claiming that you have been unfair in so judging; rather, I would be saying that your judgment is 'unfair' because inaccurate. (2004, 129-30)

Hieronymi acknowledges that sometimes a "fair judgment" or a judgment "fairly made" implies something more like a "reasonable belief." And, as we observed in Chapter Three, reasonable or justified beliefs need not be *true*. But in the above passage, Hieronymi attests to a different sense of "unfair," which is *truth*-tracking. In Hieronymi's case, your judgment of me is "unfair" in the sense of "unfairly suffered," or "undeserved," given that it is *not true*. I suggest that one's being worthy or *not* being worthy of a judgment works in just the same way; one is worthy of it when it is accurate, and unworthy of it when it is false. Moreover, sometimes one's judgment is not reasonable, and yet it is still deserved because it is true. Consider Hieronymi's example of the judge who arrives at a correct judgment of some criminal's guilt but through hasty thinking. Of course, in the eyes of the law, it would not be fair for the criminal to suffer punishment following the faulty procedure by which the judge arrived at the verdict. But the contrite criminal would know in his heart that he got the punishment that he deserves. We are interested in the fairness of a judgment in this *latter* sense.

Thus, given my view of blame for wrongdoing as morally faulting someone for wrongdoing, blame for an act is fair if and only if the act is wrong all-things-considered and the agent really is morally at fault for it. Blameworthiness for an act is the quality of being morally at fault for wrongdoing.[87]

The third and final objection to my (kind of) cognitive account that I will discuss is Hanna Pickard's (2013) objection, that since it is possible to *know* that you are blaming someone *inappropriately*, it follows that blaming someone is not constituted by cognitive judgments. This, to me, is the strongest case against any cognitive account. Consider her motivating case (cut short, so as to restrict the focus to only one of the "blaming" parties):

> A couple come home from work exhausted and stressed, after a long, hard day. One has failed to see to a minor household duty, which annoys and inconveniences the other, but for which, let us suppose, they have a reasonable excuse. Nonetheless, a

[87] Remember, I am using "morally" in the wide sense, to capture all the considerations or types of value relevant to an assessment of blameworthiness. Thus, by a judgment of being morally at fault for wrongdoing, I mean a judgment of being *all-things-considered* at fault for wrongdoing.

> critical remark is uttered, and a blazing row ensues, in which each takes their exhaustion and stress out on the other. The inconvenienced party blames their partner for the inconvenience, even though they know at heart they have a reasonable excuse. (2013, 615)

Since she argues that the inconvenienced party can rightly be said to blame the other, it appears that a cognitive account, which would identify blame with the knowledge of having *no* reasonable excuse, fails to capture what is going on. I confess, however, to having a different intuition about the case. Let me suggest that one of two things is happening. On the one hand, it could be that the inconvenienced party holds contradictory beliefs. The inconvenienced party may rightly be said to blame their partner because they believe that their partner is "at fault" or "in the wrong" for not performing their household duties, but believe also that their partner has a reasonable excuse; and so they hold inconsistent beliefs (at least absent a mistaken theory of how these concepts diverge). On the other hand—and I think that this is more probable—it could be that the inconvenienced party does not really *blame* their partner, even though it seems to make sense on the surface to describe them as "blaming" their partner for the inconvenience. This might make sense because the way that the inconvenienced party is treating the partner would *in normal circumstances* express inward blame—that is, "uttering" a critical remark and "taking it out" in the other. In these circumstances, however, it would not technically be blame (neither inwardly nor outwardly), because of the underlying judgment that the other is excused. Asked afterwards if they blamed their partner, they would likely retort: "Not deep down, no, even though it probably looked like it"—that is, if they "knew at heart" that their partner had a reasonable excuse.

I conclude, then, with a cognitive view of blame, the view that to blame someone for an act is to morally fault them—which means that they played a causal role and played that role morally objectionably—in doing wrong. It is time, now to consider a view of responsibility that can be developed in a way that requires a condition of moral competence and may be consistent with my account of blame.

5.4 Evaluating Accountability Views

Throughout this thesis, I have been working with the notion of responsibility as materially equivalent with accountability, the quality of being worthy of praise, blame, or neutral appraisal. But if moral responsibility is materially equivalent with accountability, this seems to make accountability a very good candidate for what responsibility should be identified with. Moreover, many accountability theorists are moved by Wolf's case for normative competence as a necessary condition for accountability and are among those who have developed the point about blame's being governed by norms of fairness (e.g., Wolf 1990; Wallace 1996, ch. 7; Rosen 2003; Levy 2009). So why not say that responsibility *just is* the quality of being worthy of either praise, blame, or neutral appraisal?

Some considerations in favour of accountability views of the concept of responsibility are already on the table. Not only is moral responsibility materially equivalent with accountability, but, as observed in Chapter Two, accountability views: account for our natural language association between blameworthiness and praiseworthiness on the one hand, and responsibility on the other ("responsible" can be used synonymously with "blameworthy" or praiseworthy"); explain how moral responsibility is to be distinguished from the other responsibility concepts (e.g., causal responsibility, role responsibility, and the virtue of responsibility); and account for the importance of Strawson's "reactive attitudes."

Accountability views have further benefits. They explain why the conditions for moral responsibility may include those introduced in Chapter Two (control, epistemic, quality of will, etc.), since these conditions all seem necessary for praise or blame to be warranted, as the case may be. Another advantage of accountability views is that they are theoretically rich or fruitful. Consider, for example, Shoemaker's (2015) pluralist accountability theory, according to which there are three forms of responsibility (which he, confusingly, calls "attributability," "accountability" and "answerability"), each of which with a distinct class of characteristic "responsibility responses" which (typically) aim at a distinct quality of will displayed in the agent. One of the theoretical advantages to Shoemaker's view is that he is able to account for ways in which agents, with varying degrees and kinds of agency impairment, can be more or less or differently responsible (accountable).[88]

Despite these benefits, I think that it is mistaken to identify responsibility with accountability. There are a handful of considerations against the view, each of which strongly

[88] Accountability theories may also be fruitful in accounting for the relevance to responsibility of wider types of value than moral value (cf. Kauppinen 2018).

suggest a problem that can be framed quite generally with the following: moral responsibility fundamentally concerns the agent and the way that the agent acts—not, as the accountability theorist says, the way that others (or even the agent's own self) should respond to the agent. The agent is accountable for their conduct *because* and *only* because they are responsible for their conduct (cf. Oshana 1997, 80; A. Smith 2012, 577-8).

My first objection in this connection is that accountability views do not explain the intuitive notion that to be responsible for something is *valuable*, or a kind of privilege. To be responsible seems to imply that one is empowered, that one can make a difference to oneself and the world. But now consider the accountability view. How is possessing the disjunctive quality of being either praiseworthy, blameworthy, or neutrally appraisable empowering?

A second objection is that accountability views fit uneasily with my account of blame, even if they are in principle consistent. Recall that my account of blame for wrongdoing is that it is the judgment that one is morally at fault for wrongdoing. But if accountability views hold that the relevant responsibility responses are praise, blame, and neutral appraisal, then one would expect these responses to have a similar nature (*qua* judgments) and similar *agent-centred*, propositional content. But they clearly cannot have the same agent-centred content, for praise cannot be about being *morally at fault* for anything. In my mind (and as I argue below), what unites these responses is precisely the fact that their agent-centred content has a common core—*responsibility*—the special role that the agent plays in the causal history of the act. But the accountability theorist could not say that the common core is responsibility (for then they would invite a vicious circle). And it would be difficult for this theorist to find terms paralleling being "morally at fault" or "in the wrong" for signifying the positive role that the agent played in the performance of praiseworthy conduct.[89]

Two further points about our linguistic use of "responsible" reinforce this point that responsibility is a property about the agent herself, rather than about what responses one merits. First, the *word* "responsible" picks out a quality primarily *of* or *in* the agent *more frequently* than a quality that they have in virtue of appropriate responses to them (consider the causal, capacity, virtue, and role concepts of responsibility discussed in Chapter Two). And second, note that it is very natural for us to say that someone is worthy of praise or blame *because* they are responsible. But then if "responsible" means "worthy of praise or

[89] It is no surprise that accountability theories are typically (perhaps with the exception of Hieronymi 2004) drawn to *emotion-based* theories of blame, indeed theories which give pride-of-place to the quality of will condition, following Strawson (1993 [1962]).

blame" then we have the absurd consequence that being worthy of praise or blame is about being worthy of praise or blame (cf. Peels 2017, 22-24).

5.5 A Reasons-Respondence Attributability View

To avoid these worries, I suggest that we return to the idea that responsibility is a kind of attributability—indeed, the attributability that *grounds* one's aptness as a target of praise, blame, or neutral appraisal.[90] But now we have the following challenge: how can we embrace an attributability view in a way that avoids the ledger view's objectionable metaphoricity (or metaphysical implications), as well as the real-self view's moral competence objection (together with the objections against answerability views)? A natural suggestion may be that we should adopt a version of Wolf's (1987, 382) "sane deep-self view" in which responsibility consists in the act or attitude's revealing a *morally competent* deep-self, excluding psychopaths and agent like Jojo from the morally responsible. Recall that moral competence is the general ability to recognise and respond to the "True and the Good"—or to recognise and respond to the "reasons that there are" (Wolf 1990, 123), the *normative reasons* to act in one way rather than another. Technically, Wolf's (1990, 68) view of the *concept* of responsibility is an accountability view whose *condition* is the revelation of a deep self. Nevertheless, we might wonder whether the relation itself between the morally competent deep-self and their act or attitude is the locus of responsibility.

There are two problems with this suggestion, however. The first is that because it is still a real-self view, it would invite the problem raised in §5.2 that one can be responsible *for* one's real self and so a vicious regress (or circle) looms. The second problem is more pressing: it is possible for wrongdoing to reveal the morally competent real-self (e.g., their fundamental values) and yet fail to be something for which the agent is *morally at fault* and so blameworthy. If that is the case, and if I am right that blameworthiness for wrongdoing is being morally at fault for wrongdoing, then the material equivalence of responsibility for wrongdoing and blameworthiness for wrongdoing **(**Rw↔Bw) is undermined; in this case, it would undermine (Rw→Bw), the claim that responsibility for wrongdoing entails blameworthiness. Given the intuitive plausibility and defensibility of **(**Rw↔Bw), I think that

[90] I should note here that I do not mean to align myself with what has been called *attributionism* (Levy 2005), the idea that the key *condition* on responsibility is merely something like the real-self or quality of will condition (e.g., those views discussed next Chapter).

we should avoid upsetting it. But why think that sane deep-self views have this cost? In §5.3, I argued that being morally at fault for wrongdoing entails that in the reasoning processing that led to the action, she played a morally faulty role—the agent conducted that role morally objectionably. But there seem to be cases in which wrongdoing reveals a morally competent real-self but where the "self" played no morally objectionable role in the aetiology of the act. I have in mind certain extreme cases of moral ignorance which E. Harman (2015) calls "being caught in the grip of a false moral view." Consider the Korowai cannibal, who while generally morally competent (for cannibals are surely not morally insane *by definition*), *never* had a chance to question the received wisdom about killing and eating male witches, and *never* entertained any doubts about the moral permissibility (indeed, in some circumstances, *moral duty*) of doing so. He is now mistakenly morally certain in its obligatoriness (under certain conditions).[91] This is someone, let us suppose, who never acted contrary to his moral conscience, always acted with apparent wholeheartedness or in the belief that he was doing the right thing when eating male witches. It is hard to see how he could have been morally at fault for his wrongdoing and so blameworthy for it. But the sane real-self theorist must say that he *was* blameworthy for his wrongdoing (as long as they accept (Rw→Bw), as I have argued that anyone should), because he was morally competent and his cannibalism revealed his real-self (i.e., his "second-order desires" or his "evaluative stance").

The problem seems to be that the cannibal has not exercised his moral competence *over his moral ignorance*. After all, having the capacity does not mean that it is always exercised. And it seems that if he *had* exercised this capacity in the management of his cannibalistic beliefs, and yet failed to prevent ensuing ignorance, he might well have been morally at fault for his cannibalism. My proposal, then, is that we locate the agent's moral responsibility for wrongdoing in *the very exercise* of their moral competence during the process that led to that wrongdoing. Given that moral competence is about the capacity to recognise and respond to moral reasons, a comment by Douglas Husak supports this move:

> It is natural to suppose that questions about whether persons are responsible for conduct… should be resolved by invoking the same framework that shows why they possess the capacity for responsibility in the first place. That is, agents become eligible for attributions of responsibility for a given act by exercising (or failing to

[91] Some say that the Korowai still practice cannibalism to this day (Raffaele 2006). Some say that the Korowai believed that it was their duty to consume male witches.

> exercise) their capacity to be reason-responsive with respect to it. (2016, 147)

The upshot is a "reasons-responsive" view of (the conditions on) responsibility, which receives considerable support in the literature (Wolf 1990; Fischer and Ravizza 1998; Nelkin 2011; Nelkin and Rickless 2017; Husak 2016, chap. 3; McKenna 2013; Coates and Swenson 2013; Sartorio 2016; see also this view about doxastic responsibility in: Steup 2011; Ryan 2003). A plausible way of analysing the exercise of our capacity to be reasons-responsive is along the lines of Dana Nelkin's "rational abilities" view (following Wolf 1990). On Nelkin's view:

> people are responsible when they act with the ability to do the right thing for the right reasons, or a good thing for good reasons. (2011, 7)

Like Wolf (1990, 68), Husak (2016, 143) and Nelkin (2011, 7) hold that the *concept* of responsibility is to be identified with what I have called accountability, and so the above remarks concern accounts only of the key conditions that ground accountability on their view. Nelkin, for instance, is explicit that hers is an account of *control* (Nelkin and Rickless 2017, 118). But since we are looking for a way of characterising responsibility in terms of attributability, why not say that moral responsibility consists just in that very "control"?

One obvious issue here is that precisely what counts as responsibility-relevant control has been the object of endless debate, especially between compatibilists and incompatibilists.[92] This divide also cuts across reasons-responsive conceptions of control.[93] However, I do not believe that we must wed ourselves to one of these conceptions in order to identify responsibility with the attributability involved in "the exercise of reasons-responsive control," construed broadly enough. Moreover, it is surely an advantage of an account of the concept of responsibility that it is specified in such a way that it remains open to either compatibilist or incompatibilist analyses, at least *in advance* of an argument for either compatibilism or incompatibilism.

How then shall we formulate a reasons-responsive attributability account of the concept of responsibility, so as to remain neutral between compatibilism and incompatibilism, and to

[92] For a good review of this debate, see Levy and McKenna (2009).
[93] Reasons-responsiveness views are *typically* compatibilist (see, e.g., Wolf 1990; Fischer and Ravizza 1998; Nelkin 2011, ch. 3; Sartorio 2016; McKenna 2013; Coates and Swenson 2013; Steup 2011; Ryan 2003). However, reasons-responsive control can easily be given an incompatibilist analysis (see discussion in Nelkin 2011, ch. 3; and McKenna and Coates 2019).

account for everything else on the table (e.g., the arguments against the other views)? My proposal is that we adopt a *reasons-respondence* account of responsibility:

> **Responsibility as Reasons-Respondence**
> Responsibility for some act consists in the act's being the agent's *actual response* to the relevant normative reasons for and against it.

What reasons are the "relevant reasons for and against the act" I shall save for below. For now, notice that the key notion in this view is the agent's "response" to the relevant reasons. The response establishes actual respondence to reasons (the quality of responding), not just the manifestation of one's general moral competence or reasons-responsive dispositions.[94] Shifting the focus to the outcome, or the response, prevents any straightforward implication about the requisite abilities, capacities, control, or the process that led to the response—thereby securing neutrality between compatibilism and incompatibilism. It must only make sense to describe the outcome as a response to the relevant reasons. And to the possible worry that the proposed view still suggests leeway incompatibilism on the grounds that "a response" of wrongdoing implies a *choice* while able to do otherwise, I cite compatibilists who have been happy to talk of the act's continuing to be a response to reasons on their views (see, e.g., McKenna and Coates 2019; McKenna 2013, 175; Steup 2011, 551, 555).

Of course, for any behaviour to count as an action in the first place, it has to be done "for reasons." And for that behaviour to count as an exercise of *moral agency*, it is plausible that the act has to be done for *moral* reasons (i.e., reasons that are morally appraisable). But notice that, on the proposed view, *moral responsibility* is importantly different from moral agency, because the reasons with respect to which responsibility is cashed out are *normative* reasons—the reasons which *count in favour of* or against some act—not the reasons that, as a matter of fact, and regardless of the agent's awareness, *explain* one's conduct (e.g., one's desires, motivations, emotions, or intentions; see §3.4.2).

I turn now to the question of what normative reasons are the "relevant" reasons to which the agent must have responded, a question for any broadly reasons-responsive approach to

[94] Commenting on reasons-responsive approaches to responsibility, Gideon Yaffe (2018, 343) says that: "According to all such theories, an agent is responsible for an act just in case the act manifests the way the agent recognizes, weighs, and responds to reasons for and against the act." My view is more like the following (although not quite, for reasons that will become clear in the next few Chapters): an agent is responsible for an act just in case the act manifests the way that the agent *has recognised, weighed, and responded* to the relevant reasons bearing on the act in question.

responsibility. I contend that the relevant reasons are the reasons that determine the moral status of the act being assessed. Thus, supposing that one can be responsible for a morally good act, responsibility for that act consists in the act's being the agent's response to the reasons in favour of it, concerning its moral goodness. Since we are interested in wrongdoing given its unique relevance to blameworthiness, the view entails the following about responsibility for wrongdoing:

> **Rw:** An agent's moral responsibility for wrongdoing consists in its being the agent's (poor) response to the reasons to avoid the act/omission, concerning its wrongfulness.

The notion of "response" in **Rw** might immediately sound odd, in contrast with its use in respondence to good-making reasons. Does a response to reasons not mean "action *in accordance with* reasons"? I hope to alleviate this worry with the qualification of the response as "poor" in parentheses—which is important for the explanation of blameworthiness (see below). The sense of "poor" connotes something like "incorrect." As Douglas Husak puts it, the agent responds "*incorrectly* to the balance of moral reasons" (2016, 153, my emphasis).[95] In consequence, she "[exhibits] a defect or corruption in her response to moral reasons" (p. 154).

But I suggest that even without this qualification, sense can be made of a *contrarian* response to reasons. It is no abuse of language, after all, for "response" to have a contrarian implication. In academic philosophy, a response to a piece of work is more likely than not going to be an *objection* or a *rejection* of it. Someone's hostile emotional reaction can be described as their kneejerk "response" to something with which they strongly disagree. Accordingly, I take it that there are *two* ways of responding to reasons with an action—*for* and/or *against* the relevant reasons.

But even *qua* account of responsibility for wrongdoing, it might be wondered why **Rw** is expressed negatively in terms of the agent's response to the reasons *to avoid* wrongdoing. Why is the wrongdoing not a poor response to the reasons governing *whatever they morally ought to do* in the circumstances? Sometimes a poor response to reasons to avoid wrongdoing will be a poor response to the reasons governing whatever they ought to do, but it seems that

[95] The reason why **Rw** is not put simply in terms of something like the act's being a response to "the balance of moral reasons" is that this description is not fine-grained enough. Must the agent know all the reasons? Which reasons must the agent be responsive to? I should also note that Husak holds this to describe the *corrupt deliberation* apparently necessary for blameworthiness. In Chapter Seven, I argue that responsibility does not require deliberation; habitual or automatic activity can count as a response to normative reasons.

this need not be the case. Suppose that Bill could either (a) lie to his wife, (b) omit the whole truth, or (c) tell the whole truth, where telling the whole truth is what is morally best and what he *ought* morally to do. (Omitting the whole truth would simply be failing to do what he morally ought to do, while lying would be doing what he morally ought not to do.) It is plausible that in order for Bill to be responsible for lying to his wife, it need only be his response to the reasons against (or not to) lie to his wife, and not necessarily his response to the reasons in favour of telling the whole truth. The moral of the story is that reasons against wrongdoing sometimes do not determine *which* alternative to take when there is more than one permissible alternative.[96] Similar stories could be told about variants in which the agent responds poorly to reasons to avoid doing whatever is wrong or what one morally ought not to.[97]

That the reasons in **Rw**, responsibility for wrongdoing, must be reasons concerning the conduct's wrongfulness should be clear: there can be morally insignificant reasons to avoid wrongdoing (e.g., when it would serve my interests not to harm someone), and it seems plausible that a response to *those* reasons exclusively fails to make someone *morally* responsible. Even the Korowai cannibal would at times have had purely prudential reasons not to eat male witches.

What is significant for the remainder of the thesis is the following claim: that **Rw** *entails a minimal degree of awareness of the wrong-making normative reasons*. After all, it makes no sense to say things like, "She is responding to a reason for or against acting, but she has no idea what that is, even deep down," or "That's his *response* to the reasons for or against doing so, but he is not in any way sensitive to those reasons." As Husak (2016, 188) has put it, "it is hard to see how rational individuals can be expected to respond to reasons of which they are unaware." It may be alleged that this view conflicts with one way that a writer has characterised responses to reasons—namely, Nomy Arpaly (2002, 2015). For Arpaly, the fictional character Huckleberry Finn, who helped Jim escape from slavery while believing that "the right thing" to do would be to turn Jim in, may nevertheless be praiseworthy, and responsible, for freeing him, despite his acting *against* his conscience. (This is what she calls

[96] Since I argued in Chapter Three that a wrong option implies at least one right alternative, I hold that in cases of two or more permissible alternatives, it is minimally the case that reasons against wrongdoing count overall in favour of doing any one of the permissible alternatives (as long as those alternatives do better with respect to the specific values that those reasons promote; cf. Snedegar 2018). Sometimes, however, those reasons will decide in favour of one of those alternatives (if, e.g., the reason why Bill's lying to his wife is wrong is also the reason why telling her the whole truth is right).

[97] In these kinds of case, the problem is that there may be more than one wrong option, and it seems that responsibility for doing any one of them need not require a poor response to the reasons not to perform the other wrong alternatives.

"inverse akrasia.") What makes Huck responsible for saving Jim, thinks Arpaly, is the fact that Huck frees Jim *in response* to reasons relating to Jim's humanity or personhood (Arpaly 2015, 142), even though he is unconscious of them. Now, Rudy-Hiller (2018) has interpreted Arpaly here as endorsing the claim there can be response to reasons *without awareness* of them. He would be right that the reasons are not *occurrently* aware for Huck. But Rudy-Hiller fails to notice that Arpaly (2002, 77) may still require that Huck has "dim" awareness of the reasons relating to Jim's humanity (no doubt, e.g., some non-negligible credence in Jim's humanity). Thus, it would be a mistake to interpret Arpaly as rejecting this inference from responses to reasons to *awareness* of those reasons.

But if so, and if responsibility for wrongdoing entails blameworthiness (as I have argued), then **Rw** entails *culpability internalism*, the view that blameworthiness for an act requires beliefs or credences concerning the wrongfulness of the act. In particular, it entails a version of culpability internalism according to which blameworthiness requires *true* beliefs or credences of the moral significance (or morally significant features) of her conduct. On my view then, and whatever the process is that results in the agent's "response to reasons" (paradigmatically, it is *deliberation*),[98] it is part of the process that there is an original "input" of the relevant reasons in the form of an epistemic state (a belief, credence, or motivating reason), as well as an eventual "output" in the action or intention to act that constitutes a response to those reasons. There are no responses to reasons without awareness of them.

Let me close this Section with some remarks in favour of the view. Some of its existing motivation should already be clear from our discussion. It finds a way of remaining neutral between compatibilism and incompatibilism, and it avoids the problems of the (attributability, answerability, and accountability) theories previously discussed. In relation to the particular problem of accountability views that they do not account for the *privilege* of being responsible, notice that the view offers a satisfying account: it is privilege to be responsive to reasons—for then one has the chance to do what is right or good. In this connection it also has a satisfying account of why people are worthy of praise: they have responded *well* to the reasons to do the right (or a good) thing.

What about the link between blameworthiness (being morally at fault) and responsibility for wrongdoing? Well, it seems clear that responding *poorly* to the relevant reasons not to perform wrongdoing entails being *morally at fault* for wrongdoing. The former is sufficient for the latter. Even stronger, I propose the following:

[98] But see n. 95.

Someone is morally at fault for wrongdoing if *and only if* they responded poorly to the reasons to avoid wrongdoing, concerning its wrongfulness.

This, after all, preserves **(**Rw↔Bw), the material equivalence between blameworthiness (on the left-hand side) and responsibility (on the right-hand side) for wrongdoing. Indeed, I think that it is what *explains* **(**Rw↔Bw). But why is a poor response to reasons *necessary* for being morally at fault? Consider the various ways that we have described being morally at fault for wrongdoing—for example, that it is about playing a morally objectionable role in the reasoning process that led to the act, that it may involve intentionally ignoring moral reasons or one's moral conscience, that it may involve irrationality, and that it is the opposite of acting wholeheartedly or conscientiously. I do not see how it is possible for these descriptions to *not* apply to (or imply) responding poorly to the relevant reasons in the causal history of the act.[99] That, at least, is my presumption. (In the next few Chapters, I will consider some cases that put this material equivalence to the test.)

5.6 Indirect Reasons-Respondence as Indirect Responsibility

To round out the proposed account of responsibility, we should consider an objection that could be raised against it. The objection is raised by Gary Watson (2001, 374f.) against Fischer and Ravizza's reasons-responsive account of the control condition on responsibility. Initially in defence of a reasons-responsive approach, Watson admits that:

> The idea that moral responsibility is crucially connected to the capacity to respond to reasons is a natural one. It is not an accident that the 'age of reason' appears to coincide with the age of responsibility. (p. 375)

But then he notes that: "[in] one sense, being responsive to reasons is just what it is to be a *reasonable* creature" (p. 375, my emphasis); however:

[99] Note that acting in a morally unconscientious way need not imply moral wrongdoing on my view (cf. §3.3.5), for acting morally unconscientiously only implies acting contrary to one's moral conscience—contrary to the balance of one's moral beliefs or credences whether or not they are *true*. But the same is true for being morally at fault for an act: it does not imply wrongdoing (cf. my discussion of B in §5.3). Notice that our focus here is on being morally at fault *for wrongdoing*, and likewise we are considering cases of committing a *wrong* act morally unconscientiously. Accordingly, responsibility for wrongdoing is about poor responses to *normative* reasons, not (just—for I will argue that it ultimately does require) responses to one's *own* (motivating) reasons.

> [this] points to the most immediate problem for any view that links responsibility tightly to reasons-responsiveness. We blame people precisely for their insensitivity to reasons, so reasonableness and responsibility are not to be equated. (p. 375)

Leaving aside his equation of reasons-responsiveness and reasonableness "in one sense," the objection is that sometimes we are responsible and blameworthy for failing to be sensitive to reasons. If we can be responsible for a failure to be receptive (and thus responsive) to reasons, responsibility for an act cannot be the quality that the agent enjoys when the act is their poor response to reasons. Now, to some extent, this objection is what is at stake in my discussion next Chapter of responsibility *externalist* responses to the Regress Argument. Nevertheless, let me here sketch a response to Watson by appealing to an "easy" case of responsibility for insensitivity to reasons so as to explain how I think that a reason-responsive account is still able to handle cases like these.

Consider William Clifford's (1886) classic shipowner case:

> A shipowner was about to send to sea an emigrant-ship... Doubts had been suggested to him that possibly she was not seaworthy… [He] thought that perhaps he ought to have her thoroughly overhauled and refitted, even though this should put him to great expense. Before the ship sailed, however, he succeeded in overcoming these melancholy reflections… He would dismiss from his mind all ungenerous suspicions about the honesty of builders and contractors. In such ways he acquired a sincere and comfortable conviction that his vessel was thoroughly safe and seaworthy; he watched her departure with a light heart…; and he got his insurance-money when she went down in mid-ocean and told no tales. (1886, 1)

We should agree with Clifford that the shipowner is to blame for the loss of life as well as for deciding to send it to sea. *At the time* of sending it out to sea, however, we can suppose that the shipowner was *not* sensitive to the reasons why it was wrong to send it to sea (namely, its putting the lives of passengers in jeopardy for no good reason). After all, he had "acquired a sincere and comfortable conviction that his vessel was thoroughly safe and seaworthy." But does it follow that his sending the ship to sea was *not* his poor response to reasons why it was wrong, and so *not* (on my reasons-respondence account) something for which he was responsible and blameworthy?

We could take two strategies here in order to capture the intuition of blameworthiness. The first strategy would be to help ourselves to the notion that the shipowner *would* still have sent the ship to sea, *had* he been aware of the reasons why it was wrong to do so, and that, if so, the shipowner *was* responsible and blameworthy for sending it (Timpe 2011, 20-21; Guerrero 2007, 62-63).[100] (In other words, the shipowner did not act *from* ignorance but *in* ignorance; recall how volitionists disagree on whether there must be action *from* ignorance: §4.3.1.) This counterfactual may be true, but apart from the fact that this strategy seems wrongly to make actual responsibility dependent on counterfactual responsibility (and would force us to revise our account of blameworthiness and responsibility so that it entailed actual and *would-be* responses to reasons), there is I think a decisive reason against it, given by Rosen (2008, 598 n. 14). Take Rosen's case in which Applebaum poisoned Botstein without realising it, because someone had put arsenic in the sugar jar. Suppose, however, that Applebaum *would* have poisoned Botstein this way had Applebaum known about the arsenic anyway. Perhaps he was planning to murder Botstein later that day. Is Applebaum responsible for Botstein's death? It seems obvious to me that the answer is no, because "as it happens, Botstein's death was just an accident" (p. 598 n. 14). Rosen acknowledges that "Applebaum may be open to moral criticism for his dispositional willingness to kill. But he is not culpable for *killing Botstein* in this version of the case." It does not matter what Applebaum's response to the (reason-giving) fact that there was arsenic in the sugar jar *would* have been.

The more plausible strategy, and the strategy that Clifford himself might have wanted to endorse,[101] is the strategy of establishing a *tracing* explanation of the shipowner's responsibility. To a greater or lesser extent, this has already been implicit in our discussion of Jojo and the Korowai cannibal above. According to this strategy, the shipowner's risking the passengers' lives for no outweighing reason *was* his poor response to the reasons not to send it to sea but only his *indirect* or *derivative* response to those reasons—his eventual response to the reasons that entered his deliberation *before* he convinced himself of the ship's seaworthiness. In other words, at the time of sending the ship out to sea, the shipowner did not *directly* or *originally* respond to the reasons against so doing, however his act was traceable to a benighting episode of self-deception in which he *directly* responded (poorly) to the reasons not to send the ship to sea before overhauling it. My claim, then, is that indirect

[100] Given that the shipowner case is a case of unwitting conduct, Timpe (2011, 19) would note that this is an instance of what Thomas Aquinas called "concomitant ignorance."
[101] His focus was on the shipowner's violation of a duty to question his beliefs; cf. 1886, 4.

responses to reasons are always traceable to *direct* responses to those reasons, and this is why indirect responsibility is always traceable to an instance of direct responsibility.[102]

Now, it appears that for the shipowner's sending the ship to sea to be his indirect response to the relevant reasons against doing so, a number of conditions must be fulfilled. We have already seen that tracing explanations plausibly require *counterfactual dependence* and *explanation* (see Chapter Two §2.3.5). But more importantly, two further conditions must be fulfilled at the time of the "benighting act" of self-deception in order to plausibly count as responding to the reasons against the eventual act (of sending the ship to sea):

(a) the shipowner must *foresee* the consequence of sending the ship to sea without overhauling it, and the reasons against doing so;
(b) the reasons to avoid sending the ship to sea without overhauling it must count as reasons to avoid his act, now, of self-deception.

Condition (a)—which I call the *foresight requirement*—follows from my argument above that responses to reasons require minimal awareness of reasons. To apply this consistently to future possible acts, we are forced to say that poor responses to reasons against *them* requires foresight of those reasons. But, as we saw in §4.4.2, the foresight requirement also gets support from the general claim, defended by Zimmerman, that responsibility for any consequence *C* requires foresight of *C*, as well from widespread endorsement in the literature.

Condition (b)—which I call the *reasons-transitivity requirement*—follows from the fact the agent's initial act must count as a *response* to the reasons to avoid the later act. If the reasons against the later act had no bearing on the moral status of the initial act, then the initial act could not intelligibly be said to be a response to those reasons. This is not a surprising condition. Consider after all that the reasons for the shipowner not to send a ship to sea before overhauling it are reasons against self-deception about the ship's seaworthiness (i.e., doing so could put emigrants' lives in jeopardy).

Now precisely what form one's foresight of the reasons must take for responsibility is the subject of Chapter Eight. There is a disagreement between those who believe that, for responsibility, the foresight of the act must be quite fine-grained (e.g., foresight of

[102] On K. Timpe's (2011, 20-21) view, non-direct responsibility depends *either* on a tracing explanation *or* on explanation in terms of counterfactual responsibility. As such, he might try to get around Rosen's objection to the possibility of the latter explanation, by holding that for *some* cases the latter is needed, while for other cases (e.g., the Applebaum case) tracing would be required. But without an argument for *which* cases ought to be treated in which way, I will presume that Rosen's reply succeeds.

specifically damaging that person's car whilst intoxicated) as opposed to quite coarse-grained (e.g., foresight of general recklessness). In Chapter Eight, I will side with the latter view. There, I will also address the issue of how exactly the normative reasons against the foreseen (coarse-grained) act must be represented in the agent's mind, following the account I give (in Chapter Seven) of the way that they are represented in the agent's mind when directly responsible.

I would like to wrap up this Section with two further remarks about this distinction between direct and indirect responses to reasons. Note, to begin with, that indirect responsibility and blameworthiness come apart in cases where the original response to reasons was not itself *wrong*. Borrowing an adaption from E. Curley's (1975), suppose that countless emigrants were fleeing from intense religious persecution and could only get out by ship. It would seem, then, that the shipowner's omission to overhaul the ship would not have been wrong, even though he foresaw the possibility of his ship going down mid-ocean. But suppose that in between omitting to overhaul it and setting sail the next day, the religious persecution was put to an end, albeit without his or his passengers' knowledge (or without any reasonable suspicion or culpable ignorance). Plausibly, going ahead as planned would then have become wrong in the circumstances. Nevertheless, he surely could not have been blameworthy for sending the ship to sea, even though he foresaw it and the reasons against it; he did not do anything wrong when he had the relevant foresight. Still, it seems that partially because the shipowner foresaw the risk of sending an unseaworthy ship to sea, he would have been indirectly responsible for sending the ship to sea, if only minimally responsible (akin to Sophie's responsibility for choosing which of her children to die to save the other). Thus, for any agent to be indirectly *blameworthy* for wrongdoing, they must be directly blameworthy for the initial act risking that (future possible) wrongdoing, and so the initial act must itself be wrong, given that culpable conduct is always wrong (Chapter Three §3.3). Finally, given that blameworthiness implies responsibility, the initial act must itself be the agent's direct response to the reasons to avoid *it* concerning *its* wrongfulness, and these reasons may well (and usually will) include more than just the reasons concerning risking future wrongdoing.[103]

Lastly, it is worth saying something briefly about *which* objects of responsibility count as direct or indirect responses to reasons, given the importance of this for subsequent Chapters.

[103] After all, many acts increase the risk of wrongdoing without being wrong (driving, drinking, befriending strangers, and so on).

This Chapter has focused exclusively on responsibility for (wrongful) *actions or omissions*. And in this Section, I have argued they can be either direct or indirect responses to reasons. In my view, many other things can be *indirect* responses to reasons—for example, beliefs, desires, moods, character traits, and external consequences of actions (i.e., events or states of affairs). Even stronger, it is my view that these things *cannot* be *direct* responses to reasons. At any given time *t* in which these consequences occur, or when the agent finds themselves "in" these passive states, these events or states cannot be the agent's *direct* response to reasons *at t*. This is because responding to reasons is about playing a causal role in bringing about something in response to normative considerations, but the agent could not have caused (the production/retention/occurrence of) these states/events in response to the reasons for or against them *at that time t*. Responsibility for all of them requires that the agent first exercised (or had the opportunity to exercise) their will to cause them, at that must have happened at a prior time *t-1*; they are the *effects* of actions/omissions. Thus, I am construing direct response to reasons as forms of *direct causal control.*[104] And my contention is that only *actions or omissions* at a given time *t* can, I believe, be direct responses to reasons at *t*.

With the distinction between direct and indirect responses to reasons, we seem, then, to have a satisfactory account of direct and indirect responsibility and an adequate reply to Watson. In the next Chapter, I will consider more formidable pressure against the view on the part of the responsibility externalists.

5.7 Conclusion

Once we get clear on the nature of blame, blameworthiness, and responsibility, we see that a "valence-tracking" version of culpability internalism follows. Blame for wrongdoing is the judgment of being morally at fault for wrongdoing, and so (given the notion of "fairly suffered" judgments), blame*worthiness* for wrongdoing is the quality of being the *accurate target* of that judgment. But one cannot be morally at fault for wrongdoing without playing a morally objectionable role in the way that the wrongdoing was brought about. And most importantly, one cannot play this morally objectionable role *without responding poorly* to the reasons not to bring about that wrongdoing. That is, one cannot play this morally objectionable role without *moral responsibility* for the resulting wrongdoing. But, of course, to *respond* to reasons is to be at least minimally *aware* of them, and so responsibility for

[104] Although whether this must be *voluntary* control is another matter; see my discussion in Chapter Seven.

wrongdoing, and thereby blameworthiness, entails minimal awareness of the wrongfulness of the act (i.e., culpability internalism).

Chapter 6
Resisting Responsibility Externalism

6.1 Introduction

In the last Chapter, I argued for a certain understanding of the concept of responsibility as *reasons-respondence*, from which I derived the thesis of culpability internalism—the view that blameworthiness for an act requires beliefs or credences about the moral significance (wrongfulness) of the act. We have thus already gone a long way toward rejecting the contrary thesis, responsibility externalism. But to conclude our case for culpability internalism (and defence of the reasons-respondence view), we should consider three forms of responsibility externalism about the epistemic condition: those that Rudy-Hiller (2018) has identified as "epistemic vice" theories, "quality of will" theories, and "capacitarian" theories. These are not views directly about the *concept* of responsibility, but first and foremost about its *epistemic* condition. What unites this family of views is the denial that responsibility requires beliefs or credences about the wrongfulness of the act. Accordingly, each type of view produces replies to the volitionist's Regress Argument.

For each type of externalist view, I will briefly introduce its response to the Regress Argument, and the characteristic arguments made in favour of it. Along the way, I will also add to my case for culpability internalism by fleshing out my theory of responsibility—adding, for example, an account of fair expectations—and by pointing out unique problems for each type of externalist view.

Before we consider externalist replies to the Regress Argument, let us restate the argument here:

The Regress Argument

(1) An agent S is blameworthy for performing all-things-considered wrongdoing *A* only if S does *A* contrary to S's true occurrent belief that *A* is wrong all-things-considered, based on true occurrent beliefs that *A* has the features that make it wrong (i.e., *A* is

"fully advertent" wrongdoing), or S does *A* in/from culpable ignorance of *A*'s all-things-considered wrongfulness (i.e., *A* is "culpably unwitting" wrongdoing).

(2) S is culpable for the ignorance in/from which S does *A* only if S's ignorance is the foreseen upshot of performing a "benighting" act or omission *B* for which S is blameworthy.

(3) S is blameworthy for performing a benighting act *B* only if *B* is either fully advertent or culpably unwitting wrongdoing.

Therefore,

(4) An agent S is blameworthy for performing all-things-considered wrongdoing *A* only if *A* is fully advertent wrongdoing, or the ignorance in/from which S does *A* is the foreseeable upshot, ultimately, of fully advertent wrongdoing (*B*, or *C*, or *D*, etc.).

6.2 Resisting Epistemic Vice Theories

Two externalist replies to the Regress Argument go by the name of epistemic vice theories—due to James Montmarquet (1993, 1992, 1995, 1999) and William FitzPatrick (2008, 2017). According to these replies, Premise (1) is true, but blameworthiness for unwitting wrongdoing can sometimes bottom out in the voluntary exercise of *epistemic vices* (e.g., closed-mindedness, carelessness, or inattentiveness), even if their exercise is not in any way advertent. This therefore makes epistemic vice theorists responsibility externalists. Although they share this idea in common, Montmarquet and FitzPatrick primarily target different Premises of the Regress Argument. Montmarquet challenges Premise (2) while FitzPatrick primarily challenges Premise (3). FitzPatrick also challenges Premise (2) but for a different reason than Montmarquet (as we shall see).

I turn first to Montmarquet's case against Premise (2). Montmarquet targets Premise (2) on the grounds that culpability for the ignorance in or from which one performs wrongdoing can be direct, when it exhibits epistemic vice. In all other cases, I presume that he thinks that culpable ignorance bottoms out in culpability for fully advertent benighting wrongdoing (or for some other instance of direct culpability for ignorance).[105] How can culpability for any instance of ignorance be *direct*? And how do exercises of epistemic vice play into it?

[105] Montmarquet does not consider cases of fully advertent benighting wrongdoing, but his argument for grounding culpable ignorance in directly culpable exercises of epistemic vice is based on a consideration of cases in which the benighting conduct is itself unwitting.

On Montmarquet's view, culpability for ignorance never bottoms out in an unwitting benighting act, because for every unwitting benighting act there is an ignorant belief relative to which it is justified (Montmarquet 1992, 333) (and if an act is justified by a belief, then the act cannot be directly but indirectly culpable through culpability for the belief). The problem if we are to find an original source of blameworthiness is then the problem of having to account for culpability for the belief without reference to any unwitting benighting act that led to that belief. Montmarquet's solution is to appeal to *the voluntary exercise of epistemic vice* over one's beliefs; and this, he thinks, constitutes a form of "direct (albeit incomplete)" control over one's beliefs (Montmarquet 1999, 844). In particular, Montmarquet argues that we exercise this control over beliefs by having direct control over *the way* that our beliefs are formed, rather like our control over the way that we conduct an activity like carving wood or running.[106] If we fail to exhibit "care" in our belief-forming activities (like failing to exhibit care in carving), we may be directly blameworthy for resulting problematic beliefs.[107]

In defense of the appeal to epistemic vices, Montmarquet attaches significance to Zimmerman's (1997) admission that the natural thing to which we appeal to explain Perry's culpability for unwittingly paralysing Doris, the car crash victim, is Perry's "carelessness," "inconsiderateness," or "inattentiveness" in forming the belief that pulling her out of the wreckage would be the right thing to do (Montmarquet 1999, 842). After all, as the case is described by Zimmerman, when Perry comes onto the scene:

> [a] vision of wrecked cars catching fire and exploding into roiling balls of flame fill his mind, and he feels that he must rescue the driver now or else [Doris] will surely die. So, with considerable trepidation, Perry rushes in and quickly drags Doris free from the wreck, thinking that at any moment both he and she might get caught in the explosion. (Zimmerman 1997b, 410)

Zimmerman has in mind Perry's carelessness in failing to "entertain the possibility of doing more harm than good by means of a precipitate rescue" (1994, 416). For Montmarquet, this is the locus of Perry's direct control over his belief-formation, and his exercise of that control

[106] The running example is from Montmarquet 1999, and the carving example is from Montmarquet 1993.
[107] Such a view should be distinguished from what appears to be a variant of the quality of will view (discussed below), on which a false belief can be blameworthy for the reason that it displays the agent's epistemically vicious character (e.g., her gullibility; see Owens 2000, 124). Montmarquet introduces an element of control or voluntariness into his account (absent in Owens'), and it is not clear whether Montmarquet requires that the agent actually exercised a *possessed character trait*, or simply an epistemically vicious motive or way of acting.

exhibited epistemic vice; so, according to Montmarquet, Perry is directly blameworthy for his false belief that Doris should be pulled out of the wreckage.

The view is externalist, because contrary to Zimmerman, Montmarquet would not require for Perry's culpable ignorance that Perry is *aware* of his carelessness or inconsiderateness or his failure to be open-minded to "the possibility of doing more harm than good":

> The root idea here, it seems to me, is that a certain quality of *openness* to truth- and value-related considerations is expected of persons and that this expectation is *fundamental*, at least in the following regard. The expectation is not derivative of or dependent upon one's (at the moment in question) judging such openness as appropriate (good, required, etc.)—just the opposite: it would include a requirement that one be open to the need to be open, and if one is not open to this, one may be blameworthy precisely for that failure. (Montmarquet 1999, 845)

What shall we make of his view?

Although I think that Montmarquet produces an interesting reply to the Regress Argument, I take there to be decisive objections to his argument for epistemic vice theory. His only hope, I argue, is to join forces with FitzPatrick and revise his view so that it targets Premise (3). Apart from the dubious point that every unwitting benighting act has a justifying belief,[108] Montmarquet's argument relies on the faulty premise that exercises of epistemic vice can be forms of direct (albeit partial) control over one's beliefs. I am in basic agreement with Zimmerman's rejection of this premise. Zimmerman believes that "taking care" in exercising open-mindedness is not under the agent's direct control, but more importantly that:

> even if taking such care were in my direct control, *still* we should say that my believing that *p* is not in my direct control. This is because, in such a case, my bringing it about that I believe that *p* must be a non-basic action, since I must first change my attitude from one of being 'closed' to one of being 'open'; and it is only *by way of* doing this that I can come to see the truth. (Zimmerman 2008, 189)

[108] This appears to neglect the possibility of unwitting acts whose moral significance one has only *suspending* ignorance of, rather than disbelieving ignorance; see Chapter Three §3.4.3.

But if forming a (true) belief can only come *by way of* cognitive effort to be open, then the agent's control over it must be indirect, and so Montmarquet wrongly characterises it as direct. It is simply part of the meaning of "indirect control" that it is control over *x* by way of control over *y*.

Thus, it would make most sense for Montmarquet to concede that the cognitive effort to be open is itself a discrete *act* and the failure thereof a discrete omission. But then, he would have the claim that culpability for ignorance can be the upshot of benighting acts or omissions that are epistemically vicious. And since Montmarquet argues that it is not required for voluntary exercises of epistemic vice to be culpable that the agent is *aware* of them as such (see the long quote above), Montmarquet would then count as challenging Premise (3) (that benighting acts are culpable either if they are fully advertent or culpably unwitting), on the grounds that they can be culpable despite the lack of culpable ignorance of *their* epistemic viciousness. But then Montmarquet would be defending something like Fitzpatrick's view. To FitzPatrick, then, let us turn.

While Montmarquet takes on Zimmerman, William FitzPatrick (2008, 2017) takes on Gideon Rosen. His primary target is Premise (3), on aforementioned grounds: sometimes unwitting benighting conduct can be directly culpable due to voluntary exercises of epistemic vice. As a matter of fact, FitzPatrick would also reject Premise (2) on the grounds that the agent's resulting ignorance need not be the *foreseen* upshot of exercises of epistemic vice. However, he does not offer an argument for the rejection of the foresight condition on indirect responsibility, and so I will not reassert my defence of it on the grounds of my account of indirect responsibility (see last Chapter). At best FitzPatrick relies on the intuitions supporting the plausibility of culpable ignorance solely being the upshot of exercises of epistemic vice when it is not the upshot of fully advertent benighting wrongdoing, and so I will confine my attention to discussion of this claim in his response to Premise (3). I should note also that FitzPatrick appears happy to take voluntary exercises of epistemic vice to be *acts* in the ordinary sense. However, he takes them to be acts *from* epistemic vice (from possessed epistemically vicious character traits; FitzPatrick 2008, 608). But surely *viciously-motivated acts* (not from vicious character, but from vicious motives) would satisfy FitzPatrick, even if he follows Aristotle in holding that we are predominantly responsible for our character traits, and so indirectly responsible for actions issuing from them. At any rate, what is FitzPatrick's basic position?

Consider:

> Mr. Potter, a powerful businessman who holds false moral views. He takes certain business practices—such as liquidating Bailey's Building and Loan and sticking it to the poor families of Bedford Falls—to be 'permissibly aggressive,' when in fact they're 'reprehensibly ruthless.' This leads him to do bad things, though he doesn't understand that he's acting badly, which means that he's acting out of a certain kind of ignorance. He's fully aware of all the circumstances, but he applies flawed normative principles or weightings and comes up with bad decisions. Is he culpable for his bad actions? (2008, 599-600)

FitzPatrick believes that Potter is culpable for his bad actions but acknowledges that it is improbable that Potter's moral ignorance would have been the upshot of fully advertent failures to discharge procedural epistemic obligations. Potter, FitzPatrick supposes, has been "raised badly, given skewed values early on, and never taught to reason soundly about such matters" (p. 601). Agreeing with Rosen, FitzPatrick holds that the relevant question for determining the source of Potter's culpability is:

> What, if anything, could the agent reasonably (and hence fairly) have been expected to have done in the past to avoid or to remedy that ignorance? (p. 603)

But FitzPatrick's answer is simple: avoid the voluntary exercise of epistemic vices. Had Potter's ignorance been the upshot of exercises of epistemic vices in a society that provided him the general opportunity to "work on" and avoid these vices, then FitzPatrick argues that Potter would have been reasonably expected to avoid the exercise of those vices, even if he did not exercise them fully advertently. Indeed, Fitzpatrick argues that would be "disingenuous" to claim that it would not have been reasonable to expect these things from Potter—given, he supposes, that Potter would have had knowledge about the general importance of self-scrutiny in other areas, "such as analysis of stock market and interest rate trends" (p. 607). This is so, even if Potter exists largely in the confines of a "narrow, elite social sphere" (p. 607). FitzPatrick's reasoning here echoes comments made by Montmarquet (1999, 845) concerning reasonable expectations, and so both share the claim that it is reasonable to expect agents to avoid benighting exercises of epistemic vice, even if they are unaware of them, the vices themselves, or any resulting ignorance.

FitzPatrick takes these considerations to suggest that following view:

> Ignorance, whether circumstantial or normative, is culpable if the agent could reasonably have been expected to take measures that would have corrected or avoided it, given his or her capabilities and the opportunities provided by the social context, but failed to do so either due to akrasia or due to the culpable, nonakratic exercise of such vices as overconfidence, arrogance, dismissiveness, laziness, dogmatism, incuriosity, self-indulgence, contempt, and so on. (p. 609)[109]

What, then, are we to make of this challenge to volitionism? My first reservation targets any kind of exceptionalism. Why should benighting conduct be treated any differently, as far as culpability ascriptions are concerned, from ordinary (non-benighting) conduct? It is difficult to see *what it is* about being the kind of act or omission that causes ignorance that makes it eligible for a different culpability assessment than any other kind of act or omission. At any rate, this is only a *prima facie* disadvantage for such a view, for we may well be convinced by the epistemic vice theorist's appeal to the distinct nature in which these acts are *epistemically vicious* when grounding culpable ignorance.

At this point, Neil Levy (2009) must be cited in reply to FitzPatrick. Levy contests FitzPatrick's answer to the key question about what it would have been reasonable to expect Potter to have done to avoid ignorance, by arguing that in fact it would have been unreasonable, and thereby unfair, to expect Potter to avoid his voluntary exercises of epistemic vice, because it would have meant expecting Potter to do something that he could not possibly have done *rationally*, given his upbringing and actual beliefs at the time. In effect, Levy invokes Montmarquet's initial worry in parentheses that the epistemic vice theorist violates "ought implies can," where "can" would here mean what one has the *rational capacity* to do. Potter did not, for instance, believe that he should try to see things from another perspective; "by his lights," as Levy puts it, "Potter govern[ed] his normative views adequately" (2009, 737). And so expecting Potter to avoid exercising his epistemic vices would have amounted to expecting him to do something *irrational*. However, "it is not reasonable to expect agents to do anything they can do only by way of a failure of practical rationality" (2009, 739). It is not fair to expect Potter to avoid exercising his epistemic vices

[109] FitzPatrick is one of the thinkers who (wrongly I think; see §4.3.1) takes the volitionist position to entail that direct blameworthiness for advertent wrongdoing requires akrasia.

"by chance or through a glitch in [his] agency" (p. 739). It is fair to expect him to avoid exercising his vices only if he has the *rational capacity* to avoid exercises them.

The sense in which he could not have done it "rationally" is a distinctly *internalist* sense—that is, given Potter's beliefs (credences) or motivating reasons. Sometimes, as Levy (2009, 735f.) points out, rationality can be given a more "external" reading, as when we deem "irrational" people who engage in "substantively senseless behaviour" (Scanlon 1998, 29) but who nevertheless act *consistently* with their (motivating) reasons. But Levy argues that the reasons and rationality that are relevant to establishing what it is reasonable to expect, and thus culpability, are crucially internalist. Blameworthiness depends on reasonable expectations, reasonable expectations depend on rational capacities, and the exercise of rational capacities depends on the agent's beliefs and credences.

One might worry about Levy's claim that even if there is a form of externalist rationality, the only form of rationality relevant to reasonable expectations is crucially internalist. To say so might appear arbitrary, or self-serving. Here, though, I think that something stronger can be said*: there is no form of (non-ideal) rationality that is "externalist*," and so rational capacities by definition are capacities enabled by one's beliefs or credences about what one can do. T. Scanlon (1998, 25-32) sets out particularly persuasive reasons for holding that the term "irrational" should be confined to the evaluation of behaviour that goes *against* one's *own* reasons. After all, this is the *paradigmatic* case of "irrational action." Moreover, most of the time it is the case that substantively senseless behaviour *does* run contrary to one's own reasons because one *does* have outweighing reasons against it—which arguably explains why we call it irrational. But Scanlon also argues that when substantively senseless behaviour *is* consistent with one's reasons, then there are other terms than "irrational" that would do a better job at evaluating this behaviour—for instance, the terms "senseless," "thoughtless," "delusional," perhaps even "unreasonable"—which avoid the ambiguity and the strong internalist connotations of *irrationality*.

Now, it is worth stressing how Levy's reply is also bolstered by my accounts of responsibility and blameworthiness. Potter's exercises of epistemic vice were plausibly not his *poor responses to reasons* against exercising them, because for that to be so, Potter would have had to be aware of those reasons against exercising them (including the reason that it could lead to ignorance about how to conduct oneself properly). But Potter was not thus aware. And so Potter was not responsible. Moreover, upon reflection of whether he could be described as being "morally at fault" for these exercises, if Potter was not originally at fault

for his epistemic vices, and was not aware of his exercising them, then it is difficult to see how he could have been morally at fault for exercising them.

Given that the verdict about Potter's blamelessness is the same, we should ask whether there is any deep relation between being subject to a reasonable or fair expectation to avoid wrongdoing and the wrong action's subsequently counting as a poor response to the (relevant) reasons to avoid it. I believe that there is. To expect someone to act in some way (to morally expect something of someone) is to *believe* that they *ought* morally (all-things-considered)[110] to do it and that they *can* do it. And following the notion of *fair* blame *qua true* judgment (see last Chapter, §5.3), a *fair* moral expectation on someone must then rest on the *true* beliefs that they morally *should* and *can* do the expected action.[111] With this in mind, turn to the fact that it is plausible that wrongdoing's being the agent's (poor) response to the relevant normative reasons entails that the agent could and should have avoided wrongdoing (if only "could have" in the conditional compatibilist sense). Not only have I argued that responsibility *qua* reasons-respondence for wrongdoing entails (minimal) *awareness* of the wrong-making normative reasons, but I have also argued that responsibility *qua* reasons-respondence for wrongdoing requires *control* over that wrongdoing. But if an agent has control over whether they do wrong and has awareness of the wrong-making reasons, then it is surely accurate to say of them that they *can* avoid wrongdoing. Thus, responsibility and blameworthiness for wrongdoing requires that the agent could and should have avoided wrongdoing. With this, we may conclude that responsibility for wrongdoing *requires that it would have been fair to expect* the avoidance of wrongdoing.[112] (See also the link between fair expectations and being "at fault" in: Talbert 2013, 226.)

Returning to the discussion of FitzPatrick, we should note that FitzPatrick responds to Levy by arguing that it would have been reasonable (and thus fair) to expect Potter to avoid exercising his epistemic vices because it would have been reasonable to expect Potter not to have performed vicious "character-forming" actions in the first place (2017, 41-2). But now we should ask whether it would have been reasonable to expect Potter to avoid his vicious character-forming choices in the first place. As Talbert has argued:

---

[110] I will hereafter omit "all-things-considered."

[111] R. Clarke also seems characterise the content of a moral expectation as an expectation that one can and should act in some way (2017, 67).

[112] Notice that the other way around is not necessarily true (that only someone who is responsible for wrongdoing can have been subject to a reasonable expectation) for it may be fair to expect a manipulated agent to avoid wrongdoing even though their failure to avoid wrongdoing would not reflect her true self, which may be required for the act's being the *agent's* response to the relevant reasons.

> [What] if he didn't regard his self-forming choices as bad ones? Do we say that he should have seen this and explain his failure in terms of incipient versions of the vices to which his bad self-forming choices would lead? But now we're back to Levy's worry; namely, the concern that, given Potter's incipient vices, it wasn't reasonable to expect better self-forming choices from him. (Talbert 2017a, 52)

Put simply, the same problem rears its head again. FitzPatrick could, of course, appeal to Montmarquet's case for the *fundamentality* of exercises of epistemic vice (as he does in a footnote in 2008), but why should *character-forming* actions be any different from *benighting* actions in this regard?

Therefore, I think that the epistemic vice theorists fail to provide successful replies to the Regress Argument. Not only is it preferable to treat culpability assessments for benighting and more general acts similarly, but epistemic vice theorists cannot successfully turn the notion of fair expectations against the internalist. Fair expectations go hand-in-hand with culpability internalism (and indeed an account thereof can be grafted neatly into my wider theory of responsibility and blameworthiness).

6.3 Resisting Quality of Will Theories

Another family of views gives rise to a reply to the Regress Argument—namely, "quality of will" views, of which some are culpability externalist. According to quality of will views, blameworthiness requires that a bad quality of will is on display in the action (e.g., malice, disregard, or indifference), and the question about the epistemic condition for blameworthiness is to be answered by inquiring into the epistemic condition for the display of ill will. This is analogous to the way in which some (as we have seen) make inquiry into the epistemic condition for responsibility a function of inquiry into the epistemic condition for *control*. Different quality of will theorists analyse "quality of will" in different ways. For example, blameworthiness has been analysed in terms of the act's expressing or reflecting inadequate care for what's morally significant (Arpaly 2002, 2015; E. Harman 2011, 2015; Littlejohn 2014), indifference towards others' needs or interests (Talbert 2008, 2013, 2017a, 2017b; McKenna 2012), objectionable evaluative judgments (A. Smith 2005, 2008, 2017), a bad moral personality (Hieronymi 2008), or reprehensible desires (H. Smith 1983, 2011; cf. Moody-Adams 1994). We have, of course, encountered quality of will approaches to

responsibility in this thesis (see, e.g., §3.3.3; §4.4.2), and some of those who are sympathetic to a reason-responsive approach to responsibility require ill will for blameworthiness (Arpaly 2002; Talbert 2008; Littlejohn 2014). This is not surprising: responding poorly to moral reasons appears to display disregard for what is morally significant. Finally, but if only to give a feel for the contours of a quality of will view, I should note that there is some ambiguity about whether the "display" relation is evidential or causal (King 2009, 583ff.). There is also disagreement on the issue of whether the quality of will must be a *stable* quality (rather than a "one-off" quality) across time, and on whether it can be an *isolated* quality (e.g., a single desire or evaluative judgment, rather than "the will" itself).

Quality of will theorists reject most, if not all, of the premises of the Regress Argument. While epistemic vice theorists leave Premise (1) intact, quality of will theorists reject Premise (1) on the grounds that one can be directly blameworthy for wrongdoing, even if it is unwitting, when it displays ill will. The ancient slaveholder who beats his slaves, not knowing it is wrong to do so, is blameworthy for doing so, because he displays an objectionable disregard for the humanity of his slaves. For some quality of will theorists (Talbert 2013, 234), this holds *even if* his moral ignorance is blameless (or epistemically justified), given widespread cultural acceptance of slavery or some such excuse.[113] Quality of will theorists also reject Premise (2), because they hold that ignorance can be *directly* blameworthy, when it reveals ill will (e.g., holding prejudiced or misogynistic beliefs about women; Arpaly 2002, 104). Indeed, these theorists typically do not promote tracing explanations (Talbert 2017a, 55-56), because, like real-self attributability theorists, they hold that the relevant responsibility relation between agent and object (act, belief, etc.) is an *atemporal* or *structural* relation between the agent's *quality of will* and the object of responsibility assessment. Finally, for the same reasons that they reject Premise (1), they reject Premise (3): benighting acts can be directly blameworthy if they reveal ill will. Quality of will theorists tend not to focus on these cases (probably because they tend not to require tracing explanations), but it seems to me that they would be open to an explanation of the culpability of benighting acts in terms of the display of epistemic vices (bad "epistemic" qualities of will) (cf. Owens 2000, 124). Otherwise, they might hold that bad *moral* qualities (e.g., moral vices) can be on display in benighting activity (e.g., in cases of "motivated"

[113] By contrast E. Harman (2011, 461-2) requires that one's moral ignorance—or false moral belief—is still *blameworthy*, but she holds that moral beliefs can merit blame simply if they are false, because they violate obligations to believe moral truths.

ignorance).[114] They are therefore in a clear position to reject the conclusion of the Regress Argument.

Now, my concern in this Chapter is to rebut responsibility externalism and to defend my accounts of responsibility and blameworthiness as they have developed thus far—including the newly grafted-in account of fair moral expectations. What makes this tricky to achieve in relation to quality of will theories is that (a) these theories are divided on the culpability internalism/externalism debate (where perhaps the *majority* endorse culpability internalism), but also (b) that some quality of will internalists *already* reject aspects of the theory of that I have put forward in the last two Chapters. Concerning (a), many embrace culpability internalism for the reason that caring inadequately about what is morally significant requires some awareness of what is morally significant (cf. Harman 2011, 460; Talbert 2013, 244; Littlejohn 2014, 144). In order, then, to do justice to these theories, my plan of attack will be to deal first with objections to my reasons-respondence theory, and then to challenge quality of will externalism (epitomised by A. Smith's 2005 work). My case against quality of will externalism will conclude, however, with an argument which targets *all* quality of will accounts, and so quality of will internalists as well.

I turn first to an objection that some quality of will theorists have made to the inseparability of the concept of responsibility and blameworthiness from that of *reasonable (or fair) expectations*. Some quality of will internalists have argued either implicitly or explicitly that responsibility or blameworthiness does not depend on what it is fair to expect the agent to do. We saw above that Matthew Talbert, a quality of will theorist, endorses Levy's objection to FitzPatrick that expectations to act differently are only reasonable if the agent would act fully advertently. Talbert's goal is not, however, to defend volitionism, but to mount a *reductio* of volitionism, on the grounds that a reasonable expectation that the agent could have avoided wrongdoing is not required for blameworthiness. As a matter fact, on Talbert's (2013, 2017a) view, there may be no sense at all in which it would be reasonable or fair to expect Potter, or the ancient slaveholder, to have acted differently—and yet they can still be blameworthy for their heartless business practices or mistreatment of slaves. Why? The reason is that they manifest the kind of ill will to which it would be fitting for their victims to respond with the blaming emotions like resentment. To take the case that Talbert discusses from Rosen in which Bill lies to his wife for self-interested reasons, Talbert insists

[114] E.g., the malicious doctor shows disregard for her patient's life when intentionally omitting to check the patient's blood-type before administering a drug.

that it would still be "reasonable for Bill's wife to blame him because of the way his lying expresses Bill's morally faulty judgments and attitudes" (Talbert 2013, 234), even if it would not have been reasonable to expect Bill to avoid lying to his wife (given his *blamelessly* believing that he should all-things-considered lie to his wife for reasons of self-interest). If she found out about his lies, Talbert insists that:

> it would be appropriate for her to be offended and hurt by what Bill's action expresses, for her to protest that her interests ought to rate more highly with Bill, to resent him for his callousness, to insist that he change his ways, and so on. (p. 234)

In Talbert's view, resentment (expressed in these ways) is sensitive only to displays of ill will, and not to an "inculpating history" (Rosen 2004, 309). Even if the agent is not morally at fault for their ill will, resentment can be appropriate.

We are led, then, to suppose that even if Bill's wife found out that Bill had a very good excuse for his moral ignorance (she found out that he suffered from a virus which upset his normative sensibilities, like Rosen's "Bonnie"), it would still be appropriate for her to resent Bill. I think that this is mistaken. I grant that if she were protected from the information about Bill's virus, her resentment would be *understandable*. But surely, in light of the virus, it would not be *fair*. As FitzPatrick (2017, 33) has put it, Bill would not be "deserving" of blame. I take this, moreover, to be indicated by the fact that it would be natural for Bill's wife to suspend her resentment immediately upon learning about the virus. At any rate, I deny that the relevant question to ask is whether the wrongdoer is worthy of resentment, because I deny that blame must be accompanied by a certain *affect* (see last Chapter §5.3). Blame is judgment of the agent's being *morally at fault* or *in the wrong* for something, but Bill is surely not morally at fault for his lying, given the virus.[115] Thus, Talbert fails to severe the tie between blameworthiness, fair expectations, and being morally at fault.

[115] Note that there may be a sense in which Talbert could say that Bill was "morally at fault" for the lie: the lie was morally faulty, and the lie was traceable to *a* moral fault of his—his "morally faulty judgment" (viz., that self-interested reasons to lie to one's his wife in the circumstances trump moral reasons not to). But note that describing Bill in this way would be to miss a subtlety (from last Chapter §5.3 and §5.5) about judgments of being morally at fault for an act (or at least about the way that I am using this terminology): they need not imply a morally faulty act nor a morally faulty quality of the agent, but rather a moral fault in the *way* that the agent performed the act or the *role* that she played. This is the sense of being morally at fault for wrongdoing that I contend is relevant to responsibility and blameworthiness. But I see no moral fault of this kind in the case of Bill; Bill acted in accordance with his higher-order normative judgment that self-interest trumps morality in the circumstances. How does Bill act in a morally faulty sort of way, relative to that judgment? It might be argued that Bill "let himself" act from that mistaken judgment. But how is that morally faulty, if Bill does not recognise

I turn now to discussing quality of will *externalism*, epitomised by Angela Smith's (2005) work. One of Smith's key claims in this paper is that we can be directly responsible for "spontaneous attitudes, reactions, and patterns of awareness" (2005, 236-7). The nub of her view is that:

> what makes us responsible for our attitudes is… that they are the kinds of states that reflect and are in principle sensitive to our rational [or evaluative] judgments. (p. 271)

The following is her central example.

> I forgot a close friend's birthday last year. A few days after the fact, I realized that this important date had come and gone without my so much as sending a card or giving her a call. I was mortified. What kind of a friend could forget such a thing? Within minutes I was on the phone to her, acknowledging my fault and offering my apologies. (p. 236)

Smith argues that this is a case in which she is straightforwardly responsible for forgetting her friend's birthday, even though she insists that:

> I did not consciously choose to forget this special day or deliberately decide to ignore it. I did not intend to hurt my friend's feelings or even foresee that my conduct would have this effect. I just forgot. It didn't occur to me. I failed to notice. (p. 236)

Especially indicative of her own responsibility, for Smith, is the fact that upon realising that she forgot, she felt the need to *apologise* to her friend. After all, apologies are often markers of responsibility. The way she accounts for this case, then, is by holding that she is responsible for her forgetfulness because of the way that it reveals objectionable evaluative judgments and commitments. Of course, not all "failures to notice" reveal objectionable evaluative judgments (e.g., failing to notice the pattern on that man's shirt would normally not be relevant in this way), but many do, and when they do, we can be appropriately held

---

it as such? (Moreover, in the case of the virus being solely responsible for his beliefs, it is clear that Bill could not have been morally at fault, in this sense, for his ignorance.)

responsible for them because of their sensitivity to these judgments. Since forgetting a friend's birthday involves a failure of awareness of the features which make the omission to send a birthday wish wrong, and yet she is still directly blameworthy for this omission for Smith, Smith defends a form of quality of will externalism.

A number of points should be made in reply to Smith. The first two points are already on the table. To begin with, recall the presumption of responsibility-as-reasons-respondence. Smith is not directly responsible for omitting to send her friend a birthday wish because this omission cannot be said to be her direct response to reasons why she ought not to thus omit: she has forgotten—lost awareness of—the fact that it is her friend's birthday, but we have argued that responses to reasons requires awareness of those reasons, and so her omission to send wishes to her friend is not her direct response to reasons.

Of course, Smith may argue by *modus tollens* that cases of responsibility for "failures to notice" and other involuntary reactions or attitudes show that the reasons-respondence theory cannot be right. But to do so, Smith would have to show that we have very strong theory-independent intuitions about responsibility for failures-to-notice (etc.) *and* that there is not a sense in which they can constitute our responses to reasons. Let me challenge the second claim before challenging the first. There is, after all, the possibility that they constitute *indirect* responses to reasons. It might very well have occurred to Smith, the week before, that her friend's birthday was the following week and that, if she did not set a reminder for herself in her diary, she would likely forget to send her birthday wishes. Smith's failing to remember could then be traceable to her failure to respond to the reasons against forgetting (and so failing to acknowledge) her friend's birthday. Alternatively, Smith might already have standing knowledge that she is prone to forgetting people's birthdays, and ever since this came to light, she might have had plenty of opportunities (due to the latter's occurring to her) to "work on" this forgetfulness. Her forgetting on *this* occasion could then be traceable to failures to "action" her intention to work on her forgetfulness of birthdays (or the general *type* of forgetting special occasions). In the unlikely event that neither of these circumstances obtained, I am inclined to wonder whether Angela really *was* responsible for forgetting her friend's birthday. If, in other words, we cannot trace her apparent culpability back to some prior opportunity to remedy or prevent her forgetfulness on this occasion, I struggle to see how she is responsible for it.

But what of Smith's point that it would be appropriate for Angela to apologise to her friend? Does this not suggest that Angela is responsible? Here, too, Levy has a convincing reply. Quite often we apologise for that for which we *are not* responsible.

> If I step on your foot on a crowded subway, we shall both believe that an apology is called for. That apology might take two forms. The first, in this context extremely uncommon, form consists in the acknowledgement of responsibility for wrongdoing… [But much] more commonly, I might say, 'Sorry, I didn't see you there,' or 'Sorry, I lost my balance' (so much more common are apologies of the second kind than apologies of the first, we can usually just say 'sorry,' confident that our speech-act will be understood as an apology of the second kind). (2005, 12)

It is the second type of case that Levy sees in the case of Smith's forgetfulness, provided that there was nothing that she could reasonably (fairly) have been expected to do to remedy or prevent that forgetfulness. Thus, apologies can be made intelligible without responsibility, and so Angela's apologies to her friend for forgetting her birthday do not imply or even strongly indicate responsibility for forgetting it or omitting to send birthday wishes.

The fact that I struggle to see how Angela could be responsible for her omission in the unlikely case where she could not have foreseen that she might be forgetful, might, of course, be explained by the corruption of my intuitions by commitment to my preferred theory. But my commitment to that theory is itself explained by strong intuitions about other cases, which intuitions Smith herself is willing to concede are "normal" intuitions: "If asked, most of us would probably say that choice or voluntary control is a precondition of legitimate moral assessment" (Smith 2005, 237). Why not think that these intuitions suggest a theory which forces us to explain responsibility for failures to notice by tracing?

As a matter of fact, the absence of an intuition of her culpability in this unlikely case, when suitably fleshed out so as to make tracing impossible, need not be explained by our inability to trace culpability. Thus, it would be wrong-headed to accuse my intuitions of being corrupted by theory. On certain plausible renditions of Angela Smith's case of forgetfulness (call her "Angela" to avoid confusion), other quality of will theorists would also lack intuitions of blameworthiness. On Holly Smith's view, for example, if an agent's involuntary response "reflects a small portion of her psychology, not a sufficient portion of it for us to judge her to be overall morally reprehensible for what she has done" (H. Smith 2011, 145), then the agent cannot be to blame for it. If, indeed, Angela cared deeply about her friend, had never forgotten her birthday, and was particularly busy at the time, then it is likely that Angela's forgetting her friend's birthday on this occasion did not reflect enough of Angela's "moral personality" to warrant blaming her for it.

Now, let me suggest that the fact that we would normally hold voluntary control to be necessary for responsibility suggests that the burden is on A. Smith to justify why we should reject this view on the basis of failure-to-notice cases—which suggests that we can rest content with culpability internalist replies to her arguments.[116] But there is another consideration that might suggest her having the burden of proof. Levy has objected that A. Smith and other so-called "attributionists" (e.g., Scanlon 1998) collapse the "bad" and the "blameworthy" (Levy 2005, 5–6). This is already evidenced by the fact that A. Smith can so easily switch between making a claim about the precondition for "legitimate moral assessment" and a claim about the precondition for legitimate *responsibility* assessments. Indeed, A. Smith (2008) accepts this consequence, speaking in the same breath of moral *criticism* and moral blame. But, as I argued in last Chapter (§5.3), these should not be collapsed. Moral blame for wrongdoing is holding someone morally at fault for wrongdoing, while moral criticism for it is (e.g.) judging someone "shoddy," "selfish," or "indifferent" because of it, which need entail that the agent is morally at fault for it. Levy does well to point out that quality of will theorists struggle to keep the bad apart from the blameworthy.

A. Smith's own response is to deny that collapsing the bad and blameworthy is disadvantage for quality of will or "attributionist" theories, but, given my arguments in favour of the moral faulting view of blame, I think a more plausible response for the quality of will theorist is to admit that this is a problem, but take the solution offered by Holly Smith (2011), who wields the distinction between an act's revealing a *single* objectionable attitude and its revealing a sufficiently comprehensive set of objectionable attitudes. For H. Smith, we can "appropriately think worse of a person" who expresses a single or "isolated" quality of will that is objectionable, but we cannot blame her, unless she reveals "enough of her moral personality" (H. Smith 2011, 144). Consider her key example (2011, 133-4). Clara strongly dislikes Bonnie but has always managed to reign in "nasty" comments about her hair in order to keep a good reputation (among other reasons). One day, however, "Clara's psychology teacher hypnotises Clara," the outcome of which is that Clara no longer cares about her reputation (etc.). In consequence, Clara launches a "cutting attack on Bonnie's appearance." Now, what is important is that the attack manifests a bad quality of will *of hers* (her desire to "wound" Bonnie). But I share H. Smith's intuition that she is not blameworthy. After all, the desires for maintaining her good reputation (etc.) that would normally inhibit her are not

[116] Holly Smith (2011) reflects this emphasis in explicitly attempting to model her own account of responsibility for involuntary responses on an account of responsibility for volitional choices.

operative. Thus, blameworthiness requires the display of a sufficient portion of the agent's will, not just one part of it (e.g., a single bad desire).

This, to me, is persuasive, but H. Smith's solution comes at the high price of accentuating the final, and already quite devastating, problem for quality of will theories. *Displays of ill will are simply not necessary for blameworthiness*. Two classes of cases come to mind, one I will mention here, and the other I will save for the section on capacitarianism below. The cases that I present here are cases in which the agents know what is right, but on account of blameless uncertainty, do not do what is right *in time* to prevent misconduct. Suppose that you find yourself in a variant of Peter Singer's (1993, 229) Drowning Child case, in which there are two children drowning in ponds either side of you as you walk through the park (one child per pond). Only one child can be saved. The only cost for you is getting your nice new suit wet. You are not moved by this, however. Stunned by the dilemma of *which* child you should try to save, alas you do not move quickly enough to save one of them in time before they both drown. Suppose, however, that had you decided immediately, or simply a moment earlier than you did, you would have saved one of them. It seems to me that even if you deserve some compassion following this traumatic incident, you are minimally blameworthy for failing to save one of chilren. After all, you had enough time to choose (if only a few moments), you were able to, and you knew that you had to save one of them and decide fast. Nevertheless, you displayed no ill will. In fact, part of the *reason* why you were so stunned is that you could not bear the thought of one of them having to die, due to an arbitrary decision that you had to make.

It may, of course, be objected that you displayed a bad quality of will in your objectionably weak disposition to act on reasons. But we could flesh out this case such that you actually *had* a strong disposition to react to the right reasons in these kinds of cases, such that you acted out of character. More importantly, without an objectionable disposition to react to the right reasons, the quality of will theorist appears bereft of options to explain your blameworthiness. All of your cares, values, desires, and evaluative judgments looked good. Rather, the problem appeared to be a failure *of reactivity* (to use Fischer and Ravizza's [1998] term)—a failure of action, or choice, *given* your awareness. It was not some failure of an *attitude or state in you* to meet some moral standard for these attitudes. And this problem for the quality of will theorist is, of course, exacerbated if they follow H. Smith (and Hieronymi 2008, 361) in grounding your failure to decide on time in something as big or complex as your "moral personality," for if there was any objectionable attitude that you displayed in this case (which I have argued there was not), the attitude appeared to be isolated or singular.

In summary, and besides the already strong case in favour of a reason-respondence view of responsibility, quality of will externalist views are met with the following objections: they often rule out tracing too swiftly, they collapse the bad and the blameworthy, and they require too much for blameworthiness (especially if they take H. Smith's plausible solution to the problem of the collapse of the bad and the blameworthy). The last objection also has the virtue of directly targeting quality of will *internalism*, along with the objection against Talbert that he has failed to upset the requirement for blameworthiness that it would have been fair to expect the agent to avoid wrongdoing.

6.4 Resisting Capacitarian Theories

The final version of responsibility externalism to discuss is capacitarianism. To see what the view is, however, let us first consider another range of cases that I take to show that ill will is unnecessary for blameworthiness. Capacitarians have used some of the following cases to dismantle quality of will theories.

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. (Sher 2009, 24)

On the Rocks. Julian, a ferry pilot, is nearing the end of a forty-minute trip that he has made hundreds of times before. The only challenge in this segment of the trip is to avoid some submerged rocks that jut out irregularly from the mainland. However, just because the trip is so routine, Julian's thoughts have wandered to the previous evening's pleasant romantic encounter. Too late, he realizes that he no longer has time to maneuver the ferry. (Sher 2009, 24)

Forgotten Milk. [As] I'm about to leave my office at the end of the workday, my wife calls to tell me we're out of milk. My regular route home takes me right by a grocery store, and I tell her I'll stop and buy some. Between my office and the store, I start to think about a paper I'm writing on omissions. I continue thinking about my work until I arrive home, where I realize that I've forgotten the milk. (Clarke 2014, 164)

Secret Service. Alvaro is an important member of the president's Secret Service detail. His job is to keep the president safe at all times. Right now, the president is working a rope line at a rally. There are hundreds of people along the rope line hoping for the chance to shake her hand. Alvaro is next to the president, checking for signs of suspicious behavior. His iWatch vibrates. Stealing a quick glance at it, he sees that he has a text message from his son, who is at a restaurant with some of his… high school friends… Suppose, now, that Alvaro stops for a second to read the text message, and, just at that moment, a person in the rope line lunges at the president. (Nelkin and Rickless 2017, 123–24)

It is plausible to think that in all four cases, the agent does not display any ill will in the act (or omission). And yet we seem to get a clear intuition of blameworthiness in each case. Let me first motivate the point that there is no ill will involved in these cases.

The claim that there is no ill will involved may be questioned, on the grounds that we should treat these cases as A. Smith treats her case of forgetting her friend's birthday: the agent's failure to notice or remember is grounded in an objective evaluative attitude.[117] Alessandra does not care adequately for the wellbeing of her dog. Julian does not value his passengers enough to ensure that he would never crash into the rocks. Randy—the referent of "I" in *Forgotten Milk*—does not care enough about his wife's and his own needs. And Alvaro is not committed enough to the job of protecting the president. These explanations are clearly possible but notice that the quality of will theorist has shouldered a significant burden here, for they must say that there is *no* telling of these stories such that we can preserve blameworthiness without attributing ill will. Plausibly, however, this is false; there is such a telling in each case. Consider H. Smith's plausible defence of Alessandra's will:

> Alessandra, for example, might have a long history of displaying great love and

[117] This possibility is entertained by capacitarians, but also by M. King (2009) and H. Smith (2011).

> concern for the family dog, spending hours playing with her, training her, obtaining expensive veterinary care when the dog is ill, and so forth. Her appalled response on discovering the dog dead from heat prostration further manifests her high value for the dog's welfare, and the improbability that she was even willing to risk exposing the dog to danger. (H. Smith 2011, 119)

We can say something similar about Julian, Randy, and Alvaro. Now, the quality of will theorist could deny that any stable or sufficiently robust qualities of will must be on display (e.g., traits of character), on the grounds that one-off qualities (e.g., desires) may be sufficient for attributing the ill will they think necessary for blameworthiness. For instance, Alvaro is blameworthy because his failure to notice is the product of his fleeting desire to look at the text from his son, in the context where he ought not to. But notice that this would then come at the cost of deeming blameworthy H. Smith's Clara, for her rage at Bonnie, despite her intuitive blamelessness (as argued above). Or at the very least, in order to avoid this consequence of a one-off quality of will theory, the theory would have to posit a condition that no portion of one's psychological make up could be manipulated (e.g., hypnotised), whilst avoiding the charge of *ad hocness*—a prospect that looks bleak given the resources quality of will theorists have to draw upon.[118]

The moral of the story is, I think, that these "failure-to-notice" cases need not involve ill will, and yet they involve blameworthiness, and so quality of will theories get blameworthiness wrong. But capacitarians do something tricky at this point, for they argue that not only do these cases present a problem for quality of will theorists, they also present a problem for attempts, such as mine, to *trace* blameworthiness back to culpable benighting misconduct. In consequence, these cases are often also called "non-tracing" cases (cf. H. Smith 2011). Why think that it is implausible to trace culpability in these cases? Well, for Alessandra, "the pick-up has never taken long", and she is not expecting to be caught up in the tangled tale of misbehaviour (etc.); she is just doing what she always does, and it has never been a problem. Something similar can be said about Julian: he has made the ferry trip hundreds of times; he has not deliberately decided to risk not seeing the rocks—his thoughts have just drifted to his romantic encounter the night before. As George Sher argues about these agents, "the difficulty appears to lie not in the agent's conscious will but in something

[118] Talbert bites the bullet in the face of manipulation cases, holding that manipulation in the agent's history need not affect the agent's responsibility if the agent "acts consistently with her values and desires and for reasons that she counts in favour of so acting" (Talbert 2016, 94).

that overtakes it" (Sher 2009, 25). The other two can be given similar descriptions—especially Alvaro. Alvaro may be a veteran Secret Service guard with reliable and proven instincts about when he can let his attention drift. And Randolph Clarke argues that even if Randy thought that by letting his mind wander he risked not remembering the milk, Randy would not have been under any obligation (or subject to any norm) to remind himself to do so, because "it would have bordered on compulsion to do such a thing" (Randolph Clarke 2014, 165). Notice also that if volitionism is true, it is even less likely that these cases involving tracing, for their indirect blameworthiness would have to trace back to a moment when the agent occurrently believed that it was all-things-considered wrong to take the risk of failing to notice or remember the relevant facts.

We shall return to this claim about these cases, but there is no doubt that they could quite naturally be non-tracing cases. And yet we still get that intuition of blameworthiness. What is the conclusion to draw from this? Capacitarians argue that it must be that moral responsibility is sometimes sensitive to some feature of the agent other than their quality of will or their conscious control. What unites them is the claim that moral responsibility is sometimes sensitive merely to one's *capacity for* awareness (hence the name "capacitarianism"). More accurately, capacitarianism is the view that direct responsibility or blameworthiness for an act (or omission) requires either awareness *or the capacity for awareness* of the wrong-making features of (or reasons against) the act (Sher 2009; Randolph Clarke 2014, 2017; Rudy-Hiller 2017; Murray 2017; Murray and Vargas 2020). The view is disjunctive, because capacitarians allow blameworthiness for acting contrary to one's awareness. However, the key point of contrast is the claim that having the mere unexercised capacity for awareness of the wrong-making features can be enough for direct blameworthiness. Capacitarians usually require the satisfaction of other conditions related to the exercise of the capacity. Rudy-Hiller, himself a capacitarian, describes the view as holding that when the agent is not aware of the relevant considerations, "they *should and could be aware* of them given the available evidence, the opportunity to adequately process it, and their cognitive capacities" (Rudy-Hiller 2018). Thus, there must not only be unexercised capacities for awareness, but it must be that the agent "should have" the relevant awareness, and that they have a (fair) opportunity to exercise their capacity for awareness.

With these conceptual ingredients, capacitarians argue that they are able to account for our intuitions of the agents' blameworthiness in the above cases (of Alessandra, etc.), unlike quality of will theorists and internalist tracers. I will indicate how they do so, in the context of spelling out these ingredients—namely, the notion of capacities for awareness, the idea that

the agents need a fair opportunity to exercise these capacities, and the idea that the agent *should be* thus aware.

Capacitarians generally agree on which kinds of cognitive processes or faculties constitute cognitive capacities, however they disagree on how exactly to characterise them. On the former, Clarke has a useful passage cataloguing the relevant capacities:

> Some are capacities to do things that are in a plain sense active: to turn one's attention to, or maintain attention on, some matter; to raise a question in one's mind or pursue such a question; to make a decision about whether to do this or that. These are, in fact, abilities to act. Others, though capacities to do things, aren't capacities whose exercise consists in intentional action. These include capacities to remember, to think of relevant considerations, to notice features of one's situation and appreciate their normative significance, to think at appropriate times to do things that need doing. (Clarke 2017, 68)

Most capacitarians allow both kinds of capacities,[119] however some (e.g., Sher) do not allow the first class of capacities that issue in intentional actions, for, as Sher argues, "if we did construe the cognitive capacities as ones that their possessors can choose to exercise, then we would have ushered [an internalist tracing view] out the front door only to see it reenter through the back" (2009, 114). It is not clear, however, that allowing these more volitional capacities would involve smuggling such a view back in, for capacitarians need not hold that as soon as we enter any domain of agency or choice, let alone the domain of exercising cognitive capacities, awareness conditions need to be met.[120] What matters is that there are capacities left unexercised (or inactivated), despite a fair opportunity to exercise them (or for them to operate), and norms calling upon the agent to do so. At any rate, there is complete agreement on the relevance of the second class of capacities mentioned by Clarke, those whose exercise do not consist in intentional action.

[119] Murray is an exception, however: he would not pluralise capacity. Murray thinks that all responsibility bottoms out in either awareness or the (unexercised) capacity for *vigilance*, a general and "variegated" disposition to "become occurrently aware of morally or prudentially relevant considerations that constitute a sufficient reason to act or omit" (Murray 2017, 513). This disposition also "requires more or less effort to appropriately exercise" (p. 516).

[120] At least, R. Clarke and Rudy-Hiller would be inclined to say as much, given their desire to ground their capacitarianism in a non-volitional view of control (which I'll discuss further below). George Sher, by contrast, wants to sever the necessary link between control and responsibility.

Capacitarians face the challenge of answering what it takes to have the relevant capacity. Clarke and Rudy-Hiller take a view on which the agent has the relevant capacity if *on similar occasions in the past*, they have become aware of the relevant reasons. By contrast, Sher adopts a *counterfactual* analysis of capacities:

> To say that someone was capable of remembering something he forgot is only to say that he would have remembered it in an appropriate range of alternative situations. (2009, 114)[121]

Of course, both accounts have problems,[122] but I do not expect these to be insurmountable for the capacitarian (and so they will not figure prominently in my evaluation below).

So far, this appeal to capacities looks promising for explaining the above cases. Alessandra surely has a capacity to keep track of (remember, think of) the things that depend on her—for example, her children and pets. Julian is normally vigilant enough to pilot the ferry. Randy is quite capable of remembering the milk at the relevant time (Clarke 2014, 165). Alvaro clearly has the requisite attentiveness to be hired in the president's security detail. Moreover, these agents appear to be "normal," morally competent agents. In this connection, Rudy-Hiller is right, I think, to appeal to a "commonsensical view of unexercised capacities and fair opportunity," a view on which we should assume, by default, that "seemingly normal adults do have those abilities even when they fail to exercise them, and do have such opportunity even when they fail to take advantage of it" (Rudy-Hiller 2017, 407).

For capacitarians, however, having the capacity for awareness means nothing without a fair opportunity for it to manifest. Rudy-Hiller, for instance, requires that there are no "situational factors that decisively interfere with the deployment of the relevant abilities" (2017, 408). It could, for example, be argued that Alessandra's being caught in tangled tale of misbehaviour made it too difficult for her to exercise her memory of the dog (although this would not yet absolve Alessandra if she initially had the capacity to prevent the risk to her dog, by, e.g., leaving the windows open, before leaving the car). Clarke says something similar, although he appears to argue that the situational factors need not *decisively* interfere with the exercise of these capacities for them to be exculpatory; it is enough that they "sometimes mask… the manifestation of psychological capacities without diminishing or eliminating them" (Clarke 2017, 68). In this connection, he argues that if an agent in Randy's circumstances, "had just

[121] Murray (2017) likely also takes this view, given the standard counterfactual analysis of dispositions.
[122] See Rudy-Hiller's (2018) illuminating discussion of the problems for both views.

witnessed a horrible accident, it might not be reasonable to expect her to remember or think to do certain things that she has a capacity to remember or think to do" (2017, 68).

The last key requirement, according to the capacitarian, is that the agent *should* have been aware of the relevant considerations at the time of acting.[123] The above two notions may give the capacitarian the claim that the agent *could have* been aware, but they do not yet amount to the condition that the agent *should* have met a certain "cognitive standard" (Clarke 2014). Why is such a condition indispensable? Well, just as internalist tracers require that blameworthiness for an unwitting act requires benighting *misconduct*, so capacitarians require that blameworthiness requires that the agent's awareness fell below a certain standard that they could reasonably have been expected to meet in the circumstances. If, for example, Randy had been stunned by just witnessing a car crash, it seems false that Randy *ought* to have remembered his promise to get milk, for such a standard would seem too harsh or demanding. Capacitarians disagree, however, on whether this standard is set by an *obligation* to be aware (Rudy-Hiller 2017, 415; Murray 2017, 513), or merely a *norm* to be aware (Clarke 2014, 167). Either way, though, the standard applies when the agent *can* have the requisite awareness (due to having the capacities to be aware and the fair opportunity to exercise them), but also when this awareness is needed in order to fulfil their obligation *to act* in a certain way (e.g., to rescue the dog, or to avoid the rocks) (Sher 2009, 111–12).

All the capacitarian ingredients are now on the table. Before we turn to an evaluation of the view, however, we should ask how it interacts with the Regress Argument. Some capacitarians accept Premise (1) (that culpability for an act requires that the act is either fully advertent or culpably unwitting); others reject it. Rudy-Hiller (2017, 417) accepts this claim, but denies that blameworthiness for a culpably unwitting act counts only as *derivative* blameworthiness. Thus, if Premise (1) was formulated to imply that one is culpable for an unwitting act *because* one is culpable for the ignorance, then Rudy-Hiller would reject it.[124] By contrast, Clarke (2004, 173-4) rejects Premise (1) even as I have formulated it, because culpability for an unwitting act requires not culpable ignorance but *faulty* ignorance, entailing that one could have and should have been relevantly aware, but not necessarily that one's ignorance was the upshot of culpable misconduct (which he takes to be necessary for its culpability). This disagreement applies also to Premise (3) in the Regress Argument, for

[123] For this reason, Nelkin and Rickless (2017) call capacitarians "Below-Standardists."
[124] This resembles the quality of will theorist, E. Harman (2011, 459), who holds that culpability for the act *requires* culpability for the ignorance, but denies that culpability for the act is *derivative* from culpability for the ignorance, because both are derivative from a manifestation of ill will.

capacitarians treat benighting acts and non-benighting acts in the same way (with one possible exception).[125] I should note too that Rudy-Hiller rejects Premise (2) (that culpable ignorance must be the foreseen upshot of culpable benighting misconduct). Rudy-Hiller thinks that the agent can be *directly* blameworthy for their ignorance, if the agent had "capacitarian control" over it. By contrast Clarke appears to accept Premise (2), although he is silent on the requirement of *foresight*. Whichever way they go (rejecting Premises (1) and (3), or rejecting Premise (2)), capacitarians thus also reject the Conclusion of the Regress Argument. Culpable conduct need not bottom out in fully advertent conduct; it can sometimes bottom out in faultily unwitting conduct (Clarke), or else in directly culpable ignorance, due to the agent's having capacitarian control over it (Rudy-Hiller).

Capacitarianism is intuitive. Not only does it come along with the advantage of avoiding revisionism about responsibility, but it seems to account for intuitions generated by cases that are unable to be accommodated by volitionists, other internalist tracers, and quality of will theorists. The view should also be commended for its apparent ability to be shaped by considerations of reasonable expectations (cf. Clarke 2017, 74; Murray 2017, 514) or being at fault (Clarke 2014, 170).[126] The ferry crash is surely "Julian's fault"; similarly, it would have been fair to *expect* Alvaro to prevent an attack on the president. As such, capacitarianism constitutes an impressive challenge to the link between fair expectations and *rational* capacities already established (in §6.2), as well as the claim that being morally at fault for conduct requires awareness of its wrong-making features (last Chapter §5.5). Sometimes one is not aware that they ought not to act in some way, and yet it seems reasonable to expect them to avoid doing so, and to hold them at fault for it anyway.

Indeed, capacitarianism is the toughest of the three forms of externalism to tackle, in part because its analysis of some of the cases used in support of it is, I think, consistent with affirming a reasons-respondence view of responsibility. But how can this be, if the reasons-respondence view of responsibility entails culpability internalism (as we argued last Chapter)? Let me explain. Some of the alleged non-tracing cases to which capacitarians appeal in order to justify their account of the epistemic condition for direct responsibility are cases to which some culpability internalists can appeal to justify *their* account of the epistemic condition. Recall that culpability internalism is the view that blameworthiness for

[125] Rudy-Hiller (2018) takes FitzPatrick (2008) to be a capacitarian as well as an epistemic vice theorist.
[126] Typically, though, capacitarians use the terminology of "moral fault" to describe morally substandard ignorance (cf. Clarke 2014, 170; Rudy-Hiller 2017, 416ff.). Having a morally faulty *x* should be distinguished from being morally at fault for *x*.

some act requires a belief or credence about the moral significance of the act, where this entails a belief or credence in either the act's having a moral status, or in its having features which in fact make it have its moral status. Now consider both Alessandra and Randy. While caught up in the tale of misbehaviour at the school, Alessandra plausibly still *believes* that the dog is stuck in the car on a boiling day, and that she has been away from the car longer than usual—such that she would recognise these as (normative) reasons to head back to the car immediately if asked whether there was any reason to return to the car. She has *dispositional* beliefs in features which make her delay in returning to the car morally objectionable. But this means that ascribing direct responsibility to Alessandra *is* consistent with affirming culpability internalism (of this bare-bones variety), because having a dispositional belief in the act's having a morally significant feature suffices for having a belief concerning the act's moral significance.[127] Or consider the case of Randy. As Randy comes up to the store where he can stop for milk, Randy surely still *believes* that he has promised his wife to get milk on the way home and therefore that he *should* stop for milk, morally. It is just that these are *dispositional* beliefs, because he is distracted by his philosophical thoughts. But a belief is still a belief, and so a story on which Randy is directly blameworthy for failing to stop for milk is still consistent with culpability internalism (of this bare-bones variety). Now what matters for our purposes is whether Alessandra's or Randy's omissions count *as responses to reasons* (why they are committing wrongdoing), even though they are not conscious of these reasons to do otherwise (to rescue the dog, or stop at the store). I think, in the end, that these omissions *can* be construed as responses to reasons, though only indirectly. However, I shall save my defence of this claim until the next Chapter, where I discuss forms of culpability internalism—especially forms of internalism which allow for blameworthiness despite the presence of only *dispositional* beliefs in wrongdoing (or wrong-making features). Suffice to say, for my purposes here, that a reasons-respondence internalist view of responsibility *may* (as I have shown) entail that the cases of Alessandra and Randy involve direct responsibility and blameworthiness.

The cases of Julian and Alvaro are different, however, for capacitarians argue that the agents are directly blameworthy, even though they completely lack beliefs or credences in reasons why failing to avoid the rocks (this time), or failing to prevent *this* attack on the

[127] Of course, the fact that the dog is stuck in the car on a boiling day could have been *completely* forgotten by Alessandra—that is no longer held as a belief in her head—in which case a story on which she has direct blameworthiness would be inconsistent with responsibility internalism. But this would only constitute an objection to my point if it were the case that forgetting facts always meant losing all awareness of those facts, even dispositional awareness, which seems false.

president, is bad. Of course, they have beliefs and credences in reasons why these general omission-*types* would be bad (hypothetically), but they have no beliefs and credences in the reasons why *these (token)* omissions are bad, because they have no awareness of performing these omissions in the first place. But since they perform these omissions while having (had) the capacity and fair opportunity to be aware of the reasons against them, and when they ought to be aware of them, capacitarians argue that the agent is directly responsible (and blameworthy) for their omissions. The reasons-respondence view entails, however, that they cannot be directly blameworthy for them, because I argued last Chapter that a response to reasons requires awareness of the reasons to which one is respondent. Does this therefore give us a reason to reject the reasons-respondence view?

I do not think so, for although there is an intuition of Julian's and Alvaro's blameworthiness, it is not at all clear that there is an intuition of *direct* blameworthiness. In my view, these cases can be described as *indirect responses to reasons* and thereby given an internalist tracing explanation. The reason why capacitarians report intuitions of direct blameworthiness is, I think, that (to use H. Smith's language) "the temporal gap" between the benighting act and the unwitting omission is "infinitesimal" (H. Smith 1983, 547). Still, the gap necessitates an appeal to indirect responsibility. Julian, to start with, surely believes that he ought to keep his attention on piloting the ferry throughout this trip (as on any other trip), and for the reasons that make this obvious. He surely also regards (or at least *would* regard) this as giving him a reason not to let his mind wander to his romantic encounter the night before. I think, then, that a case could be made for Julian's allowing his mind to wander being his *direct* response to the reasons he has not to do so, even though he was not *occurrently* aware of those reasons; and therefore for holding that his eventual omission to avoid the rocks was his *indirect* response to reasons (not to perform that omission), because the reasons not to let his mind wander included the reason that doing so would risk performing putting passengers' lives at risk by crashing the ferry into the rocks. I put things in this way, of course, to match the brief account of indirect responsibility and blameworthiness that I gave last Chapter, requiring foresight of the eventual act (or omission) and the reasons against it. Now exactly the same kind of story could be given, I think, to the case of Alvaro in order to account for his blameworthiness for not preventing the attack on the president: this omission could be an indirect response to the reasons not be in a position where he is unable to prevent an attack on the president. (I will return to these cases in Chapter Eight.)

The upshot of our discussion of capacitarianism is so far that it is probably false that only capacitarians can account for the agents' blameworthiness in the four cases above. This is less of a problem for, than a defeated advantage of, capacitarianism. But there is a problem that I think is particularly poignant for capacitarianism. And there are some other worries arising from the *complexity* of such a view (e.g., how to account for the relevant capacities, and for when the agent has a *fair*, as opposed to an *unfair*, opportunity to exercise their capacities), but these other worries seem less serious. If anything, all these would show is that if there is a view of the epistemic condition that is *simpler* and achieves just what the capacitarian wants to achieve (which I think in the end that there is), the simpler view should be preferred.

The problem that I think is poignant for capacitarianism is that it seems to produce "false positives" or verdicts that the agents are responsible or blameworthy for their behaviour when they plausibly are not. Consider an interesting example from the command responsibility literature in international criminal law. In a 2017 paper, legal scholar Darryl Robinson criticises the "had reason to know" (HRTK) test sometimes employed by international tribunals to assess the (legal) responsibility of commanders for their subordinates' war crimes Robinson points out that the HRTK test for commander responsibility would be satisfied if the commander had "possession" of reports that had made it to his office with details about these war crimes. But according to the way that the Chambers have understood the term, Robinson points out that possession means only that the information "needs to 'have been provided or available' to the commander" (2017, 644). "[The] commander need not have '*actually acquainted* himself' with the information" (pp. 643-4). As a matter of fact, it would be sufficient for the reports to have "made it to his desk, even if exigent demands of his work understandably delayed him from reading the reports" (p. 645). Robinson, however, criticises this test as being "over-inclusive," for surely the commander who, on account of these extenuating circumstances, did not get to read the reports in time for effective remedial action, is not negligent here, and so not legally culpable.

Now, what has this to do with capacitarianism about the epistemic condition for *moral* responsibility? My intention is clearly not to make a contribution to the command responsibility literature, but I think that the case, or at least a variant thereof, is a useful one for assessing what capacitarians should say about the commander's *moral* responsibility (regardless of what test we want to employ to assess his *legal* responsibility), and accordingly, about the responsibility of agents in similar types of cases. Suppose that we have a commander whose subordinates have committed a terrible atrocity (the unjustified killing of

civilians) without knowledge of it. A report indicating these crimes (presumably not from the perpetrators themselves) makes it to his desk and ends up at the bottom of a pile of other reports. Now suppose that the "exigent demands of his work" (including the reading of other reports) delay him from reading the relevant report, but that he could *easily* have picked it up to read about the incriminating information in time for effective remedial action (e.g., because he has recently developed the habit of flipping the pile of reports to work on the bottom ones first). The commander seems to satisfy the capacitarian's conditions. He has the capacity to acquire the information (e.g., he can read the reports; he has recently flipped the pile of reports), he seems to have had a *fair opportunity* to acquire it (a number of days to possibly pick up this report), and one of his work obligations is to know what his subordinates are up to. Suppose, however, that he does not read the report in time for effective remedial action against what happened.

It would seem that the capacitarian should say that he is responsible (and blameworthy) for this omission. But my intuitions say that he is not. After all, the information has not made it to the commander. And even though the workload still allows it to be the case that he could easily—in fact, let us say that it was *probable* that he was going to—read the report in time for effective remedial action, his failure to have read it does not seem traceable to any blameworthy conduct. He has carried out his job as usual, to a decent, standard. It just seems *unlucky* that he has failed to read the report—and unlucky, of the kind that threatens a judgment of responsibility for his subordinates' war crimes. Robinson seems to have this intuition guiding his judgments about the commander's legal responsibility: the "test hinges too dramatically on whether *other actors or external events* bring the alarming information into the nebulously-defined 'possession' of the passive commander" (p. 646). Responsibility is made too external, and in consequence, too subject to luck.[128] I think that the same observations could be made about other cases like it, involving the clear lack of awareness of relevant facts, even though it would be easy to acquire this awareness, and one *should* acquire it. Think, for example, of a case in which (a) a major incident appears in your peripheral vision—a high-profile burglar breaks into building in the distance—such that you would usually and could easily spot it, and (b) you are there to spot incidents like that (e.g., as a security guard), but (c) you are focused unluckily on something unrelated in your foreground that seems suspicious (an escalating disagreement between two shadowy men). It would be

[128] There are some luck-based objections like this to capacitarianism in the literature. Levy, for example, argues that whether or not the relevant capacities for awareness are exercised is a "chance occurrence… over which she cannot exercise (and does not possess) control" (2017, 255).

easy for you to spot the break-in and it is your job to spot breaches of security but since your attention seems justifiably to be focused on something else, it would seem unfair for someone (e.g., your boss) to blame you for not responding to the break-in. The information has not properly "entered in" to your head. And yet capacitarianism seems, wrongly, to entail that you are blameworthy for this omission (all things being equal): you had the capacity and fair opportunity to be aware of something of which you ought to be aware.

I conclude my case against capacitarianism. Overall, it is unlikely that only capacitarians can account for the cases raised in defence of their view; and the view involves too much luck, evidenced in mistaken verdicts of blameworthiness.

6.5 Conclusion

I have argued in this Chapter that three forms of culpability externalism about the epistemic condition cannot shake my case for culpability internalism and fail as replies to the Regress Argument. Apart from the ways in which my accounts of responsibility and blameworthiness put pressure on these views from the start, the three forms of culpability externalism have special problems of their own. Epistemic vice theorists introduce an odd exceptionalism in culpability assessments for benighting conduct and cannot turn the notion of fair expectations to avoid wrongdoing against the culpability internalist (indeed a notion that fits nicely with my accounts of responsibility and blameworthiness). Quality of will accounts of the epistemic condition were considered, and then found that they variously: fail to dismantle the claim that fair expectations are necessary for blameworthiness, rule out tracing too swiftly, collapse the bad and the blameworthy, and require too much for blameworthiness (in some failure-to-notice cases and in cases where one fails to act quickly enough). Finally, I considered the more plausible capacitarian theories, which I nevertheless found wanting, in light of the fact that they (also) rule out tracing too swiftly and produce mistaken verdicts of blameworthiness (premised on too "external" a conception of accessibility). The moral of the story is that we should reject culpability externalism, in favour of culpability internalism. In the next Chapter, I will begin an investigation into what form of culpability internalism we should embrace, and whether that should be a kind of volitionism after all.

Chapter 7
The Epistemic Condition for Direct Blameworthiness

7.1 Introduction

Having set out a reasons-respondence theory of the concept of responsibility in Chapter Five, in Chapter Six, I then defended this theory and its entailment of culpability internalism against culpability externalists of the "epistemic vice," "quality of will," and "capacitarian" varieties. The job of defending culpability internalism is now done; we can answer "yes" to the first part of the thesis' guiding question of whether blameworthiness for conduct depends on the agent's beliefs or credences about the wrongfulness of her conduct. But we have not yet answered the question of *the way in which* blameworthiness so depends—namely, the question of what epistemic states *in particular* are the relevant epistemic states on which blameworthiness depends. Recall that, for volitionists, blameworthiness for wrongdoing is, or is traceable to, *fully advertent* wrongdoing—wrongdoing while in the true occurrent belief that one's conduct is all-things-considered wrong, based on (occurrent) beliefs in the normative reasons why it is wrong. Should we accept this answer?

In the next two Chapters, I will argue that we should not accept this answer but rather adopt the strategy of "weakening" the volitionist's epistemic condition. To take this approach is almost surely to adopt a version of what F. Rudy-Hiller (2018) calls "weakened internalism" about the epistemic condition.[129] It is to hold that some kind of *partial awareness* of wrongdoing can satisfy the epistemic condition on blameworthiness. There are at least four different versions of weakened internalism: P. Robichaud's (2014) non-decisive reasons view, C. Sartorio's (2017) awareness of *de dicto* moral significance view, A. Guerrero's (2007) moral risk view, and the dispositional belief-in-wrongdoing view defended

[129] We have already seen how some internalists are quality of will theorists. I will not consider quality of will internalists in this Chapter, on account of the objections that I raised against *any* quality of will view last Chapter (e.g., that ill will is unnecessary for blameworthiness; §6.4-6.5).

by, among others, I. Haji (1997), R. Peels (2011), D. Husak (2011), and K. Timpe (2011). Although I will argue that these views all have something importantly right to them (and not just in their presumption of culpability internalism), it is my object in the next two Chapters to argue for a novel account. In this Chapter, I focus on the epistemic condition for *direct* blameworthiness, and in the next Chapter I focus on the epistemic condition for *indirect* blameworthiness. The account that I defend in this Chapter is the following: an agent satisfies the epistemic condition on direct blameworthiness for an act if and only if, at the time of the act, the agent has right and outweighing motivating reasons to refrain that are either explicit (i.e., occurrent) or consciously accessible through (what I call) "deliberative attunement."

My method of argument will be much as it was in the last Chapter—to make use of the theories already on the table, as well as to pinpoint specific problems for each of the aforementioned alternative views, in order to argue for the superiority of my own. For a reminder of the theories on the table, I cite the reasons-respondence theory of responsibility (defended in Chapter Five), the moral faulting view of blame (defended in Chapter Five), and a corresponding notion of fair moral expectations (defended in Chapter Six). Factoring in my account of indirect blameworthiness next Chapter, the final payoff is, of course, that the view critically sidesteps the volitionist's revisionism about responsibility. Precisely how this is so will be spelled out in Chapter Nine (the Conclusion).

The structure of the Chapter is as follows. In §7.2, I consider Robichaud's view and argue that although Robichaud is right to appeal to motivating reasons and "non-decisive" motivating reasons to refrain, his alternative is too lenient in that it allows for motivating reasons to do otherwise that are irrelevant to the wrong-making features of the act. Given that non-decisive motivating reasons include reasons other than beliefs in wrongdoing, this Section also constitutes a denial of the dispositional belief-in-wrongdoing view. In §7.3, I then discuss Sartorio's account which falls victim to the objection that it narrows down the relevant range of reasons too far (to those that are about the *de dicto* moral significance of the act), and that it mistakenly dispenses with the requirement that there be apparent alternatives. Guerrero's view is then discussed in §7.4 as a candidate for striking a balance between Robichaud's view and Sartorio's view, but in the end faulted for its implication that there can be blameworthiness in "epistemically irresolvable" moral dilemmas. §7.5 is then devoted to the issue of whether the relevant motivating reasons can be *implicit*, where I side with dispositional belief-in-wrongdoing theorists in holding that the relevant reasons can be implicit, but argue contrary to these theorists that there must be *something* occurrent in the agent—which I call a state of "deliberative attunement"—at the time of acting.

7.2 Non-Decisive & Wrong-Related Reasons: A Response to Robichaud

A good place to start is with the debate between Robichaud and Levy on the particular question of the epistemic conditions under which it would be fair to expect someone to avoid wrongdoing. Recall Levy's argument (from last Chapter §6.2) that "it is not reasonable to expect agents to do anything they can do only by way of a failure of practical rationality" (Levy 2009, 739), with the culpability internalist conclusion that blameworthiness must require that the act is morally criticisable by the agent's lights (morally criticisable relative to the agent's beliefs and credences). Remember though that Levy took these considerations to support volitionism. For us to have the rational capacity to avoid wrongdoing, we must believe that we *ought not* (morally, all-things-considered) to perform the act, and believe that occurrently (Levy 2016, 265). Anything else undermines our *rational* capacity to avoid wrongdoing, such that we could only avoid it "by chance or through a glitch in [our] agency " (Levy 2009, 739). In summary, it is reasonable to expect avoidance of wrongdoing only if we have the rational capacity to do so; and we have the rational capacity to do so only if we have the occurrent belief that we morally should not perform the (wrongful) act. The first claim is plausible, for reasons given last Chapter. But what about the second?

Proponents of the dispositional belief-in-wrongdoing view would contest Levy's condition that the belief in wrongdoing must be *occurrent* in favour of its needing to be only dispositional or implicit, but a more radical response rejects the requirement of belief in wrongdoing altogether, targeting both volitionism and the dispositional belief-in-wrongdoing view. Robichaud (2014) is the key antagonist here. But before I outline his argument, let us recall our earlier account of a concept that is central to his (and my later) account, the concept of "motivating reasons":

M: An agent has a *motivating reason* for or against an act at some time *t*, *if and only if*, at *t*, (i) they believe or have a non-negligible credence in *p*, (ii) they would, in principle, express their belief or credence in *p* in answer to the question of why they should or should not act (or an equivalent question), and (iii) this belief/credence is capable of featuring in the explanation of why they act (or omit).

Remember, importantly, that they are different from the reasons "out there" that *in fact* count in favour of acting or refraining from acting—the normative reasons, to which we must be respondent to be responsible. Motivating reasons can align, but can also fail to align, with

one's normative reasons. They are also not necessarily the reasons that feature in the explanation of the agent's conduct, however they must be capable of doing so (just as beliefs/credences are capable of doing so).

With that reminder, let us turn to Robichaud's argument against Levy. On Robichaud's understanding, Levy (and the dispositional belief theorist)[130] requires for blameworthiness that we have a "decisive" motivating reason against some act, where we have a decisive motivating reason against φ-ing if we take ourselves to have a reason that is "strong enough as to decisively support [not-φ-ing]" (Robichaud 2014, 142). Levy requires a decisive motivating reason against wrongdoing because believing that one morally should not φ gives one such a reason, according to Robichaud. However, Robichaud argues that we need only have (what he calls) a "sufficient, non-decisive" motivating reason against φ-ing to have the rational capacity to avoid φ-ing, a reason against φ-ing that is *strong enough as to make not-φ-ing rational* but not strong enough as to decisively support not-φ-ing. To use his example, although we do not believe that we have an *obligation* (or that we morally ought) to check the functionality of our brake lights every time we go to drive, we may believe that "it would be good" to check them and may have the rational capacity to check them solely for that reason. "It would be good" or alternatively "it would be safe," or "I haven't checked them in a while" (my examples), would then function as *non*-decisive motivating reasons to check them and not to ignore them, in contrast to decisive motivating reasons such as "it would be wrong not to," "I overall ought to," or "I have an obligation to" check them.[131] If *credences* can also constitute motivating reasons, then they would plausibly also constitute non-decisive reasons (e.g., a credence of 0.5 in the brake's being in good condition). Now, in the circumstance of having only *non-decisive* reasons to check the brake lights, it is surely true that checking the brake lights could be *rational* for us (absent outweighing reasons to do otherwise). But then it is false that we have the rational capacity to check the brake lights only if we have a *decisive* reason to check them.

The claim that non-decisive motivating reasons not to φ give us the rational capacity not to φ is surely plausible. The belief that checking the brake lights would be good seems like it would give a rationale for checking them and the ability to check them. However, Levy is not convinced. Levy (2016) responds that acting for non-decisive reasons is too "chancy" to

[130] I will hereafter until §7.5 omit reference to defenders of this view

[131] Since others use the term "sufficient reasons" to describe *decisive* reasons, I will drop "sufficient" and simply call them *non-decisive* reasons instead. They are reasons which are *adequate* reasons on which to act, but not (in the usual logical sense) sufficient.

count as making the act one which you have the rational capacity to perform. This is because when someone has non-decisive reasons to act one way, "the agent has what she takes to be equally good reasons to choose either option" (2016, 4), and when she has equally good reasons to choose either option:

> the effect of any of a wide number and variety of situational and internal primes and influences may be decisive [in causing one to act]... It follows that in a large proportion of nearby possible worlds, the agent will choose differently; thus her actual choice is chancy. (pp. 4-5)[132]

Crucially, when her choice is "chancy," her choice is no longer under her control or the exercise of her rational capacity to avoid wrongdoing, and thereby undermines any legitimate ascriptions of blameworthiness.

I have a few responses to this. To begin with, I agree (albeit for different reasons; see below §7.5) that blameworthiness is undermined when the agent has "equally good reasons to choose either option", but I do not think that their reasons need to be equally good to be non-decisive. Neither, apparently, does Levy: later in the paper he recognises the possibility of non-decisive reasons being stronger or weaker (and so plausibly stronger or weaker relative to other non-decisive reasons), but says:

> when it is genuinely the case that an agent has sufficient but not decisive reasons to choose from two or more conflicting options, chancy factors will play a decisive role in how she chooses. (p. 5)

Thus, even if someone's non-decisive reason to avoid wrongdoing is stronger than their reason to do so but still not strong enough as to be decisive, Levy nevertheless rules out blameworthiness for wrongdoing on the grounds of luck ("chancy factors"). To me, however, whether the agent's reasons are equally strong or imbalanced is critical for blameworthiness. Here is my intuition of blameworthiness (and responsibility) for wrongdoing when the agent has outweighing reasons: if the Battalion 101 officer began having doubts about whether he would be violating human rights by shooting upon innocent Jewish women and children, and

[132] To avoid misunderstanding, notice that the sense of "decisive" here means "sufficient to cause someone to act," not the sense of decisive in "sufficient, according to the agent, to make the act what one ought/ought not to do."

those doubts wound up giving the officer a non-decisive but outweighing motivating reason, inclining him, say, 55% in favour of not taking part in the shootings, then it seems pretty clear to me that he could be to blame for participating in the shootings—morally at fault, and poorly respondent to the reasons not to. Below I will give some reasons why *equally* strong reasons for and against might, by contrast, *excuse* wrongdoing. Still, I am agreeing with Robichaud that non-decisive reasons to avoid wrongdoing can be epistemic grounds for blameworthiness. A thorough defence of the possibility of blameworthiness for wrongdoing despite having only non-decisive reasons against it, would require some discussion of Levy's argument that there is too much responsibility-undermining luck in these cases. I do not have the space to engage in this discussion, except perhaps to mention that it is by no means obvious that Levy's metaphysics of control needs to be accepted. On leeway incompatibilist accounts of control (of which I am fond, for independent reasons), cases in which one is torn between conflicting motivating reasons to do different things are paradigm cases of responsibility-relevant control. This point is made, for example, in the work of R. Kane (2007), R. Swinburne (2012, 197ff.), and Z. Cogley (2015). Such a conflicted state provides room for the agent's exercise of agent-causal power, on agent-causal accounts like Swinburne's. And on Cogley's (2015, 133-6) view, it is *built into* libertarian (incompatibilist) control that in a large portion of nearby possible worlds, the agent would act differently in identical circumstances.

At any rate, I should register a concern about the way that Levy pushes the debate up a level to the metaphysics of luck and control in the first place. For many of those working on the epistemic condition—and for all that I have argued in this thesis—it is still an open question what *kind* of control is required for responsibility, and whether responsibility entails the absence of any luck. For this reason, I will not attempt to contest Levy's metaphysics of control and luck.

A final point to note in response to Levy is that, as Levy himself appears to recognise,[133] the notion of "rational capacities" seems to have been given a perfectly acceptable analysis by Robichaud in allowing for non-decisive motivating reasons to give rise to this capacity. Indeed, I am convinced that it is sufficient for a capacity (to avoid wrongdoing) to be *rational* that the agent has a non-decisive motivating reason (or set of reasons) to avoid wrongdoing that is *at least as strong as* the (set of) reason(s) to commit wrongdoing. Since, as I have

[133] Levy clarifies that the "sense of 'capacity' [he] had in mind [in 2009] was explicitly tied to what agents may reasonably be expected to do or have done" (2016, 4).

already indicated, I will argue that the reasons to avoid wrongdoing must be *outweighing* (for blameworthiness), it follows on my view that sometimes the agent can be excused for wrongdoing despite having (had) the rational capacity to avoid wrongdoing. For this reason, I will not appeal to the notion of rational capacities hereafter.

We have good reason to suppose, then, that we *can* be blameworthy for wrongdoing if we have merely non-decisive motivating reasons to avoid it. Can we go further? Can an agent be directly blameworthy for wrongdoing if a belief or credence of theirs does not even amount to a non-decisive motivating reason to avoid wrongdoing but ought to? I do not think so; the line should be drawn between motivating reasons and beliefs/credences that have no reason-giving weight for the agent.[134] Talbert (2008) disagrees. Talbert argues that an agent who is "unreasonable" in the following way may nevertheless merit the emotional reactions constitutive of blame:

> [Consider] a case in which an agent [on this one occasion] has knowledge of his drink's toxicity in the sense that he is aware of certain facts – like the facts about how his drink will interact [harmfully] with his physical organism – but the agent cannot appropriately grasp the reason-giving status of these considerations. In this case, the agent would be unable to see that he has a reason to refrain from drinking a liquid even if others were to insist that the liquid has the property of toxicity and that this counts as a good reason not to drink it. (p. 527)

Talbert argues that such a "one-time unreasonable drinker" (p. 528) may nevertheless merit blame for drinking it (presumably as long as drinking it is *wrong*), and that recognising the drinker's blameworthiness is critical for appreciating how (philosophical) psychopaths may still merit blame even though they cannot, in principle, recognise certain moral considerations as reason-giving. But on my view, *even if* the drinker can correctly be described as unreasonable (which I doubt), I think that it is a mistake to blame him for it. As far as he can tell, the facts about the liquid's toxicity make no difference to him. Imagine being entirely unmoved by such a fact. To be as morally blind as to fail to see that as a reason seems to me

[134] Note that this distinction is not to be confused with the distinction between beliefs/credences that the agent *believes* constitute reasons and beliefs/credences that the agent does *not believe* constitute reasons. That is not the relevant contrast, because the latter may still constitute motivating reasons (on the grounds that the agent would in fact cite those beliefs in answer to the question of why they should or should not act). The relevant contrast is between motivating reasons, and beliefs/credences *that have no reason-giving weight for the agent*. (A belief may have reason-giving weight for the agent even if the agent nevertheless fails to *believe that* it is reason-giving, understood as such.)

to excuse the drinker (as long as his failure is not his moral fault). Indeed, on my view, he could have drunk the liquid *conscientiously*, under the circumstances, and this would have ruled out his being *morally at fault*, and so *blameworthy*, for it (given my prior arguments linking being morally at fault and being unconscientious; see §5.3, §5.5, and n. 115 in §6.3).

Bearing all of this in mind, then, I propose the following constraint on an adequate account of the epistemic condition on directly culpable conduct:

> ***Non-Decisive Reasons****:* An agent's direct blameworthiness for wrongdoing requires that, at the time of acting, the agent has decisive *or non-decisive motivating reasons* to refrain from wrongdoing.

Returning now to Robichaud's account, we must now ask *which* reasons are the relevant reasons to which we must be responsive. Robichaud's view appears to entail that *any* non-decisive motivating reason to avoid wrongdoing is enough epistemically for us to be subject to a fair expectation to avoid wrongdoing. But is just *any* non-decisive reason going to give us that? Given that Robichaud does not appear to constrain the type of content that may constitute the content of non-decisive motivating reasons, Robichaud appears to think so. But herein lies a problem, echoing a point that Rosen (2008, 593-4) has made about the relevance of "unrelated" reasons. What if the non-decisive motivating reasons in question do not map onto the normative reasons why the act is wrong? We have already seen (in Chapter Five §5.5) that the relevant reasons to which one must be respondent to be responsible for wrongdoing are the reasons *concerning the act's wrongfulness*—and not, for example, its imprudence. But consider another case in which the agent's motivating reasons *are morally* significant, but nevertheless do not map onto wrong-making features.

Breaching a Liquor Ban

Suppose that Jack brought a bottle of beer to a friend's party, not knowing (nor truly believing) that the venue had a liquor ban. Still, let us suppose that Jack had the opportunity to question whether he should bring the beer to the party, because his partner, Jill, felt that bringing only one bottle of beer was stingy (mean-spirited) or selfish. True, he felt that the risk of being stingy was a good (non-decisive) reason not to bring it. But in the end, he decided to bring the beer, thinking it unlikely that his friends would care. It is true: they did not care for reasons of stinginess, and so bringing the beer was not wrong because stingy. It was wrong because of the ban. When Jack and Jill found out about the ban, Jill blamed Jack:

"I told you, you shouldn't have brought it!"[135]

But was Jack really directly blameworthy for breaching the liquor ban? It seems not. The only way that Jack could have avoided breaching the liquor ban was by refraining from bringing the bottle of beer for a reason *entirely unrelated* to the normative reason why it was wrong in the circumstances. Why is this problem? Jack's bringing the bottle of beer to a venue with a liquor ban *is not* his response to the normative reason not to bring it (i.e., that it would violate the ban), because this normative reason *does not play a role in the reasons he takes himself to have*—the motivating reasons—not to bring it. As such, Jack is not responsible for violating the ban, and so not blameworthy. We also get the same verdict about Jack's blamelessness on a consideration of fair expectations and blaming as moral faulting. If Jack could have avoided violating the ban only by heeding an unrelated reason not to bring the beer, then it seems clear that Jack could have avoided violating the ban only *blindly* or *accidentally*. Indeed, had he omitted to bring the bottle, it would have made sense for Jill to say to Jack, upon learning about the ban: "Lucky you didn't bring it!" But if Jack could have avoided breaching the ban only blindly or accidentally, then it was not fair to expect him to avoid breaching the ban, and so he was not blameworthy for breaching it. Finally, Jack is surely not *morally* at fault for violating the ban in the circumstances.[136]

What about a case in which Jack's bringing the beer *was* stingy, after all (and everything else remained the same)? In this case, I say that Jack could have been blameworthy only for *bringing the beer because it was stingy*, not for *bringing the beer because it would violate the liquor ban*. Thus, talk of blameworthiness "for being stingy" would be correct while blameworthiness "for violating the ban" would not strictly be correct (although talk of blameworthiness *in* violating the ban would sound better). I intend this as a variant of Rosen's (2008, 593-4) suggestion that when an act is wrong for two or more unrelated reasons, we should, as far as culpability ascriptions are concerned, split the act into two or more discrete acts (or "wrongs") corresponding to the differing reasons why the act is wrong. In response, however, I do not see that we need to split the act into two fine-grained acts—as

[135] Someone might contest the claim that breaching a liquor ban is *morally wrong*. If that is a problem, I only invite the objector to dial up the moral severity of breaching the ban. Imagine, e.g., that the venue's owners are recovering alcoholics, and bringing beer would trigger a significant relapse.

[136] Some, who hold the view rejected in Chapter Three that risking doing wrong under strictly *moral* uncertainty can itself be wrong, might want to pin Jack for bringing it despite risking stinginess or selfishness. After introducing the concept of being at moral fault (in Chapter Five), I am happy to say that moral recklessness can make someone *at moral fault*, but if they do not thereby do any wrong, then they cannot be morally at fault *for wrongdoing*, and so cannot be blameworthy.

long as the sense in which the act is wrong is specified in the explanation of the reason for which one is blameworthy.

To wrap up our discussion then, I agree with Robichaud that direct blameworthiness does not require having *decisive* motivating reasons against one's conduct, but I contend that Robichaud lets in too much by failing to narrow down the relevant range of reasons to those that concern the act's wrongfulness—namely, to those that are "wrong-related." Thus I defend the following further requirement on an adequate account of the epistemic condition:

> ***Wrong-Related Reasons****:* An agent's direct blameworthiness for wrongdoing requires that, at the time of acting, the agent has motivating reasons to refrain from wrongdoing *concerning its wrongfulness.*

7.3 Wrong-Sensitive Reasons: A Response to Sartorio

Suppose that Jack was aware of the liquor ban and that the ban was a (decisive or non-decisive) motivating reason for him not to bring the beer. Would he also have required, for blameworthiness, the further reason that he "*morally should not* bring the beer," or that "it would be *morally* good not to bring the beer"? It does not seem so. Minimal awareness of a *feature which, in fact, makes the act wrong* (e.g., "there might be a liquor ban") can, for the agent, have enough reason-giving force as to motivate the agent to act in accordance with that (in fact, moral) reason, without the additional *moral* belief/credence "therefore I morally should," or "it would be morally wrong not" to, act in the way suggested by the reason. After all, the agent may be mistaken about what morality requires, without being morally incompetent. We have also seen that on a reasons-respondence theory of responsibility, the objects of response are *normative reasons*, not the verdicts or moral statuses determined by one's normative reasons. Finally, consider that we often bypass verdictive moral judgments such as "I morally should not" or "it would be morally good to" act in some way in our reasoning. As T. Scanlon (1998, 96-7) has envisaged, we "pass the buck," as it were, from morally significant reasons to actions, without verdictive judgments in between. To cite some examples, you judge that you are hungry and that you have access to food, so you eat food; I pay back my friend simply because I judge that she lent me some last week; you judge that it would be honest and so you tell me the truth. This suggests that normative reasons often have

enough force to propel us to action, without additional judgments of moral status. To put my point in language often used in the literature, the agent need not have *de dicto* moral awareness (awareness that the act is morally right, wrong, good, bad, etc.) to be blameworthy for it. *De re* moral awareness (awareness of the features which, in fact, determine the moral status of an act) can be epistemically sufficient for blameworthiness. In particular, I think that awareness of a morally significant feature or normative reason is sufficient when the agent *takes it* as a reason for them to act in some way—that is, when it becomes a motivating reason for them. Thus, I would put the point above as the point that *de re motivating reasons* can be epistemic grounds for blameworthiness.

This is another reason for resisting volitionism and other forms of weakened internalism (e.g., the dispositional belief-in-wrongdoing view) which require *de dicto* moral awareness. One of these views is C. Sartorio's view (2017), according to which the agent must be aware of the "moral significance" of the act, where the moral significance of the act can be analysed in different ways depending on the circumstances (a point to which we will return), as long as the awareness is *de dicto*. Sartorio cashes out her view in the following way:

> [Being] aware of the moral significance of our behavior—could be satisfied in different ways in different circumstances. In circumstances where we act wrongly, it could be satisfied by the awareness that we were acting wrongly, or by the awareness that one ought to have behaved differently. In circumstances where we don't act wrongly, and perhaps are aware that we don't act wrongly, it could be satisfied simply by virtue of recognizing that we are acting from morally reproachable reasons. (2017, 20)

Indeed, the kinds of cases that Sartorio takes to be counterexamples to any belief-in-wrongdoing theorists are precisely the kinds of cases in which the agent does not do wrong, but believes that they act for morally reproachable reasons (i.e., has [higher-order] beliefs that their motivating reasons are morally reproachable). Since this can be enough epistemically for blameworthiness, she rejects volitionism.

Now it is surely plausible that believing that one acts for morally reproachable reasons, say, rather than that one acts wrongfully, can be enough epistemically for blameworthiness. But notice that every one of her examples of "awareness of moral significance" involve awareness of *moral* considerations (morally reproachable reasons), *moral* wrongs, or *moral* norms, considered as such. Sartorio's view therefore falls victim to the objection of the

sufficiency of *de re* moral awareness. Absent any beliefs about moral considerations, someone can still be blameworthy if they believe that the act would not be "safe," "respectful," "honest," "kind," "just" (etc.), regardless of what they believe about morality as such or about the *moral* relevance of these descriptions.

Now, although *de re* motivating reasons can be sufficient for blameworthiness, the important question for constraining the range of relevant motivating reasons is whether *de re* motivating reasons are *necessary*. My answer lies in the negative; I think that *de dicto* motivating reasons can be sufficient. However—and this is significant—I *do* think that one's *de dicto* motivating reasons must be "anchored" in *de re* moral *awareness*, whether or not one can access that awareness and therefore take it as reason-giving. That is, *de dicto* motivating reasons must be *based upon beliefs or credences about features which, in fact, make the act have its moral status*, whether or not those beliefs or credences are accessible.

Let me illustrate. Suppose that Jack is ignorant of the ban but has an inkling (or a low-level credence) that bringing the beer is wrong, albeit for no reason that he can put his finger on.[137] If he were asked *why* he had this inkling of wrongdoing, he would not be able to answer. Suppose further that Jack deems the fact that it *could* be wrong to bring the beer to be a reason not to bring it, and so its possible wrongfulness is a motivating reason for him not to bring it (likely of the "non-decisive" kind, given that bringing the beer was deemed only *possibly* wrong). Now, in this case, Jack's motivating reason is *de dicto*, but he does not have a *de re* motivating reason to avoid bringing the beer because he does not have access to a normative reason why it might be wrong. Still, let us suppose that having been to the party venue before, Jack has some faint memory somewhere in the back of his head of the venue's having a liquor ban. Although now not directly accessible, Jack still has access to this memory *by proxy* through his *de dicto* motivating reason (that bringing the beer to the venue could be wrong). Thus, although Jack's motivating reason (the possible wrongfulness of bringing the beer) is not *de re*, I think that we can safely say that he does still have a minimal degree of *de re* moral *awareness*. Reflecting on this case, my intuition is that Jack can be blameworthy for acting contrary to his *de dicto* motivating reason, precisely because it is based upon his *de re* moral awareness. Some might say, of course, that it is *psychologically necessary* for a *de dicto* moral judgment or motivating reason to be grounded in a *de re* moral belief or credence,[138] but all the better for my point.

---

[137] Intuitively, this inkling would have given him only a non-decisive reason not to bring the beer.
[138] Recall Dancy's view that "to say that it is good or right is merely to express a judgement about the way in which other considerations go to determine how we should act" (2004, 16–17).

I have cited only an intuition in favour of Jack's blameworthiness in this case. Other grounds in favour arise from the reasons-respondence theory of responsibility itself. If Jack did not have any belief or credence in the (normative) reasons why bringing the beer is wrong, then given that responses to reasons requires minimal awareness of reasons, it would be impossible for his behaviour to be a *response* to those reasons and so Jack could not have been responsible for his behaviour.

To keep track of our discussion so far, I have been arguing for the following refinement of the requirement of ***Wrong-Related Reasons*** (to be added alongside ***Non-Decisive Reasons***):

> ***Wrong-Sensitive Reasons****:* An agent's direct blameworthiness for wrongdoing requires that, at the time of acting, the agent has motivating reasons to refrain from wrongdoing that are, or are based upon, the *normative* reasons why the action (or omission) is wrong.

Thus, wrong-sensitive reasons are reasons that are, or are based upon, wrong-making normative reasons. The terminology of wrong-sensitivity has already been introduced in §3.4.1 (however, here we are stipulating a further condition on wrong-sensitivity that any *de dicto* motivating reason must at least be *based upon* minimal awareness of a wrong-making feature, in the way that I explained above).

Now Sartorio would question why blameworthiness should be restricted to cases where the normative reasons are *wrong*-making. Sartorio argues that in cases involving no (objective) *wrongdoing*, the agent can still be blameworthy for the act if the agent was aware that it issued from morally reproachable reasons. But the possibility that one does not commit wrongdoing yet still acts from morally reproachable reasons seems dubious, since *bad motives* are plausibly always wrong-making (Slote 2001, 14). Imagine that Jimmy tries to kick Lara, aware of his morally reproachable reason for doing so (to harm her gratuitously), but misses. Jimmy seems to do something wrong, even though he misses Lara. In these cases, the morally reproachable reasons generate normative reasons why Jimmy does wrong, and, provided that Jimmy recognizes them as reasons for him not to kick Lara, Jimmy therefore has wrong-sensitive motivating reasons not to do so, and so can be blameworthy for trying to harm her.

However, Sartorio is interested not only in "Frankfurt-style" counterfactual intervener cases in which the agent has no objective alternative but to commit some bad act, but also in "*Nelkin*-variants" of Frankfurt-style cases where the agent *believes* (truly) that that they have

no such alternative, given the presence of the counterfactual intervener (see Nelkin 2004). Nevertheless, Sartorio thinks that the agents can be blameworthy for this bad act. Recall that in Chapter Three (§3.3.6), I argued that for an act to be wrong and indeed something for which the agent is blameworthy in a Frankfurt-style case (which may not even involve the "micro-alternative" of *trying* to avoid the bad act), it must *appear to the agent* that she has an alternative: relative to the agent's (false) *beliefs or credences*, the agent must have an alternative. But Sartorio is arguing here that the agent in these cases may be blameworthy for the act even despite awareness of *no* alternatives—and so *a fortiori* a lack of apparent alternatives. Thus, Sartorio's argument here challenges my presumption that apparent alternatives are necessary in a Frankfurt style case for the agent to do something wrong and indeed blameworthy. What is her argument?

Sartorio's thought is that even in a Nelkin variation in which an agent Jones becomes *aware* of the fact that a mad neuroscientist will intervene if Jones falters in his attempt to shoot Smith, Sartorio argues that Jones may still be blameworthy for shooting Smith if he "makes the choice completely on his own, on the basis of his own reasons (morally reproachable reasons, such as a desire for revenge), in exactly the same way he would have made it if he hadn't been aware of the neuroscientist's presence" (2017, 19). In this case, Jones becomes aware of the neuroscientist's intentions "at some point during the process" (resulting in the shot) but in a way that leaves Jones unaffected, preserving his acting on the basis of his *own* reasons.

Sartorio states that she does not intend to question whether belief that one has alternative options for action is required for *deliberation*, only whether it is required for *blameworthiness* (2017, 7-8). As we shall see, I do not think that blameworthiness requires deliberation, and so this is a fair move. But first, I should register that, contrary to some (e.g., Nelkin 2004), I do not see how it is possible for Jones to truly ignore the presence of the neuroscientist *if Jones engages in any practical deliberation*. I grant with Nelkin (2004), that Jones, once aware of the neuroscientist's presence, could still evaluate the reasons for and against shooting Smith. But how could we be sure that Jones' own evaluation of these reasons was not somehow affected by his awareness of the neuroscientist? Perhaps Jones was determined not to let it affect his deliberation, believing that he ought to deliberate *as if* the neuroscientist was absent, or believing (oddly) that he ought to deliberate in the way that he *would* if the neuroscientist was absent. But to say as much would be to concede that awareness of the neuroscientist *would* affect Jones' deliberation, making it no longer immediately a *practical* deliberation but a kind of *hypothetical* deliberation. But even if it were granted that the

"deliberation" was no different in kind, it is not clear that his deliberation would still remain unaffected by his awareness of the neuroscientist. I would expect this awareness to *infect* his evaluation of the reasons for and against shooting Smith, in such a way that he would not give the right weight to the reasons against the act. That is, at least, as long as Jones had not made up his mind already that he was to kill Smith out of revenge. But then if he had made up his mind already, then this does not show that awareness of alternatives is not required *for deliberation*, for he may well have made up his mind *while believing* that he had an alternative. Alternatively, immediate practical deliberation may still be possible for Jones, but only about whether to shoot Smith for a morally reproachable reason or not, or about whether to bother with a vain attempt to resist the inevitable, or to let the neuroscientist force him to shoot (so as to avoid culpability). Recall from our discussion in Chapter Three (§3.3.6), however, that these would be "micro-alternatives"; and it seems that for these, too, Jones would have to be aware of *them* to be blameworthy for failing to perform them.

Now, directly culpable conduct often issues from deliberation, but I allow that sometimes it does not. If the agent becomes occurrently aware of the wrongfulness of the act, fails to deliberate, and acts anyway, the agent can still be blameworthy (for not immediately reacting in accordance with their awareness). See also the cases below in which I argue that the agent is blameworthy while *deliberatively attuned* (§7.6), but not necessarily while actively deliberating. But if direct blameworthiness does not require deliberation, arguing that deliberation requires apparent alternatives does not establish that direct blameworthiness requires belief in alternatives; and Sartorio denies this last claim. Sartorio thinks that Jones could have become aware of the neuroscientist "at some point during the process" and could have remained impervious to that fact. But I struggle to see how even outside of the context of deliberation, Jones' *becoming aware* of the neuroscientist would not affect this process leading to the act. Part of the problem, I think, is that I am not sure what Sartorio is thinking is the relevant "process" (the alternative to deliberation). Let us suppose that you can become aware of something either occurrently (consciously) or dispositionally (unconsciously). An example of the latter would be the case that I mentioned in §3.4.2—that before reflecting upon it, I surely believed (and was aware) that "Kazakhstan is north of New Zealand"; yet surely I *became* aware of that at some point (or gained the disposition to think it or assent to its truth upon reflection or when asked). Now if Jones becomes aware of the neuroscientist *occurrently*, then I do not see how he could not have immediately attached significance to that fact and began thinking about what to do *by* deliberating about his micro-alternatives. But then we have the same problem as above: that deliberation requires apparent alternatives.

The most plausible way of defending the possibility of Jones' remaining impervious to his new awareness is by saying that he—somehow—*became unconsciously aware* of the neuroscientist, and that its non-occurrence is partly what explains his imperviousness to it in the process leading to the act. It is extremely difficult to determine how this could have happened, however, without something in his environment or phenomenology triggering the formation of his new belief (e.g., his having a "vision" of the neuroscientist). But if there *was* such a trigger, it seems that he would have noticed it. And if he had noticed it, then he would have become occurrently aware of it. But then, by my argument above, he would have begun deliberating about it or thinking about it in a way that would have disrupted the way that the process would have panned out otherwise. (Note also the difference between becoming unconsciously aware about the relative location of Kazakhstan and becoming unconsciously aware of the neuroscientist that the former plausibly came about when I learnt some other propositions that entailed it. However, I do not see what analogous set of beliefs Jones could have learned such that they entailed becoming unconsciously aware of the neuroscientist.) Thus, I cannot see how Jones could have become aware of the neuroscientist without this awareness affecting the process leading to the act. At the very least, in the absence of further detail from Sartorio, I think that there is good reason to remain sceptical about the psychological possibility of remaining impervious to new awareness of a counterfactual intervener.

All of this does not show, however, that *belief* in alternatives is required for direct blameworthiness. It shows only that there must be the lack of *disbelief* in alternatives, and more strongly, that the alternatives must be *apparent* to the agent in some way, relative to the agent's beliefs or (non-negligible) credences. This becomes important later (in §7.6) where it is doubtful that in some cases of direct blameworthiness the agent believes that they have an alternative available to them.

Thus, ***Wrong-Sensitive Reasons*** remains intact. Either the agent has awareness of her morally reproachable reasons for acting but *fails to have* apparent alternatives, and so fails to be directly blameworthy, because blameworthiness requires apparent alternatives (if only micro-alternatives), *or* the agent has awareness of morally reproachable reasons and apparent alternatives, and so would be blameworthy for wrongdoing, given that these morally reproachable motives would be wrong-making features of that wrongdoing.

7.4 Outweighing Reasons: A Response to Guerrero

So far we have argued that a blameworthy agent's motivating reasons against his wrongdoing can be non-decisive but must be wrong-sensitive. This brings us to Alexander Guerrero's (2007) "moral risk" view, which incorporates (or comes extremely close to incorporating) both of these elements and proposes a further epistemic condition. What is Guerrero's view? And what shall we make of it?

The central claim in Guerrero's paper is what he calls "Don't Know Don't Kill" (DKDK):

> "[if] someone knows that she doesn't know whether a living organism has significant moral status or not, it is morally blameworthy for her to kill that organism or to have it killed, unless she believes that there is something of substantial moral significance compelling her to do so." (2007, 78-9)

Along with (volitionist) cases in which one *knows* or *truly believes* that a living organism has "significant moral status" (e.g., a person, panther, or petunia), one can still be directly blameworthy if Guerrero's DKDK is met with respect to that living organism. Indeed, Guerrero states that he endorses "neighbourhood principles" with equivalent conditions concerning other kinds of moral risks (e.g., "Don't Know Don't Invade"; 2007, 94) but for the sake of simplicity, I will restrict my focus to DKDK. DKDK entails two key epistemic conditions:

(a) knowing that one does not know whether a living organism has significant moral status or not, and
(b) the lack of belief that there is something of substantial moral significance compelling one to kill an organism or have it killed.

On (a), Guerrero allows for other interpretations of "knowing that one does not know." To remove the higher-order element of this moral uncertainty, the agent might have a non-negligible credence ("0.35-0.65" credence) in the organism's having significant moral status, or the agent might believe that there is a non-negligible possibility or risk of its having this feature (p. 79). This condition is what merits my characterisation of Guerrero as holding that the agent's motivating reasons against wrongdoing "can be non-decisive" but "must be

wrong-sensitive" (although he does not put his view in terms of motivating reasons).[139] Knowing that one does not know whether a living organism has significant moral status would typically constitute only a *non-decisive* motivating reason against having it killed (unless one were a hard-line "Uncertaintist" [Harman 2015] believing that one should never risk wrongdoing under moral uncertainty when given the chance to take a lesser moral risk). And the content of the required beliefs or credences is *wrong-sensitive*, at least in cases where the organism actually has significant moral status. One key difference from the view that I have developed so far, however, is that Guerrero believes that DKDK is *sufficient* for a killing to be culpable, whereas I hold that blameworthiness requires wrongdoing and that wrongdoing, whether of a killing or any other action, is never determined (directly) by the balance of one's moral beliefs or credences (§3.3.5).

On (b), Guerrero states that the lack of belief that "there is something of substantial moral significance compelling her" to act is best given the interpretation that it is a lack of *reasonable* belief that there is something of substantial moral significance compelling her to act (p. 80).[140] Thus, for Guerrero, an unreasonable belief in a morally compelling reason to act would still not excuse the agent who risked killing something with significant moral status. The sense of "compulsion" here is only something like "moral necessity," where:

> the balance of moral reasons (and prudential reasons, insofar as we are morally permitted to take these into account) tips in favor, and perhaps heavily in favor, of [the act]. (2007, 81)

All up, then, if an agent has a living organism killed while uncertain whether it has significant moral status and while lacking reasonable belief in overriding moral (normative) reasons to have it killed, the agent is directly blameworthy for the killing.

To see how the view works and its motivation, consider Guerrero's main example of "Douglas" who is morally uncertain about whether pigs have significant moral status (and thus whether he should have one killed for dinner). Douglas deliberates, but "after some effort, he is still unable to come to an answer to this question" (p. 76). In the end, he resolves to have it killed, even though there are "plenty of other food options that he knows are

[139] He comes close to putting his view in terms of motivating reasons, however. Consider how he characterises the "lack of belief in a compelling moral reason to act" condition, below.
[140] A reasonable belief, for Guerrero, is one that is the upshot of a "a reasonable effort to think about whether the thing apparently compelling them to act actually was of substantial moral significance" (p. 80).

morally permissible, which would be nutritious and which he would enjoy eating, and which would not require killing pigs or any other animals" (p. 76). Now suppose that it is objectively wrong for Douglas to have the pig killed, and the key normative reason is that the pig has significant moral status (or that it has the non-moral features which make it have that status). Is Douglas blameworthy for having it killed? Guerrero thinks so, and I think that is right (as long as Douglas took his uncertainty to be reason-giving). On my view, Douglas appears to have the (wrong-sensitive) motivating reason against killing the pig that it would *risk* killing something which has significant moral status. In consequence, it seems that having the pig killed can be said to be Douglas' response to normative reasons why it is wrong, and so something for which he is responsible. It also seems to fair to expect Douglas to avoid having the pig killed, because his understanding of the moral risk of killing the pig and his appreciation that he has a less risky alternative seems to make it most rational for him.

Guerrero's view has therefore much to be said for it. Guerrero does well to point out a clear counterexample to volitionism in culpable wrongdoing *under moral uncertainty*; others have also followed suit (e.g., Peels 2014, 491; arguably also Robichaud 2014 with his insistence on the relevance of *non-decisive* reasons). I also want to praise Guerrero for introducing something like condition (b), that the agent must lack the reasonable belief there are overriding normative reasons to take the risk anyway. Since we are often morally uncertain, omitting such a condition would mean that we are blameworthy far too much of the time. Moreover, just as taking a risk under *factual* uncertainty is justified if one recognises overriding reasons to do so (e.g., risking causing harm by exceeding the speed limit to get your passenger in cardiac arrest to the hospital), so taking a risk under moral uncertainty (a moral risk) is justified if one recognises overriding reasons to do so (e.g., risking offending someone by saying what needs to be said). Finally, this seems to give Guerrero's view an advantage over Robichaud's view, for Robichaud seems to allow blameworthiness in cases in which one has non-decisive motivating reasons against wrongdoing but outweighing motivating reasons to commit wrongdoing. But Guerrero's view is not perfect.

If it is not already clear, an initial gripe that I have is that merely taking a moral risk, given the lack of reasonable belief in a compelling reason to do so, is insufficient for blameworthiness, on two counts: I have argued that the blameworthy agent must satisfy the control condition on one's conduct being a response to normative reasons, and also that culpable conduct must be *wrong*. Motivating reasons to refrain are also required on my view,

not just beliefs/credences suggesting the avoidance of wrongdoing. But apart from these issues, I think that the more serious problem with Guerrero's view is caused by how Guerrero has characterised epistemic requirement (b). Consider the following example (a type of example familiar to the moral risk theorist: Moller 2011; Weatherson 2014; Field 2019).

Abortion Dilemma

Suppose that pregnant Nora is deeply morally uncertain about whether or not to abort her twelve-week old foetus. She has considered all the arguments and evidence available to her for either side and is still struck with moral uncertainty about it. In fact, her credences are split exactly fifty-fifty between supporting and opposing having an abortion, and for reasons that would in fact make abortion right or wrong.[141] Suppose, moreover, that in this situation, she has no commitment to a principle about what to do under moral uncertainty (e.g., she does not believe that it would be a greater moral risk to abort the foetus than to going to full term).[142] Let us suppose that this may be because she knows that if she has the child, it will experience immense suffering in the future. Suppose, now, that Nora decides to abort her foetus; and assume, for the sake of argument, that she then does something objectively wrong.

Is she directly blameworthy for aborting the foetus? Guerrero must think so, for Nora meets his conditions: Nora is morally uncertain about whether to abort the child and lacks reasonable belief in a compelling moral reason not to abort the foetus. Robichaud must also think that Nora is blameworthy, given that Nora's arguments and evidence against having an abortion would give her non-decisive motivating reasons to avoid having an abortion. I struggle, however, to see how she is to blame at all. As far as Nora can tell, there is *no moral difference* between the alternatives (even, e.g., in terms of what she should do under moral uncertainty).[143] Indeed, it would seem entirely *unfair* to expect her to do one of the things

[141] E.g., she has 50% credence in its being wrong because she has a 50% credence in the foetus' being an actual human person.

[142] Some theories of moral uncertainty hold that one ought rationally or morally to choose the least morally risky option (see discussion in Geyer 2018; Bykvist 2017). Others take "non-hedging" views. E.g., on J. Gustafsson's and O. Torpman's (2014) view, the most rational or morally conscientious choice is the one that is required (or recommended) by one's "favourite moral theory" (e.g., deontology), the theory in which one has most credence, even if it is only a 0.4 credence, say (because one's other credences are divided between a number of other theories—e.g., 0.3 in utilitarianism, 0.3 in virtue ethics). Thus, I am supposing that Nora is not even committed to this kind of principle.

[143] Some (e.g., Gustafsson and Torpman 2014; see above) might argue that, on the basis of their theory of what to do under moral uncertainty, there *is* a relevant moral difference between the options, and so she *can* be to

that, as far as she can tell, is just as morally problematic as the other (even if it might be fair to outwardly reason with her in order to try to tip her one way over the other). Thus, as in ordinary irresolvable moral dilemmas, I propose that in such *epistemically* irresolvable dilemmas, the agent cannot be blameworthy for choosing the wrong option.

I take this conclusion to be supported by my accounts of responsibility and blameworthiness. Does Nora respond *poorly* to the reasons not to have the abortion despite being stuck on the fence about the strength of her reasons? I do not think that she responds *poorly* to them. She might have been highly motivated to act in accordance with what is right; it is just that she could not tell which action to take. Indeed, assuming that she did not culpably create this epistemic conflict, she does not seem to play a morally objectionable role in her response to the relevant reasons, and so she does not seem *morally* at fault for her choice either.

The same conclusion that the agent is not blameworthy should, I think, apply to those who have normative or "radical evaluative" ignorance, believing correctly that their act is impermissible for some actually decisive reason, but failing to accord that reason the right weight in their reasoning and instead according the wrong weight to other morally irrelevant (e.g., purely self-interested) reasons. Recall Rosen's "Bonnie," who is affected by a virus which blocks her usual belief that moral considerations are overriding, and who elbows you into the curb in an effort to beat you to the taxicab in the belief that it is *all-things-considered* the thing for her to do, despite also believing that it was *wrong*, *from an other-regarding perspective*, to do so (because it was cruel, heartless, unfair, etc.). In my view, if Bonnie's wrong-sensitive motivating reasons are outweighed or neutralised by self-interested motivating reasons (remember outweighed *for her, given her normatively upset psychology*), and she cannot (due to the virus) appreciate the way that the (other-regarding) normativity underlying her wrong-sensitive reasons trumps the (self-regarding) normativity underlying her irrelevant motivating reasons, then Bonnie cannot be said to respond poorly to the normative reasons against her conduct. From her corrupted normative standpoint, Bonnie treats her wrong-sensitive reasons as many psychopaths treat them, like fungible social

---

blame, even if she is not herself committed to any principle about what to do under moral uncertainty. My reply to this is that direct blameworthiness is a function of the weight of her *actual* motivating reasons, not on how she *ought* to act under moral uncertainty. Suppose the truth of their principle. My claim here would be that her ignorance of their moral uncertainty principle would be like ignorance of an ordinary first-order wrong-making normative reason in favour of one of the options. Thus, if Nora herself was committed to this principle (in a variation of *Abortion Dilemma*), then it may be fair to expect her to avoid wrongdoing. I am not sure what to say, however, about a case in which this principle is *false*, and yet she believes it, takes it as a reason, and this tips the balance of her motivating reasons in favour of avoiding the abortion. This is an area for further inquiry.

norms, not like the reasons that they really are for any morally competent and enlightened agent. From this perspective, she does not even really respond to the reasons, as they are, in all their overriding normative strength.

Some culpability internalists may object that it is still *reasonable or fair to expect* agents such as Nora and Bonnie to avoid wrongdoing. Perhaps they could take the line that Rosen anticipates in favour of blaming Bonnie, according to which it would be appropriate for Bonnie to think, "'Blame is what you get when you break the moral rules. I knew that in advance, so blame is perfectly appropriate. In this case, however, it's a price I'm willing to pay'" (Rosen 2003, 81). Quite apart from the fact that Rosen is envisioning blame as something that I have already rejected as equivalent to blame (resentment), I would invite the objector to try to empathise with Bonnie, to climb into Bonnie's head. (I am here developing a point I made back in §4.3.2.) Imagine that you are Bonnie, deliberating about whether to push past that stranger to secure the taxicab. You know that it is morally wrong to do so, because it would be unfair to them and cause considerable harm. And yet these considerations do not strike you, in the way they strike others, as *decisive* reasons to refrain from doing so. Rather, they strike you at best with the same force that ordinary *pro tanto* other-regarding reasons strike you when they are overridden by self-regarding reasons. The reason that my friend could use some help would not strike you with much weight, if although you would usually help, you are burnt out, away on holiday, and desperately in need of a break. The reason that lying to a family member is wrong would not strike you with the normal weight, if you knew that telling the truth would endanger your life and limb. Similarly, as virus-afflicted Bonnie, the fact that pushing past the stranger would be unfair and extremely harmful to them would not strike you with the normal weight, given that you will benefit from taking the taxicab instead. But to feel the weight of this reason only in the way would feel the weight of the *pro tanto* reasons just mentioned is precisely what it would take, in my mind, for you to judge that it is all-things-considered permissible to secure the taxi cab for yourself and fail to grasp the decisiveness of the other-regarding reasons not to elbow the stranger to the curb. But if that is so, then it seems to me that we cannot expect you to avoid pushing past the stranger, and we cannot morally fault you for make the decision that you do (just as we cannot fault you for hiding the truth to save life and limb).[144]

[144] I should mention too that someone might object that we can *morally* fault you for it, while reserve judgment that we can fault you *all-things-considered*. But I have already indicated that I have conflated the two for the purposes of this thesis and in my initial discussion of being at moral fault in Chapter Five.

I take the general point here to be supported by considerations that Rik Peels has made about whether ignorance of an obligation as prevailing *all-things-considered* can excuse its violation. He argues that ignorance of an all-things-considered obligation to gather evidence on *p* "may excuse me for violating my all-things-considered obligation to gather evidence on *p*, despite my knowing that I have [a *pro tanto*] obligation to gather evidence on *p*" (Peels 2014, 486). Of course, on my view, this ignorance may not excuse its violation if *all up* one's motivating reasons favoured fulfilling the obligation, but I have intentionally put the case of Bonnie as one where she does not have outweighing *motivating* reasons against elbowing the stranger into the curb. (Notice the implication: sometimes those who are radically evaluatively ignorant may not be so ignorant as to fail to recognise that their motivating reasons weigh in favour of avoiding wrongdoing. In such cases, they may or may not have an excuse for their partial evaluative ignorance.)

I therefore propose the following requirement on any adequate account of the epistemic condition on direct blameworthiness for action—in addition to the requirements of ***Non-Decisive Reasons*** and ***Wrong-Sensitive Reasons***.

> ***Outweighing Reasons*****:** An agent's direct blameworthiness for wrongdoing requires that, at the time of acting, their motivating reasons overall count in favour of refraining from wrongdoing.

It is not enough, in another words, that she merely lacks outweighing motivating reasons to perform the wrong act, for this pronounces wrongdoers in epistemically irresolvable moral dilemmas blameworthy. Rather, the agent must actually *have* outweighing reasons to refrain.

I would like, at this point in the discussion, to move towards combining the three conditions of *Non-Decisive Reasons, Wrong-Sensitive Reasons,* and *Outweighing Reasons* into an overall account of the epistemic condition. To do so, I would like to argue that there is an important relation between the conditions of *Wrong-Sensitive Reasons* and *Outweighing Reasons* that has been implicit (in the cases above) but that I would like to make explicit. In particular, I argue that the agent's motivating reasons must be outweighing *because* of the collective weight of her wrong-sensitive motivating reasons. Call this thesis: ***Outweighing Reasons* because *Wrong-Sensitive Reasons.*** What explains why the blameworthy agent's motivating reasons must *overall* favour an alternative is the fact that a certain subset of those reasons is wrong-sensitive. (Of course, sometimes the agent's wrong-sensitive motivating reasons will exhaust her set of motivating reasons. But most of the time, there will be other—

e.g., self-interested[145]—reasons to factor into the equation of the strength of one's overall set of motivating reasons.) Why think that the *wrong-sensitive* reasons must tip the balance of one's reasons overall in favour of avoiding wrongdoing?

Imagine that although one of the Battalion 101 policemen had outweighing motivating reasons to avoid participating in the massacre, only about 20% of the cumulative weight of those reasons came from *wrong-sensitive* reasons. The rest of the strength came from purely self-interested reasons. Suppose that the policeman took only as a very weak non-decisive reason against executing Jewish women and children that they were humans like him, explaining to his friend that the main reasons he had for opting out were that he could not be bothered, wanted more sleep, did not want blood on his uniform, and had an aversion to gore (whether from animal or human flesh). Suppose, moreover, that he had some ideological ignorance given a mid-level credence in Nazi ideology. In this case, it seems clear to me that it would have been unfair to expect him to avoid wrongdoing, given that his wrong-sensitive reasons would have been swamped by self-interested reasons. Expecting him to avoid wrongdoing in the circumstances would have meant expecting him to avoid wrongdoing predominantly for the *wrong* (note: not wrong-*sensitive*) reasons, but that would have meant encouraging moral irrationality. I also find it hard to see how his participation in the massacre anyway would really have been a response to the *normative* reasons to avoid doing so; he did not really appreciate those reasons in all (or most) of their normative strength. This is essentially the same point that I made above with respect to those like the Bonnie or Bill who have radical evaluative ignorance. We might find it hard to imagine what it is like for these powerful moral reasons to be swamped by self-interested reasons, but I believe that if we were really to step into the policeman's shoes we would see that, as far as he could tell, what he did was genuinely "morally alright." Thus, it seems clear that we need the requirement of ***Outweighing Reasons* because *Wrong-sensitive Reasons*.**

Note that this requirement *does* allow for some slippage between the weights that the agent *should* assign to their wrong-sensitive reasons and the weights that they actually have for her. For example, given the presence of decisive normative reasons against the massacre of innocents (e.g., that it was cruel or would violate fundamental human rights), the policemen should have assigned *decisive* weight to these motivating reasons. Nevertheless, what I am saying matters for blameworthiness is that these wrong-sensitive reasons need only have been strong enough to tip the overall balance of motivating reasons (moral or otherwise) in favour

[145] Or self-interested reasons where these reasons are not wrong-making.

of avoiding the shootings, even if they collectively only weighed 51% in favour of avoiding wrongdoing (for the policemen).

We have therefore arrived at the view that direct blameworthiness for wrongdoing requires *right* and, even if *non-decisive*, *outweighing* motivating reasons to refrain (where "right" reasons are "wrong-sensitive" reasons). This conjoins ***Non-Decisive Reasons*** with ***Outweighing Reasons*** **because** ***Wrong-sensitive Reasons*****.**

There are some details that need to be ironed out which I intentionally avoided doing above to reduce the complexity of the discussion. One is what role the agent's (higher-order) beliefs *about* their reasons play in this account. Still another concerns what we should say about the volitionist's requirement of *occurrent* beliefs. Let me address the first issue for this rest of this Section. The second issue needs its own Section (§7.6), where I will provide my final formulation of the epistemic condition on directly culpable conduct.

Imagine a variant on the case of Nora above, involving "Dora." In this case, Dora is one motivating reason away, as it were, from making her dilemma (about whether to abort the foetus or go through the pregnancy) an *epistemically irresolvable* dilemma. As it stands, the weight of her motivating reasons falls on the side of going through with the pregnancy, such that if she were to commit the abortion, she would be blameworthy for doing so (given that, in this case, abortion is wrong). But now suppose that Dora somehow acquires the (mistaken) higher-order belief *that she has greater reason to abort the child*. Let us ask: does this higher-order belief even-out the balance of her motivating reasons, such that they now do not decide either way (so that she would not after all be to blame for having the abortion)? It is quite possible—and perhaps more probable than on first impression—to have mistaken beliefs about one's motivating reasons, even though one's motivating reasons are the reasons that one takes oneself to have. Being mistaken about the collective weight of one's reasons, such as in the case of Dora, is perhaps not uncommon, given that it is sometimes difficult to bring to mind *all* of one's reasons, let alone compare their weight. Notice that the point here is not the more obvious one that we can have mistaken beliefs about the reasons that *there are* for some action—that is, mistaken about *normative* reasons. Rather, the point is that we can be mistaken about our *motivating* reasons and mistaken about the *collective weight* of our motivating reasons.

In response, I think that we should say that Dora's higher-order *belief* that she has greater reason to abort the child would not even-out the balance of her almost-balanced motivating reasons (tipping them in favour of the abortion). That tipping effect, however, may well obtain if "having greater reason" was itself a *motivating reason* for her—that is, if she *took*

the alleged fact that her motivating reasons favour aborting the child *itself* to be a reason to abort the child—then that *would* even-out the balance of her motivating reasons. We are trying to aggregate *motivating reasons*, after all, not mere beliefs. I say only that the higher-order motivating reason "may well" tip the balance of her motivating reasons, because it is extremely difficult to say with any confidence—and because it is beyond the scope of this project to offer an account of—how we should determine the comparative weight of any given motivating reason (the "comparative weight," being the weight of the reason compared and contrasted with that of other reasons). Moreover, quite besides the issue of the weight of that reason in particular for Dora, we should ask whether a single higher-order motivating reason in favour of one option *would* be enough to fill the gap left by possible first-order motivating reasons in favour of that option. I am sure, however, that if a single higher-order motivating reason made *some* contribution to the overall weight of one's reasons (if only a negligible contribution), and if it made sense to aggregate first-order and higher-order motivating reasons, then the first-order level could be such that *all that was needed* was just one motivating reason of any level in favour of the option in question to even-out the balance of one's reasons.

My point here is just to say that there *are* such things as higher-order motivating reasons, and that they can make a difference to the weight of one's reasons "all up," and therefore to whether or not someone is blameworthy. But it is more likely, I think, for the agent's high-order states, if she has them at all, to amount only to *beliefs* about her motivating reasons, beliefs which fail to count as *motivating reasons* about her motivating reasons. Higher-order states of this kind seem to be acquired only by particularly reflective agents, or by agents on occasions when significant reflection is called for (such as when deciding whether or not to have an abortion).

7.5 Implicit Reasons & Deliberative Attunement: A Response to the Dispositional Belief Theorist

Volitionists require *occurrent* beliefs in the action's wrongfulness, or as Levy (2009, 736, n. 16) calls them, "explicit" or "conscious" reasons, for direct blameworthiness.[146] But

[146] Following remarks made in §3.4.2, I am taking the descriptions "occurrent," "conscious," and "explicit" to be equivalent. Also, we have already noted subtle differences between a *belief* in *x* and *x*'s being a *motivating reason* for the agent (see M in §3.4.2); and the latter in particular is required for the epistemic condition (§7.2). Despite this difference, I will sometimes use them interchangeably in this Section.

dispositional belief-in-wrongdoing theorists challenge this aspect of volitionism in favour of the claim that dispositional beliefs in wrongdoing (or implicit motivating reasons concerning wrongdoing) satisfy the epistemic condition. What shall we say on the matter?

Start with Zimmerman's argument in favour of occurrent belief (initially quoted in Chapter Four):

> if a belief is not occurrent, then one cannot act either with the intention to heed the belief or with the intention not to heed it; if one has no such intention, then one cannot act either deliberately on or deliberately despite the belief; if this is so, then the belief plays no role in the reason for which one performs one's action; and one incurs culpability for one's action only if one's belief concerning wrongdoing plays a role in the reason for which one performs the action. (1997b, 421-2)

Now, a controversial premise in this argument is the premise that if one cannot act either deliberately on or deliberately despite the belief, the belief cannot play any role in the reason for which one performs the action. This premise is indeed a liability which dispositional theorists have sought to exploit on the assumption that Zimmerman is right about his final premise. Peels provides the go-to objection:

> Imagine that I am teaching a class on evolutionary theory and, in the course of my lecture, tell the students that Darwin's *On the Origin of Species* was published in 1859. This is something I tell them because I believe it, because I take it to be something that the students ought to know, because I believe that there are students in the room, because I believe that I can transfer knowledge to my students by telling them something, and so forth. But, clearly, I need not consciously consider all these reasons in order for it to be true that my telling the students that Darwin's *On the Origin of Species* was published in 1859 is based on those reasons. (2011, 581)

Thus, dispositional beliefs can play a role in the reasons for which one acts (see also Haji 1997, 537ff.). Peels then argues that if this is true in general (i.e., in non-moral cases), then it is true in cases of blameworthiness. Let us adapt this case accordingly. Imagine for example, that Rik, the biology professor, also believes *truly* that it would be wrong to tell them that piece of information, because they have a test later in that lecture in which that piece of information will be the answer to one of the key questions. Imagine that Rik only believes

this dispositionally, however. Even so, it seems possible for this belief in wrongdoing to play a role in the reasons for which he acts—not necessarily because it is a *reason why* he acts, but because it could *make a difference* to the reasons for which he acts, or because Rik could be said to be acting "despite" this belief (which Zimmerman takes to be sufficient for "playing a role in reasons for action"). But if so, it seems that dispositional beliefs in wrongdoing are not irrelevant to blameworthiness.

Peels offers further support for this view in that it accounts for some cases that motivate A. Smith's (2005) quality of will view and capacitarian views (e.g., Angela's forgetting her birthday, and Randy's *Forgetting the Milk*; see last Chapter). Agents in these cases often fail to "activate" their dispositional beliefs (Peels 2011, 581).

Quite besides the worry that I have about the active, and indeed metaphorical, language of "activating" dispositional beliefs (given that activating some process would seem to involve intentionally making it occurrent, and so consciousness of it in the first place), I think that Levy provides a plausible reply to Peels. Levy thinks that the notions of rational capacities and fair expectations help adjudicate the dispute:

> [Not] all internal reasons contribute equally to settling what it is reasonable to expect us to do. Our internal reasons include unconscious reasons, reasons which guide much of our behavior. However, we can only reasonably be expected to do what we can do by an explicit reasoning procedure, a procedure we choose to engage in, and when we engage in explicit reasoning we cannot deliberately guide our behavior by reasons of which we are unaware, precisely because we are unaware of them. To be sure, when we engage in deliberation, unconscious reasons continue to play a role—making some options salient for us, for instance, and others relatively pallid—but if we do not take an unconscious internal reason into consideration, we cannot be aware of our oversight, nor can we take steps to correct it. (2009, 736, n. 16)

Thus, contrary to Zimmerman, Levy thinks that "unconscious reasons" can still guide and play a role in our behaviour. But contrary to the dispositional belief theorist, he thinks that this is not yet enough to show that one can be culpable for failing to heed them. The reason is that we are culpable for an act only if it was reasonable (in my preferred terminology, fair) to have expected us to have acted differently at the time, and it was reasonable to expect us to have acted differently only if we could have acted differently by way of an *explicit* reasoning procedure. An explicit reasoning procedure is one we *consciously choose* to engage in, but it

seems that part and parcel of conscious engagement in reasoning is that the motivating reasons with which we consciously reason are *occurrent*. Imagine that moments before sharing the piece of information about Darwin's *On the Origins of Species*, the only reasons that bear on whether Rik should tell his students the relevant fact are the ones in favour of telling his students, and it *does not occur to him* at the time (he forgets) that they have a test later with that piece of information as an answer to one of the questions. Levy would then insist that Rik could not engage in an explicit reasoning procedure with the output of avoiding sharing that piece of information. The only reasons with which Rik could work, in order to consciously engage in reasoning, are the ones that are occurrent.[147] In support of this point in later work, Levy (2017, 255, n. 3) appeals to empirical work on "mind-wandering," according to which people cannot stop their minds wandering unless they gain "meta-awareness" of *the fact that* their mind is wandering.

Although I think that Levy is right that we cannot engage in an explicit reasoning procedure without explicit reasons, I think that Levy is mistaken in thinking that it is fair to expect someone to have acted *differently* only if they had *explicit* reasons to act differently. (Thus, I also think that an agent can be morally *at fault* for wrongdoing, and respondent to the normative reasons against wrongdoing, even though the agent has only dispositional reasons for acting differently.) My argument proceeds from the inescapable intuition that I have of blameworthiness for, and of being subject to a fair expectation to avoid, some actions or omissions that the agent has only *dispositional* reasons to avoid in cases that are clearly non-tracing cases. Consider the following two cases, which bear a strong resemblance to capacitarian cases.[148]

Forgetting My Car's Location. Recently, on my commute home, when I arrived at the bus-station I hopped onto another bus that would take me home, completely forgetting that I had driven my car to the bus-station and parked it there in the morning. The reason that I forgot was (no surprise) that I was in a Randy-type state—I was daydreaming philosophically—and I had not taken my car to the bus-station on my morning commute for some time. Still, before

[147] My interpretation of what Levy has in mind in "explicit reasoning procedures" is that they are procedures in which the agent must assess the weight of their reasons and act accordingly, or run through a practical syllogism concluding in an action, where the reasons (or premises) are clearly consciously present to the agent. Imagine a puzzle where you have all the pieces before you; you just have to put them together. Similarly, in an explicit reasoning procedure, all of the relevant reasons are "before you"; you just have to put them together to see what action they recommend. But if an agent is not conscious of her dispositional reasons, how can the agent run them through such a process?

[148] The four capacitarian cases that I discussed last Chapter will be accounted for soon and in the next Chapter.

this incident, I had never forgotten that my car was at the station when I had chosen to commute that way. And on that day, it never occurred to me before I reached the bus station on my way home that my car was parked there, nor that I would risk not thinking of it upon arrival at the bus-station.

Now, as it happened, there was nothing significant riding on my failure, and so I would deny *wrongdoing* and so blameworthiness for this mistake. Nevertheless, at the time, I certainly held that I was responsible for it, and that it was *my fault*. And I think that if there *were* something significant riding on my remembering the location of my car, it *would* have been fair to expect me to avoid taking a bus all the way home, and indeed to *blame* me had I failed to remember.

Consider another case in which I have a very strong intuition of blameworthiness, a variant of one found in R. Clarke (2017, 65).

Burning the House Down: I almost always remember to turn off the oven after I use it, but sometimes I forget, and when I do, there is never any moment when it occurs to me that there is a risk that I will forget to turn it off. Suppose that one day (once again) I forget that it is switched on when tidying after cooking and leave it on overnight. Suppose, however, that in the middle of the night, one of its safety features malfunctions, emitting extreme heat onto a kitchen surface, and setting it on fire. Suppose that my wife and I awake just in time to get ourselves out of the house before the house burns down.

I think an intuition of my blameworthiness is inescapable here. My wife, my landlord, and I myself would be perfectly entitled to blame me for what happened upon learning the cause and blame me morally. I also have the intuition that it *would* have been fair for them to expect me to turn the oven off. Nevertheless, I did not act contrary to an occurrent belief to turn it off, and there was no time at which leaving the oven on *occurred* to me to be a risk. Am I—contrary to these virtually inescapable intuitions—actually off the hook for my omissions and their resulting consequences, as Levy seems committed to holding?

I am driven by the strength of these intuitions to argue to the contrary. But I should first say something about why I think that I am entitled to make this move. Intuitions of the blameworthiness of unwitting omissions like these are reported to be very strong for both philosophers and laypersons alike. Capacitarians typically report them—not surprisingly, given the kinds of cases that they discuss (see last Chapter §6.4). Quality of will theorists

recognise their intuitive pull and try to accommodate them (see, e.g., Smith 2005; Talbert 2017b; Björnsson 2017). Pluralists about blameworthiness try to accommodate them (Pereboom 2016; cf. Zimmerman 2017). And their being shared by laypersons is also well documented in the literature (as being revealed by "ordinary moral practices," Talbert 2017b, 21; Clarke 2014, 162; cf. "the standard perspective" in Björnsson 2017). However, just as volitionists try to stick to theory in the face of these intuitions, so do some non-volitionists, such as Talbert (2017b) and Rudy-Hiller (2019). Since Talbert is a quality of will theorist, Talbert deems culpable only those unwitting omissions that reveal a "morally criticizable lack of concern" (2017b, 22), and it seems likely that neither of the cases above demonstrate a lack of concern (e.g., for my wife's safety or my landlord's property), so Talbert would resist accommodating these intuitions of blameworthiness. I have, however, already made theoretical moves against the quality of will approach in general (§6.3) together with such an approach to unwitting omissions in particular (§6.4) (as well as argued, against Talbert in particular, that reasonable expectations go hand-in-hand with responsibility judgments; §6.3) and so I do not see a good reason to stick to that sort of theory in the face of these intuitions. Of course, Talbert has an empirical explanation for why we might regularly have these intuitions: *people have the tendency to ascribe ill will when there really is none* (pp. 20-33; cf. his example of our tendency to attribute an instance of dangerous driving to a careless idiot). But although that seems undeniable, the relevance of this consideration depends on the very link between quality of will and blameworthiness that my arguments have called into question. Moreover, I am not so convinced that (in my mind at least) an inclination to find ill will in the cases above is what explains *my* intuition about these cases; the inclination is, rather, about the omissions' being *under the agent's (my) control.*

The more challenging attempt to resist the intuitions in question is found in Rudy-Hiller (2019), who appeals precisely to a conception of control as necessary for blameworthiness, but part of which involves an "awareness-of-risk" condition involving awareness of the risk of a cognitive failure (e.g., failing to remember the location of the car or failing to notice that the oven is on), and a "know-how condition," involving awareness of how to avoid that cognitive failure in the circumstances. Rudy-Hiller would argue that neither condition is satisfied in these cases and so neither case involves blameworthiness. This challenge is particularly poignant because this paper signifies a *retreat* from his earlier (2017) capacitarian position on which these cases—cases of "slips" (since they involve unwitting omissions but plausibly without ill will or tracing; 2019, 727)—qualify as cases of blameworthiness. Quite against the grain of relaxing theory to accommodate intuitions of

blameworthiness, an intuition of the *blamelessness* of slips is what motivates his turn to a more demanding theory of the epistemic condition. His challenge is especially pertinent for my purposes, since the more demanding theory to which he turns is an *internalist* theory, requiring an awareness-of-risk and a know-how condition. And it is pertinent because the kind of awareness that he requires for these conditions to be fulfilled appears to be *occurrent* (although this is not made explicit),[149] and like Levy, this awareness is made necessary to ground reasonable expectations to avoid the omissions in question.

I think that in *Forgetting My Car's Location* and *Burning the House Down*, we can agree that there is no occurrent awareness of the risk of failing to remember the car's location or to turn the oven off (if not to notice that the oven is still on after use), nor is there any occurrent awareness of how to avoid these "cognitive failures." (These can be added to the list of other absences of occurrent awareness in these cases, including the lack of occurrent awareness that I *qua agent* am risking not taking the car home or risking setting the house on fire.) But *without* that awareness, Rudy-Hiller would argue that I am "in the dark regarding the risks associated with allocating cognitive resources in certain ways and therefore… in the dark regarding the *need* to exercise that capacity" (my emphasis, p. 731). Moreover, since there has never been any reason to worry about the consequences of the kinds of cognitive failures at issue (I have not caused any damage due to leaving the oven on), I am "entitled to rely on the good functioning of [my] cognitive capacities without having to put in special effort to shore them up" (p. 732). But now Rudy-Hiller asks the following pair of questions, one concerning the fairness of blame, and the other concerning the fairness of expecting that I anticipate the risk:

> How could it be *fair to blame* someone for a cognitive failure that resulted from her [relying on the good functioning of her cognitive capacities] when she was entitled to do so and hadn't had the opportunity to learn about the possible risks involved? (p. 733)
>
> How could she *[reasonably] be expected* to anticipate [the relevant risk] in the absence of some relevant previous experience that alerted her of the need to exert more vigilance in the particular circumstances she was in? (p. 735)

[149] The examples that he uses appear to involve occurrent awareness (see his example of "Elizabeth"; p. 724) and he stresses the importance of *circumstantial* awareness of the risks involved and how to avoid them. It is difficult to see how this is possible without occurrent thought.

Although these are aimed at capacitarian accounts (e.g., Murray's 2017 account of vigilance), I am going to attempt an answer to both questions in terms of my internalist theory. The challenge that Rudy-Hiller poses in *these* terms is the challenge that for the agent to be directly blameworthy in these cases, the agent must be occurrently aware of the risk of not being aware of the relevant normative reasons for or against one's conduct at the time (the risk of not noticing that the oven is on or remembering that my car is at the station) and so of the need to do something (e.g., pay attention, or try to recall how I commuted in) to guarantee that awareness. Is that true? I do not think so.

On my view, an agent can be directly blameworthy if they perform some act or omission, against which they have only implicit (right and outweighing) motivating reasons, even when they are not occurrently aware of the risk of not recognising the relevant normative reasons and so of the need to ensure that awareness, in circumstances under which they are *consciously attuned to their situation qua situation calling upon deliberation about whether or not to act.* I am going to call this state a conscious state of "deliberative attunement" to one's "choice situation," and I propose that deliberative attunement is required for someone to be blameworthy for an act against which they have only implicit motivating reasons. Alongside ***Non-Decisive Reasons*** and ***Outweighing Reasons because Wrong-Sensitive Reasons,*** I therefore propose the following requirement:

> ***Deliberative Attunement***: An agent's direct blameworthiness for wrongdoing requires that, at the time of acting, she is deliberatively attuned to the situation *qua* situation calling on deliberation about whether or not to perform the action (or omission).

The key argument from fair expectations here is that it can be fair to expect someone to act in accordance with their reasons, even if those reasons do not occur to them and it does not *occur* to them that they need to consult those reasons to determine what to do *and yet* the agent is still *consciously* attuned to the situation as a choice situation (as one ideally calling upon deliberation). The upshot is that although the dispositional belief theorist is right that direct blameworthiness does not require *explicit* reasons to act otherwise, they would be wrong to say that *nothing* needs to be explicit or conscious at the time of acting. However, the answer does not lie in Rudy-Hiller's (2019) requirement of occurrent awareness of the risk of cognitive failure and how to avoid it. A kind of "global" occurrent mental state—called "deliberative attunement"—is all that is necessary, and this is what allows one to

access one's implicit motivating reasons such that we can be the subject of a fair expectation to heed them.

What is "deliberative attunement"? The best place to start, I think, is to contrast the state that we are (often) in when having to make decisions with the state that we are in when *no* decisions are required. Contrast finding a place to park your car with tucking yourself into bed at night; or making a big career decision, with sitting back and watching a play. In the first activity of each pair of contrasts, we find ourselves needing to be engaged, responsive to our environment, and reflective on our reasons to act in certain ways rather than in other ways. They are choice situations that call upon deliberation—upon a conscious consultation of reasons for and against one's options in order to engage in an explicit reasoning procedure to determine how to act. I say that these situations "call upon" deliberation, either in the sense that they *require* deliberation or *recommend* deliberation. They sometimes require deliberation because they are unfamiliar to the agent or they generate too much uncertainty about how to act; one cannot always rely on one's habits or practical intuitions to determine how to act. But at other times, the situations only *recommend* deliberation because, although one could do the right thing intuitively or by force of habit—that is, *non-deliberatively*—deliberation would *ensure* or maximise the chances of one's acting rightly. (Notice that it is when the situations do not *require* deliberation that the agent is "entitled"—in Rudy-Hiller's 2019 sense—to rely on their reasons-responsive capacities. Nevertheless, *it would be better* for them to deliberate to ensure right action.) Normally, we appreciate these situations *as* situations that call upon deliberation. But when we do, our appreciation is not reducible to occurrent beliefs with the content that we ought to deliberate or reflect on our reasons, nor reducible to explicit motivating reasons in favour of deliberation (e.g., "I need to stop and think before I buy this"), nor even to a disposition for these beliefs or motivating reasons to be explicit. Rather, this appreciation is reducible to a kind of global conscious state or "mode" that is "active," "operative," or "online"[150] in the background, which may cause us to deliberate or consult our reasons, but which also disposes us to have an distinct array of non-propositional experiences—for example, of feeling "alert," "attentive," "wired," "ready-to-act," "switched on," "responsive," "vigilant," or sometimes "reflective," "hesitant," "uncertain," "indecisive," or "stuck." Call this array of experiences, a phenomenology of

[150] This sense of "online" would be different from the sense used by N. Levy—see below.

deliberative alertness.[151] (To probe somewhat further into this phenomenology, what seems to unite these experiences is also a feeling of being *unsettled* with our situation. When we are deliberatively attuned, we know that we cannot sit back and watch the world go by. We feel, in other words, that we need to *settle* something about our situation—and that is, of course, to settle how to act—although we need not have the occurrent thought that we need to settle anything.) I want to say, too, that deliberative attunement disposes us to experience this phenomenology of deliberative alertness, either as a persistent "feeling" of alertness, or as a succession of individual experiences whilst in the situation to which we are attuned (I am open to either possibility). It *does* matter, however, that one would not have this state of deliberative attunement if one did not encounter its phenomenology. The state cannot be purely functional, at least in conscious beings like us. A further feature is that it is triggered by features of one's situation which signal it as a situation calling upon deliberation. This is due to the combination of one's goals, desires, or current activities and one's environment (as one comes up to a road sign, sits down to write something, sees work that needs doing, or becomes conscious of the consequences of one's options).[152] Finally, it appears true of being in this state that if we were asked whether we should stop to think before we act, whether there was anything to remember or to think of, whether it would be good to weigh up our options or consult our reasons—that is, to deliberate—we would often answer affirmatively, and give reasons in favour of doing so, even if we also took ourselves to have reasons to "wing it" or "decide on the go." This state itself appears to be associated with (higher-order) motivating reasons in favour of the mental action of bringing our (first-order) motivating reasons to bear on our conduct. And these higher-order reasons may themselves be explicit but need not be.

Contrast this state of deliberative attunement to the state in which we find ourselves when tucking ourselves into bed at night or sitting back to watch a play (or having a drink with a friend, or quietly enjoying our commute, etc.)—that is, when no significant decisions are required. We are not really in choice situations, and we do not appreciate our situation as ones that call upon deliberation. We are not disposed to deliberate or consult our reasons, or

[151] I should note that I think that an element of weariness or fatigue may also accompany deliberative attunement, especially when one finds that one's mental resources are rapidly diminishing (e.g., when tired at the tail-end of a workday). But in these cases, a sense of being *alert enough* to make a decision or of the *unfinished* nature of one's current activity would indicate deliberative attunement.

[152] I have found D. Pereboom's (2016, 185) account of vigilance (on which, more below) particularly illuminating for how to formulate my conception of deliberative attunement. That deliberative attunement is triggered by certain environmental features is a condition parallel to one that he offers on vigilance, and I also owe the language of "attunement" to Pereboom.

to have thoughts to the effect that we should deliberate. Nor do we encounter the phenomenology of deliberative alertness (and "unsettledness") identified above. Instead, we feel "weary," "ready to rest or switch off," "relaxed," "at peace," or "settled." There is also nothing in our situation that triggers deliberative attunement (at least in the absence, e.g., of an email that reminds us to do something before bed, or a request to shift one's legs by the person coming down the aisle). Moreover, we do not take ourselves to have (higher-order) reasons to consider our (first-order) reasons. Rather we would take ourselves to have outweighing reasons to "switch our brain off." The difference between this state and a state of deliberative attunement is critical to my argument.

There is a literature-based reason why I have described the state of deliberative attunement as one in which we might feel "vigilant." We have already seen an appeal to this notion in Rudy-Hiller (2019). The capacitarian, S. Murray (2017), appeals to vigilance as the central *capacity* that is required for culpability in the absence of awareness of wrongdoing. Murray describes it as follows:

> The vigilance of an agent consists of a disposition to become occurrently aware of morally or prudentially relevant considerations that constitute a sufficient reason to act or omit. (2017, 513)

It involves "attunement to the moral environment" (p. 516). Notice Murray's "vigilance" is similar in scope to what I am calling deliberative attunement. Both involve a level of attunement to the moral environment, a disposition to be occurrently aware of morally significant normative reasons bearing on one's conduct. However, deliberative attunement involves the disposition to be occurrently aware of *motivating* reasons bearing on one's conduct, and so normative reasons only to the extent that they constitute motivating reasons for the agent (i.e., the moral environment, filtered through the lens of what appears reason-giving). What is importantly different, however, is that deliberative attunement has that persisting background phenomenology of deliberative alertness. It is not *just* a disposition to have occurrent thoughts. Vigilance, as I see it, however, is narrower than Murray's general disposition to be occurrently aware of the relevant normative reasons. Rather, I think that vigilance is better defined by Derk Pereboom, as:

> a persisting attunement to protect, which features, among other things, a standing disposition to respond to danger, triggered by indications of danger in the

environment. (2016, 185)

Vigilance shares in common with deliberative attunement a kind of "persisting" alertness, with a "disposition" to act in certain ways, and with the function that it is "triggered" by features of one's environment. But vigilance concerns attunement to *danger*; by contrast, deliberative attunement is attuned to the need to make decisions and to the instrumental benefit (or requirement) of deliberation. I have also tried to describe deliberative attunement in terms of its persisting *phenomenology*, and its being grounded in motivating reasons to deliberate. This, at any rate, is why deliberative attunement may, but need not, involve vigilance (as Pereboom has defined it). I should note finally that the commonality between deliberative attunement and Murray's wide-scope vigilance means that deliberative attunement gets to enjoy some of the empirical support underlying Murray's concept of vigilance in the neuroscience of cognitive control (2007, 517-521).

Consider now our two cases above—*Forgetting My Car's Location* and *Burning the House Down*. Recall that these are cases in which I have right (recall: wrong-sensitive) and outweighing *implicit* motivating reasons to avoid what I end up doing (taking the bus home rather than my car and leaving the oven on.) Nevertheless, in both cases, there is a critical moment or period of time when I am attuned to the situation *qua* choice situation calling upon deliberation (a situation only *recommending* deliberation, given that the situation is not unfamiliar). In the first case, it is when I need to decide what to do next once I have arrived at the bus station on my way home. In the second, it is when I go about tidying up after my cooking. Of course, in both *types* of cases, I normally trust myself to make the right decision automatically, or without deliberation, but I certainly do not have outweighing reasons against deliberating. Rather I normally show signs of deliberative attunement. I normally feel alert and wary of the need to make the right decision or carry out the next task to fulfil my goal. The reasons why I should act in one way rather than another are frequently occurrent to me ("oh that's right, I left my car at the bus station this morning"), and if asked whether I should stop and think about what I am doing, I would probably retort that "it wouldn't hurt." It seems reasonable then to attribute to myself deliberative attunement in both *Forgetting My Car's Location* and *Burning the House Down*. And, thus, the door is opened to my being blameworthy for my unwitting conduct.

Is it *fair to expect* me to remember my car's location (and take it home), or to expect me to turn the oven off, even though I only have implicit reasons in favour of doing so? I think that the answer is "clearly yes." Although it is not occurrent to me that "I should not get into the

bus home because I left my car at the station this morning," nor even that "I should think about what to do next, given that I am now at the bus station," I am nevertheless conscious of the situation *qua* choice situation calling upon deliberation. Deliberative attunement is, if you like, my *general* posture, mode, or mental state at the time of needing to make the decision. And crucially, precisely *because* I am in this conscious state, it is *fair to expect* that I do the right thing—by thinking of the reasons that I have for doing the right thing and actually doing it. (Note that I do not think that it needs be fair to expect *that I think of my reasons*, only *that I do the right thing*, given that I *can* do so by thinking of my reasons; more on that below.) What is significant about this proposal is that fair expectations are still tied to a conscious or "occurrent" state, but a kind of global mental state sometimes without explicit propositional content. The implication is that I agree with Levy (and Rudy-Hiller 2019) that fair expectations track what one can do from an occurrent state, but that they do not necessarily track what one can do given the occurrence of the reasons themselves or given the occurrence of the need to consult them.

Conversely, it is my contention that it would be unfair to expect me to perform these acts if I was not deliberatively attuned. Imagine, for example, that although it was my responsibility (obligation) to clean up after myself, one of our guests had poisoned one of the drinks, causing me to sick up, lose most of my control, and become bed-ridden for the rest of the evening. It seems, then, that it would not have been fair to expect me to turn off the oven; and that is because I would not have been attuned to the situation qua situation requiring that the oven is turned off, even though I had a decisive albeit implicit motivating reason to turn it off, and even though I arguably still had my obligation to turn it off. Alternatively, suppose that just as I arrived at the bus station, I saw a horrific accident involving another bus crashing into the bus station and injuring several people.[153] Suppose that in consequence, I could not "think straight" and I found myself getting onto the bus home rather than walking to my car. After witnessing the incident, I would have been attuned to the situation *qua* situation calling upon deliberation *about how to respond to the incident*—whether or not to stay and help—rather than attuned to the situation *qua* situation calling upon deliberation about *whether or not to take the bus home*, even though I had a decisive implicit reason not to take the bus home. This case illuminates how deliberative attunement is focused on a specific decision or problem needing to be settled, or to a specific choice situation rather than another.

[153] R. Clarke (2017, 65) discusses a variant of his case (identified last Chapter as *Forgetting the Milk*), where the agent witnesses a "horrible traffic accident." For Clarke too, this counts as an excuse for omitting to buy milk.

In consequence, one can be deliberatively attuned to a choice situation whether to *x*, while simultaneously not attuned to a choice situation whether to *y* (even though both decisions are needing to be made at the same time). At any rate, notice that both these examples of incidents which undermine the required deliberative attunement are rather unlikely. For the most part, and in the absence of odd external factors like the ones above, we have deliberative attunement with regard to the decision about *x* when we are in a choice situation whether to *x*.

This observation is crucial, I think, in accounting for the capacitarian cases mentioned in the last Chapter: *Hot Dog* and *Forgetting the Milk.*[154] Concerning these two cases, capacitarians get the intuition that Alessandra is *directly* blameworthy for not rescuing the dog from the car, even though she gets caught in a tangled tale of misbehaviour, administrative bungling, and so on. Likewise, capacitarians get the intuition that Randy is *directly* blameworthy for omitting to stop at the store when coming up to it, even though he is distracted by philosophising. But even though these agents have *implicit* motivating reasons not to omit at the time of their omissions, it seems to me that they are not *directly* blameworthy for the omissions to rescue the dog or stop at the store; rather they are likely to be *indirectly* blameworthy for these omissions *via* direct blameworthiness for the earlier omissions to open the car windows (for ventilation) or to take the dog, and to take steps to ensure that one would remember to buy milk (e.g., setting a reminder on one's phone, or keeping the task before one's mind). Alessandra and Randy are, of course, blameworthy for their eventual omissions only to the extent that they are *indirectly* blameworthy for them, and I will argue that they are indirectly blameworthy in the next Chapter. For now, however, let us focus on the claim that they are *not* directly blameworthy for their eventual omissions (and the consequences of those omissions), even if they are *directly* blameworthy for their earlier actions/omissions. This claim goes against both capacitarian analyses and the dispositional belief analyses (in terms of direct blameworthiness. Take Alessandra, for example. It seems to me that being caught in the tangled tale at the school is quite enough to undermine Alessandra's deliberative attunement to the situation *qua* situation requiring a decision *about whether to stay at the school or go back to the car.* She is deliberatively attuned, instead, to the situation of how to respond to the tangled tale of misbehaviour and administrative bungling. She does not feel any unease or uncertainty about her decision to leave the dog in the car, but not because she regards that decision with confidence and equanimity. Rather,

[154] Since I argued in Chapter Five that the other two cases discussed in the context of these cases (*On the Rocks* and *Secret Service*) are better taken to be cases of indirect blameworthiness, I discuss them in the next Chapter.

she has become immersed now in another problem—the problem involving her children—a solution to which requires her full attention. Nevertheless (as she will no doubt feel most acutely herself) Alessandra is plausibly blameworthy for the harm to her dog. However, she is not *directly* blameworthy for the harm to her dog, because her failing to return to the car in time is neither by deliberate choice nor while deliberatively attuned to the choice situation of whether to return to her car in time. Thus, I think that we should look earlier to find the source of her blameworthiness, and I contend that a plausible candidate for such a moment would be at the point when she decides to leave the dog trapped in the car in the first place. Plausibly, as a normal adult who is familiar with the risk of harming dependents when left in a boiling car without ventilation, Alessandra would take this very risk as a very strong—and I think outweighing—motivating reason not to leave the dog trapped in the car, even if she is not occurrently thinking about this risk (having left the dog locked in the car countless times before). Moreover, at the time of making the decision, she *does* seem attuned to her situation *qua* situation recommending deliberation about *how to go about her (usual) pickup of the kids,* given the day's temperature and the presence of the dog. This is the problem she must now solve, and she is attuned to her situation *qua* situation calling on a solution to this problem. Her situation calls upon a consultation of her reasons in favour of the different options available to her for going about her pickup of the kids, and one of those reasons is the decisive (although implicit) reason, in favour of not leaving the dog in the car, namely, that leaving the dog in the car *would risk harming the dog, through heat prostration.* Although she is not conscious of this reason, she could easily have become conscious of it given her conscious state of deliberative attunement (and its associated phenomenology). And plausibly she has an obligation to prevent precisely the sort of risk that she poses to the dog, and yet violates it. This, then, appears to be the locus of blameworthiness. For reasons of space, I shall not detail how I would apply the same thinking to *Forgetting the Milk,* but suffice to say that I think that the locus of blameworthiness is either at the point where he notices his mind wandering into philosophy whilst driving, or at the point where he fails to take the relevant precaution (against forgetting the milk) moments after promising to his wife that he would buy milk on the way home. These are the moments in which he has a decision to make and is deliberatively attuned to his choice situation.

Now the volitionist, and Rudy-Hiller (2019), will likely object that without the agent's occurrent belief that they should deliberate or consult their reasons for what to do next, the agent cannot really engage in an explicit reasoning procedure with the aim of doing the right thing, and so cannot fairly be expected to do so. To this objection, I have four replies. The

first is to dig my heels in—but with the reassurance of remarks that Robichaud has made who arrives at the same impasse. *Either* the relevant motivating reasons must be explicit and "we would not be able reasonably to expect agents to perform certain automatic actions" (Robichaud 2014, 150) such as the actions performed in the two (original) cases above, *or* the motivating reasons to do otherwise can be implicit and we can account for cases of automatic actions (and other capacitarian cases, see next Chapter), all the while avoiding responsibility revisionist implications. Robichaud himself regards the implications of the former "unacceptably strict" (p. 145) and I am inclined to agree, on the inescapable strength of my intuitions of blameworthiness and of what it is fair to expect in the relevant cases. My second point is that even if I were convinced that these intuitions should not be as strong as they are—and that an analysis of fair expectations cannot help to decide the matter—I would appeal to my intuitions of being straightforwardly *morally at fault* for my habitual actions (or omissions). Surely, I am morally at fault or "in the wrong" for burning the house down. I would also appeal to what I believe are the favourable implications of my reasons-respondence theory of responsibility, to which I will turn in a moment. A third reply attaches significance to two (apparent) concessions made by volitionists themselves. The first is Zimmerman's (1997b) odd—and often neglected (but see Timpe 2011, 11)—qualification that his argument that culpable conduct requires conscious advertence to the reasons has "one possible exception," namely, and not surprisingly, in cases of "routine or habitual action" (1997b, 422). For Zimmerman, if these cases do not involve conscious advertence to one's reasons (as I have argued, but on which he remains agnostic), Zimmerman concedes that his argument for the requirement of occurrent awareness would not apply to them. I do not think that we should overlook Zimmerman's (1997b) shakiness about the requirement of occurrent belief on directly culpable conduct. It culminates in the end with his rejection of this requirement in 2017 where he defends the view that directly culpable conduct is "willing" conduct, allowing acting merely *in* ignorance (and likely cases of habitual conduct). The other example is Levy himself, who appears to de-convert from volitionism in his book, *Consciousness and Moral Responsibility*" (2014, 31), and argues that requiring occurrent beliefs is "too demanding."[155] Instead, he argues that the agent must be conscious of the "morally significant facts" that play a role in explaining the act's moral valence, where by "conscious," he means that the morally significant facts must be "online"—that is operative in the actual reasons for which the agent acts (the agent's *explanatory* reasons)—and

[155] Rudy-Hiller (2018, n. 3) appears to interpret Levy in this way.

"personally available," that is, "easily and effortlessly retrieved" (without any special prompt, such as in response to the question posed to them, "what are your reasons for doing that?"; 2014, 33). My final reply is, I hope, the most significant. As I have stressed, deliberative attunement is characterised by a phenomenology of deliberative alertness (and unsettledness). It is precisely because we *feel* something—this alertness, this readiness-to-deliberate, a sense of being unsettled about our situation, or a sense of uncertainty about how to act—that we rationally *can*, and that it is fair to *expect* us to ensure, that we do the right thing. I say that we *can* do it, by heeding this feeling and consulting our implicit reasons for and against the act (e.g., noticing that feeling of uncertainty about how to act, and so asking ourselves for our reasons). Our deliberative attunement allows us to notice this phenomenology of alertness, since this phenomenology is a noticeable aspect of our stream of consciousness in the circumstances (either as a stable "feeling," or as a succession of individual experiences, as I noted before). Just as gazing upon the beautiful countryside allows us to notice features of the countryside and act in accordance with that occurrent awareness, so being deliberatively attuned to the situation allows us to notice this phenomenology of deliberative alertness and then to deliberate and make our implicit reasons explicit for an explicit reasoning procedure. No *proposition* initially needs to be occurrent in our minds to ground a fair expectation to do the right thing. Thus, we have an occurrent state giving rise to the possibility of a reasoning procedure brought about by the agent herself, quite as the volitionist wants. However, the state is wide enough as to capture cases in which there is no *occurrent belief* in the wrongfulness or wrong-making features of the act, or in the need to be aware of these features.

Thus, the claim is that, in virtue of its phenomenology of deliberative alertness, only deliberative attunement can ground the relevant *capacity* to bring our implicit reasons to light and do the right thing. Moreover, this deliberative attunement and occurrent phenomenology also provides us with a *fair opportunity* to bring our implicit reasons to light, such that without it, we would not have this opportunity. Indeed, I would say that deliberative attunement is what enables the "information" stored in our implicit motivating reasons to be "personally available" in Levy's sense, entailing availability for "easy and effortless retrieval" (2014, 33). Notice that in light of this appeal to capacities and opportunities, this view shares features with capacitarianism. But because, on my view, the agent can deliberate from this state partially because they have this phenomenology of deliberative alertness, it is really up to the agent whether or not to consciously choose to engage in deliberation, and so I need not provide an account, as capacitarians do, of the conditions under which one has the

capacity to have occurrent awareness of one's reasons. Although I have said that in deliberative attunement one is *disposed* to have this occurrent awareness of reasons (where I meant that one *would* be aware of them in a suitable range of similar situations calling upon deliberation), I have also said that the agent can *bring about* occurrent awareness of these reasons *on account of* the agent's already present phenomenology of deliberative alertness (or unsettledness). And the only reason the capacitarian must provide such an account of capacities is that they cannot help themselves to something that they agent can bring about, *given an occurrent state*. I should note finally that my view differs from some capacitarians, too, in that it is no part of my view that the agent is always *obliged* to be occurrently aware of their reasons, however I am open to the claim that they always *ought*, only in the non-obligatory sense corresponding to the *recommended* nature of deliberation, to be occurrently aware of their reasons. For that way, the agent can ensure that they do the right thing, rather than risk failing to do so by entrusting themselves to habit.

Someone might well push back on the idea that deliberative attunement allows one to have the capacity and opportunity to retrieve one's reasons, on the grounds that a mad neuroscientist can prevent access to one's reasons during the agent's deliberative attunement. I acknowledge that this would undermine blameworthiness, however my reply is that either it makes no sense to say that the agent has *motivating reasons* to refrain in these circumstances (because the agent could not in principle express beliefs/credences as reasons), or simply that deliberative attunement *without interference* is sufficient for having the relevant capacity and opportunity to recall reasons.[156]

So far in this Section we have defended the requirement of deliberative attunement on the basis of a consideration of fair expectations and of being morally at fault. But is deliberative attunement required for wrongdoing to be a poor response to reasons (i.e., something for which the agent is responsible)? Is a response to reasons possible without deliberative attunement? I think it is safe to say that this is not possible. Consider the variants of the two cases above in which I get poisoned, or I get into the bus home after seeing a horrific accident. It is plausible that my omission to turn the oven off, or to walk to my car, are not my poor responses to the reasons why these omissions are wrong in the circumstances. I am too distracted to be appropriately described as respondent to those reasons. Consider also that in these cases without the relevant deliberative attunement, I (the agent in these

[156] The appeal to the lack of interference from manipulation is, of course, a common appeal in the literature (as we saw with H. Smith (2011) in Chapter Six).

circumstances) would take myself to have outweighing motivating reasons in favour of denying that the situation calls upon deliberation—that is, apt for reflection on my reasons. I would deny that reflection on my reasons would be *relevant* or *fitting* in the circumstances. But surely for someone to respond to normative reasons why the act is wrong, they must deem reflection of reasons of that kind to be apt, relevant, or fitting in the circumstances. Consider an analogy: for a public servant to respond to advice given to them, they must deem reflection on that advice to be apt, relevant, or fitting. And so, it seems, for reasons-respondence: by lacking deliberative attunement, the agent does not deem the situation as one befitting reflection on one's reasons, and so whatever the agent does cannot be their *response* to the underlying normative reasons. (In my mind, this holds also for when the agent ends up doing the right thing. The agent may have been *unconsciously guided* by her implicit motivating reasons, but if she was not deliberatively attuned to her situation, then her right action could not have been her right *response* to those reasons.) We can see here how I am requiring that a "response" includes an element of consciousness (although the agent's response to reasons need not of course involve deliberate choices to heed/defy explicit motivating reasons).

But perhaps a stronger criticism of this kind comes from the other direction, from the point of view that right and outweighing implicit reasons plus deliberative attunement is not enough (together with the other responsibility conditions) to guarantee a response to the relevant normative reasons. It may be necessary. But it might be objected that the agent's wrongdoing would not be correctly described as a "response" to the relevant normative reasons if the corresponding motivating reasons are not explicit for her, even if she has deliberative attunement. It is natural, after all, to assume that if someone "responds" to some information, they are (or have been) conscious *of that information*. This is a fair worry, but I have several points in reply. The first is that, since Chapter Five, I have deprived "response" of any connotation of "conscious decision" at the time of the response by proposing that an act (or any other consequence of an action—e.g., a belief) can sometimes be said to be an *indirect* response to reasons when it is consciously foreseen. A second point is that sometimes we use the term "response" to describe a point that philosopher A makes in reaction to another philosopher B, even though A predated B and could not have consciously or deliberately responded to B. For example, it makes sense to say that David Hume has a good *response* to William Paley's "Design Argument," even though he predated Paley. Admittedly, neither of these responses entirely dismiss this worry. But I would like to make one more response, which encourages us to take a step back. Perhaps this is one of the areas

where theory does not provide enough information to derive implications for specific cases, leaving us to consider our basic intuitions about these cases, or to consider the implications of other aspects of the overall theory (e.g., intuitions about fair expectations and being morally at fault). But in relation to these "other aspects" of the overall theory, I would find it surprising that at this point, what counts as a response to normative reasons diverges from a consideration of the implications of fair expectations and of being morally at fault. After all, so far, we have seen that they have consistently arrived at the same verdict. It may in the end be contended that it is *too much* of a stretch to say that acting contrary to the relevant motivating reasons, when deliberatively attuned to the situation, constitutes a response to the underlying normative reasons, when one's wrong-sensitive motivating reasons are not explicit. But all this would show, I think, is that "response" was the wrong word for the "transaction" (Yaffe 2018, 344) between agent, normative reasons, and conduct, that I am contending must obtain for the agent to be responsible for their conduct (or, perhaps, the wrong word for the "exhibition" of a "defect" [Husak 2016, 154] in one's response to reasons in order for them to be blameworthy for their conduct). In the absence of a better word for this transaction (or defect-exhibition), I continue to use the notion of the "the agent's response."

7.6 Conclusion: The Epistemic Condition for Direct Blameworthiness

I have now come to the end of my discussion of deliberative attunement and implicit reasons. It is time now to combine the above requirements together into a formulation of the epistemic condition on directly culpable conduct, a formulation which lists these as necessary conditions but also as *jointly sufficient.* In my mind, there are *no* other requirements for the epistemic condition on direct blameworthiness than the ones that we have discussed above. The consequent account is therefore *complete*. Combining ***Non-Decisive Reasons, Outweighing Reasons because Wrong-Sensitive Reasons,*** and ***Deliberative Attunement,*** we get:

> **E-DB:**[157] An agent satisfies the epistemic condition on *direct* blameworthiness for wrongdoing, in some choice situation C, *if, and only if,* at the time of acting: (1) the agent has decisive or non-decisive motivating reasons to avoid wrongdoing that are,

[157] Where "E" stands for "epistemic condition" and DB stands for "direct blameworthiness for wrongdoing."

> or are based upon, normative reasons why the act is wrong; (2) these reasons tip the balance of one's motivating reasons overall in favour of avoiding wrongdoing; (3) these reasons are either explicit or implicit; and (4) the agent is deliberatively attuned to C *qua* situation calling on deliberation about whether or not to perform the action (or omission).

A couple of features of E-DB need addressing. It will be noted, following ***Deliberative Attunement***, that condition (4) requires that the agent must have deliberative attunement for their wrongdoing to be blameworthy *whether or not* the agent's motivating reasons are implicit. But I only argued in §7.5 that deliberative attunement is required when the agent's motivating reasons are *implicit*. Why, then, say that the agent must be deliberatively attuned when they have *explicit* right and outweighing reasons to refrain?

Imagine that in a variant (on the case above) in which I get distracted upon arrival at the bus station by a bus crashing into the station, I nevertheless have the persisting explicit (or occurrent) reason for walking to my car that I left the car at the station that morning. Since I argued that in the original bus-accident variant, I am not deliberatively attuned to the choice situation of whether to take the bus or car home but rather to how to react to the bus accident, am I still excused in *this* variant for not taking my car home (when something of moral significance depends on my doing so), even though I am consciously thinking about the fact that I left the car at the station in the morning? I hope it is clear that the answer is "no." But this need not cause me to revise my judgment that deliberative attunement is required, for it seems clear that *because* I am occurrently aware of the location of my car (and take that as reason-giving), I *am* deliberatively attuned to the choice situation about whether to take my car or a bus home *as well as* to how I should respond to the bus accident. I do not see how I could not be. Thus, on my view, explicit motivating reasons *trigger* deliberative attunement, and so regardless of whether the reasons are explicit or implicit, deliberative attunement is required.

Now, E-DB is an account of the conditions that are necessary and *jointly sufficient* for satisfying the epistemic condition on directly culpable conduct. This is because I do not see that there are any other epistemic requirements that need to be satisfied for direct blameworthiness. Of course, I welcome any attempts to show that I have left something out, but in the absence of any such attempts, I remain confident that E-DB is a complete account of the epistemic condition on direct blameworthiness.

What this view entails about culpability revisionism will be teased out in Chapter Nine after I have set out and defended a matching account of the epistemic condition on indirect blameworthiness for wrongdoing (next Chapter). After all, if the agent fails to be directly blameworthy for her wrongdoing, she may still be *indirectly* blameworthy for it, as I argued in Chapter Five (§5.6).

Chapter 8
The Epistemic Condition for Indirect Blameworthiness

8.1 Introduction

Last Chapter, I put forward my account of the epistemic condition for directly culpable conduct, E-DB. This condition is satisfied if and only if, at the time of the (wrong) act, the agent has right and outweighing motivating reasons to refrain from wrongdoing that are either explicit or at least consciously accessible through deliberative attunement. Now, E-DB is intended as an account of the epistemic condition on direct or original blameworthiness for *any* type of action, but last Chapter we applied the account exclusively to ordinary "non-benighting" acts which do not lead to ignorance (or further unwitting wrongdoing). It is the object of this Chapter to extend this account into the benighting realm by considering the following question*: what if the agent commits wrongdoing but, at the time of acting (or omitting), fails to have right and outweighing reasons to refrain that are either explicit or consciously accessible?* Paralleling volitionism, I propose an epistemic condition on indirect blameworthiness for this kind of unwitting wrongdoing, according to which this condition is satisfied with respect to that wrongdoing if and only if the agent lacks right and outweighing motivating reasons to refrain from wrongdoing that are either explicit or consciously accessible, but the agent performed a benighting act (or omission) for which the agent was directly blameworthy—that is, an act that satisfied E-DB—but where among the motivating reasons *against* the benighting act was the fact that the benighting act would increase the likelihood of eventual wrongdoing, which the agent also had right and outweighing motivating reasons against. Although my account bears resemblance to other internalist tracing accounts, it makes original claims about precisely when blameworthiness should be traced and to what state it is to which we should trace.

Once I have concluded with a presentation and defence of my account of tracing, I will have put in place the last piece of the puzzle of my theory of the epistemic condition. How this constitutes an overall response to the Regress Argument and its revisionist implications (if it is not already evident) will be a matter for the Conclusion, next Chapter.

The plan for this Chapter is as follows. In §8.2, I give examples of the two relevant types of cases in which the agent does not have right and outweighing reasons to refrain that are explicit or deliberatively accessible. In §8.3, I outline my basic account of how the epistemic condition for culpable conduct can be satisfied in cases of these kinds, and I support my account by appeal to direct intuitions as well as to my accounts of blameworthiness, responsibility, and fair moral expectations. Finally, §8.4 is devoted to the question of whether the account requires that for an unwitting act to be indirectly culpable, the state of ignorance in or from which the agent acts must itself be culpable (as volitionists and other tracers require). There I argue that, in fact, ignorance need not be culpable for the conduct issuing from it to be indirectly culpable.

8.2 Cases of Wrong & Outweighed Reasons

There are two significant types of cases in which the agent fails, at the time of acting, to have right and outweighing reasons to refrain that are either explicit or consciously accessible through deliberative attunement ("consciously accessible" for short).

(I) Wrong Reasons: The agent has motivating reasons to refrain that are either explicit or consciously accessible, but they are not based on normative reasons why the act is wrong (i.e., there are no "wrong-sensitive reasons"; there are only "wrong reasons," that is, reasons of the kind that are not relevant to blameworthiness).

(II) Outweighed Reasons: The agent has wrong-sensitive motivating reasons to refrain that are either explicit or consciously accessible, but they are *outweighed* or at least neutralised by explicit or consciously accessible motivating reasons in favour of the act (i.e., there are only "outweighed reasons").

There are other possibilities, of course—for instance, the possibility that the agent has *no* motivating reasons to refrain, full stop[158]—but I shall confine my attention to the more likely cases of (I) and (II).

Let us consider examples of each, beginning with cases of (I), in which there are only "wrong reasons" to refrain. Recall that the "right" reasons which must tip the balance of one's reasons in favour of an alternative must be *wrong-sensitive* reasons, or reasons which are, or are based on, normative reasons why the act is wrong. To have reasons to refrain that are *not* these reasons is to have "wrong" reasons to refrain—"wrong" in the sense of not being reasons of the relevant kind to the charge of blameworthiness. And then we saw that these reasons must either be explicit (before the agent's mind in deliberation), or if not, consciously accessible through deliberative attunement. We discussed an example of wrong reasons in the last Chapter: Jack, in *Breaching a Liquor Ban*, is entirely oblivious to the features that make avoiding bringing the bottle of beer to a venue with a liquor ban wrong. In other words, there is no normative reason that makes not bringing the beer to such a venue *wrong* which *Jack takes to be a reason* not to bring it. The only motivating reason that Jack does have is the reason that it would be possibly stingy or mean, but I argued that this was not wrong-sensitive (in this case). There are other cases of wrong reasons. Many of the cases of factual ignorance that we have discussed count among them (e.g., Zimmerman's Perry, the case of the shipowner, Arthur Coningham's friendly fire, Applebaum's poisoning Botstein), for these are cases in which the ignorant agent has reasons to refrain but which do not include the morally pertinent wrong-making facts (the fact that one could paralyse someone who you are salvaging from a car accident, the ship's unseaworthiness, the fact that there are friendlies on board, the fact that the sugar jar has arsenic in it, etc.). Or consider some cases of moral ignorance. Imagine one of the Battalion 101 soldiers having as his sole motivating reason not to commit the shooting that it would make him nervous of getting hurt in retaliation. Or consider an American slaveholder, whose only reason not to beat his slave is that she is about to be sold at an auction, where it is custom to present slaves as well treated. The slaveholder would not regard as reasons not to beat her any of the facts that actually do make beating her wrong, such as the fact that it would needlessly harm her, that she is equal in moral standing, or that beating her is cruel. Finally, consider more complex cases. One set of cases arises from the possibility (entertained in §7.5 last Chapter) of having wrong-sensitive *beliefs or*

[158] E.g., cases in which one is mistakenly certain that the wrongdoing is *obligatory*, and one recognises no self-interested or prudential reason to refrain from that action.

credences but failing to recognise them as reasons, and so failing to have wrong-sensitive *motivating reasons.* This may, in fact, be the best way to explain much of past moral ignorance. Montmarquet's (1995, 46) seventeenth century Spanish priests, who took Mayan babies from their mothers, baptised them, and killed them in order to save their souls, probably had wrong-sensitive beliefs—for example, the beliefs that doing so would bring great sorrow to their mothers, or that these babies are humans—but failed to take them as reasons to avoid killing them. Another set of cases are cases in which the agent has motivating reasons that are *morally relevant* but which are not *wrong-making*; the motivating reasons might make an alternative morally *better* "within the realm of the morally permissible" (Harman 2016), but not morally *obligatory*. Imagine that just before committing an unfaithful sexual act, the only consciously accessible motivating reason possessed by the husband is his knowledge that his late grandmother would prefer not to see him act like that. Although this may be one relatively insignificant or "swamped" reason which contributes to why remaining faithful would be better than unfaithful sexual activity in the circumstances, it is not clear that it makes this unfaithful activity *wrong*. (The husband would have to be severely morally blind to not have any other accessible motivating reason to avoid infidelity.)

In all of these cases, the wrong-sensitive reasons being denied of these agents are either explicit or accessible through deliberative attunement. By contrast, some wrong-reasons cases *involve* the possession of implicit wrong-sensitive reasons, but the agent does not have deliberative attunement to the situation *qua* situation calling upon deliberation, and so their reasons are not *consciously accessible* (although they may be "accessible" in another, irrelevant sense).[159] The capacitarian cases that I discussed last Chapter, *Hot Dog* and *Forgetting the Milk*, are good examples, as well as the variant of *Forgetting My Car's Location* (from last Chapter) in which the crash at the bus station takes my attention away from whether to take the bus or my car home. In all of these cases, the agent becomes so distracted by something else that the "choice situation" to which they are deliberatively attuned changes. When Alessandra is caught in a tangled tale of misbehaviour and administrative bungling at the school, the choice situation to which she is deliberatively attuned requires making a choice between different ways of responding to the incident, not a choice of whether to remain in the school or return to save the dog. Thus, although she has a decisive motivating reason to return to the car, it is not accessible through deliberative

[159] Think of the sense in which the report of subordinates' war crimes *can be accessed* by the commander if only the commander was to pick it up (from §6.4).

attunement and so she cannot be directly blameworthy for her omission to return to the car in time.

There are, thus, many cases in which one fails to have right reasons that are explicit or consciously accessible through deliberative attunement. And these are cases in which I propose that if the agent's action or omission is culpable, its culpability must trace back to something for which the agent is directly blameworthy.

The other key family of cases involving the failure to have right and outweighing reasons to refrain is the family of type (II) cases involving the failure to have *outweighing* reasons, despite the presence of "right" reasons (that is, reasons of the right "wrong-sensitive" kind). We have already discussed a key example of this kind in *Abortion Dilemma*, where Nora's wrong-sensitive motivating reasons against the abortion are not outweighing. Even relative to higher-order beliefs or credences about what she ought to do under moral uncertainty, it would not be more morally conscientious for her to have the abortion than to carry the foetus to term. In this case, I argued that Nora cannot be directly blameworthy. I also argued similarly about cases of radical evaluative ignorance—for example, the case of Rosen's virus-affected Bonnie who elbows you into a curb believing that that is all-things-considered the thing to do, or Bill who believes lying to his wife is the all-things-considered thing to do. These are cases in which the agent commits wrongdoing whilst having wrong-sensitive reasons not to do so, but where the agent does not properly *appreciate* those reasons so that they are not *outweighing* for the agent. There are plenty of other cases of this kind too. Addicts have wrong-sensitive reasons not to engage in their illicit meth abuse, consumption of pornography, binge-drinking, or gambling of family finances. They often become "numbed" or "desensitised" to the wrongness of their habits such that their reasons do not weigh in favour of avoiding giving into them. (Of course, in many such cases, the agents still have right and outweighing reasons to avoid giving in, but their numbness or desensitisation makes these reasons fail to *be explicit* in their reasoning. These agents would not be excused on my account, however, if their reasons are nevertheless consciously accessible.) Many cases of moral ignorance also seem to involve a failure to have outweighing wrong-sensitive reasons. The Spanish priests may have taken as reasons not to kill the Mayan babies that they were loved by God or that they could be raised to be effective evangelists in South America, but took to be greater reasons the fact that baptising and having them killed would secure their salvation. Finally consider complex cases in which it appears that one's wrong-sensitive reasons are outweighing but where they are actually *outweighed*. The agent might irrationally judge that they have greater reason to avoid wrongdoing, when their reasons actually decide

in favour of committing the wrong act. In these cases, too, I propose that we must trace blameworthiness if we are to find them blameworthy.

8.3 The Epistemic Condition for Indirect Blameworthiness

Even though agents may fail to be directly or originally blameworthy for their wrongdoing, it is my view that they can be indirectly or derivatively blameworthy for it. How? The primary purpose of this Chapter is to defend the following account of the *epistemic condition* under which they can be indirectly or derivatively blameworthy (leaving aside any control or causal conditions). In particular, and to match E-DB, I propose the following:

E-IB:[160] An agent satisfies the epistemic condition on *indirect* blameworthiness for wrongdoing *x,* at time *t,* given wrongdoing *y* at an earlier time *t-1*, *if and only if,* (i) at *t,* the agent does not have right and outweighing motivating reasons to refrain from wrongdoing that are explicit or consciously accessible; (ii) at *t-1,* the agent has right and outweighing motivating reasons to refrain from performing *y* that are either explicit or consciously accessible; (iii) these motivating reasons include the fact (or something that for the agent entails) that *y* increases the likelihood of performing an action/omission of a relatively coarse-grained type, *W,* to which *x* belongs, and (iv) these motivating reasons include right and outweighing motivating reasons to refrain from performing actions/omission of type *W*.

We may illustrate E-IB by considering the case of Nora. Nora satisfies the epistemic condition on indirect blameworthiness for having the abortion (which we are assuming to be wrong for the purposes of illustration), only if having the abortion was the causal upshot of some wrong act at an earlier time—for example, intentionally indoctrinating herself with pro-abortion beliefs—and at *that* time, she had right and outweighing reasons not to indoctrinate herself that were explicit or consciously accessible; and they included the reason that doing so would increase the risk of later abortion (or simply of being "pro-abortion" in future conduct), and they also included right and outweighing motivating reasons to avoid the later action/omission (e.g., the reason that an abortion would violate a foetus' right to life).

[160] "E" stands for "epistemic condition" and "IB" stands for "indirect blameworthiness for wrongdoing."

There are two key differences from the volitionist account of tracing. First, on my proposal, we need not try to trace blameworthiness as soon as the agent lacks an occurrent belief in wrongdoing based on belief in wrong-making features, but as soon as the agent fails to have right and outweighing motivating reasons to refrain that are either explicit or consciously accessible. Thus, as I argued in the previous Chapter, even if the agent committed wrongdoing but lacked an occurrent belief in her wrongdoing based on belief in its wrong-making features (i.e., did not perform fully advertent wrongdoing), she may still count as directly blameworthy for wrongdoing if, for example, (a) features which, unbeknownst to the agent, make the act wrong, constitute outweighing motivating reasons for her to refrain which are explicit (i.e., *de re* reasons suffice for reasons to refrain); or (b) it is true relative to the agent's explicit motivating reasons that refraining is less morally risky (or more morally conscientious); or (c) the agent has right and outweighing reasons to refrain that are merely implicit but consciously accessible through deliberative attunement. Second, on my proposal, when the agent lacks right and outweighing motivating reasons to refrain that are explicit or consciously accessible, any culpability must come from a wrong act/omission from which the agent had right and outweighing motivating reasons to refrain that were explicit or consciously accessible, and not necessarily from an instance of fully advertent wrongdoing. Moreover, the *foresight* involved in (iii) and (iv) need not be occurrent but consciously accessible on E-IB.

Let me turn now to the justification of E-IB. Since condition (i) specifies the "indirectness" of the blameworthiness by definition, I start instead with the justification of condition (ii), the requirement that for the agent to be indirectly blameworthy for the eventual wrongdoing *x*, the agent must have had right and outweighing motivating reasons to refrain from doing the act *y* which led to *x*. This follows from three claims: indirect blameworthiness always traces back to an instance of direct blameworthiness (see §5.6's discussion of the failed strategy of grounding non-direct responsibility in counterfactual responsibility); indirect blameworthiness (and responsibility) always traces back to direct or original blameworthiness *for actions or omissions*; and direct blameworthiness for *any* act/omission, benighting or not, requires right and outweighing motivating reasons to avoid it (which are explicit or consciously accessible).[161] Now this last claim is a plausible combination of E-DB and the view (defended in part against epistemic vice theorists in §6.2) that benighting conduct should be treated in the same way as ordinary conduct more generally, where culpability

[161] I will hereafter omit this qualification for ease of expression.

ascriptions are concerned. And the claim that original blameworthiness is always blameworthiness for actions/omissions follows from my argument in Chapter Five (§5.6) that direct or original responses to reasons are exclusively *actions/omissions*. This was further bolstered by the appeal to deliberative attunement in the last Chapter as attunement to a *choice* situation, where choice situations relate only to actions/omissions (and secondarily any consequences thereof).

Condition (iii) holds that indirect blameworthiness for wrongdoing *x* requires that one's motivating reasons against the initial (benighting) act *y* include the fact (or an equivalent) that *y* raises the likelihood of performing an action/omission of a relatively coarse-grained type *W* to which *x* belongs. This condition involves *foresight* of *x* and so satisfies the *foresight requirement* (that there must be foresight of the consequence for which one is responsible, defended in §5.6). It may be wondered how E-IB factors in *culpable lack of foresight* (or "inadvertence to risk"; Nottelmann 2007), but this is built into E-DB: if someone committed a benighting act *y* while *lacking* foresight of future unwitting wrongdoing, they can be culpable for *y* only if the conditions in E-IB are satisfied *with respect to y*, where culpability for *y* is traceable to another benighting act/omission *z* and, at the time of performing *z*, (among the other conditions) one *had* foresight of performing an action/omission of type *W* to which the final wrongdoing *x* belonged. There is a detail about this reply that clashes with another internalist tracing view, due to C. Ginet (2000). On Ginet's view, the original source of blameworthiness need only involve foresight of *the next benighting act* in the culpability chain linked ultimately to the eventual unwitting wrongdoing, rather than foresight of the eventual unwitting wrongdoing itself (or its general type). My reply to this is threefold. One, insomuch as foresight need only be foresight of an action/omission of a relatively coarse-grained type (see the examples of moral ignorance discussed below), then Ginet's intuition is likely to be satisfied anyway.[162] Two, Ginet's view violates the plausible principle that indirect responsibility for any consequence requires foresight of that consequence (under a general description). And three, my account of an act/omission's being an indirect response to reasons requires a response to the reasons that determine the moral status (i.e., wrongfulness) of *that* act/omission.

[162] On Ginet's view, advertent benighting acts must involve foresight of "the sort of harm" caused by the next act in the culpability chain. But on the coarse-grained view of foresight defended below, it would be enough that the original benighting act involved foresight of an increased risk of "causing harm" in general, however distant that consequence is.

Three further questions must be asked about condition (iii). Why may the foresight be consciously accessible (this applies to condition (iv) as well)? Why must the foresight have the content that the initial action/omission *raises the likelihood* of later action/omission? And why does the foresight need to be of the increased likelihood of some action/omission only *of some relatively coarse-grained type W*?

Considering the first of these questions, the answer is that if I am right about the relevance to direct blameworthiness of implicit but consciously accessible motivating reasons, then (*pace* Zimmerman 1986, 206; Ginet 2000, 270, 275) I see no reason why foresight would require occurrent belief. Others (e.g., Fischer and Tognazzini 2009, 546; Timpe 2011, 20-21) feel similarly. If a doctor only *implicitly* foresaw that not reading the new research would mean providing a possibly risky prescription in the appointment next week, she may be indirectly blameworthy for eventually giving this prescription if there was a moment during the week—for example, when hearing the announcement of new research—in which she became deliberatively attuned to her situation as requiring a choice about whether to update herself with the new research or (e.g.) watch the television with her partner. At that moment, it did not *occur* to her that she should update herself in order to prepare for her appointment in the following week, but it was certainly the case that she *would have given that as a reason* for updating herself if asked for her reasons. Thus, the relevant foresight need not be occurrent, but can be implicit—as long as it is consciously accessible through deliberative attunement.

I turn now to the claim that indirect blameworthiness for wrongdoing (or for any consequence) requires foresight of an *increased likelihood* of that wrongdoing. This claim is commonly found in tracing theories (Nottelmann 2007, 191ff.; Nelkin and Rickless 2017, 120; Timpe 2011, 14-15; Peels 2017, 177). I believe that this is for a reason given by Nottelmann. To take his example, suppose that adding some "cheapening ingredient" into canned soup tragically results in a consumer's loss of life. Now compare the competing intuitions of responsibility between a variant of the case in which you do not foresee that adding the ingredient increases the risk of anyone being poisoned by the soup, and a variant in which you *do* foresee an increase in that risk, even if it is "a very tiny increase (say 0.00001) in risk" (p. 177).[163] Assuming the fulfilment of the other responsibility conditions, it seems clear to me that in the latter case you are blameworthy (if only minimally) for the eventual death, but you are *blameless* in the former case in which you do not foresee an

[163] Nottelmann makes this point with respect to another (more far-fetched) case.

increased risk. Moreover, you would remain excused even if you believed that there was some remote chance of poisoning someone with the added ingredient as long as you believed that the risk was equal or greater *without* the added ingredient.[164] To be responsible for the harm, therefore, you must have foreseen an *increased* risk of causing harm in performing the initial act.

The reason why I add in parenthesis that the foresight can be foresight of something that *for the agent entails* an increased risk, is that some people would not have exactly *that* content included in their motivating reasons (see this also in my account of blame, B, in §5.3). However, if they believed that an initial action/omission "might" (Ginet 200) lead to the foreseen action/omission, and if they *would* endorse an increased risk—and these facts constituted a motivating reason for them—then that would satisfy the increased risk condition.

I turn now to the justification for holding that foresight of the eventual action/omission need only be foresight of performing an action/omission under a *relatively coarse-grained* type-description. This is a controversial matter. I side with Fischer and Tognazzini (2009), King (2017, 272), Nelkin and Rickless (2017a, 126), and Nottelmann (2007), who believe that foresight of a consequence need not involve highly fine-grained content. On the other side, Zimmerman (1986, 206), Vargas (2005, 275-6), and Timpe (2011, 16-17) think that the foresight must be quite fine-grained. Compare foresight of "telling a racist joke" with foresight of "doing a racist thing"; or foresight of "killing Mayan babies on the next expedition" with foresight of "killing infants. The matter at hand is whether foresight must be relatively fine-grained for indirect blameworthiness, and I think there is good reason to deny that it need be. For instance, it seems intuitive to me that a teacher can be indirectly blameworthy for coming up with the wrong answer to a highly important question raised by a student, if the teacher's answer was the upshot of a failure to prepare for class despite recognizing the need to be well-prepared. To be rightly held responsible, the teacher need not have foreseen the specific question to which she gave a wrong answer, or even have foreseen responding wrongly to a student's question. She need only have foreseen the risk of being useless to the students as a consequence of not preparing. I take this to be supported by a consideration of fair expectations. *Given* that the teacher believes that she should be well-

[164] Nottelmann raises this objection against Zimmerman's (1986, 206) account, which only requires foresight of "some probability" that a risk will obtain.

prepared for the tutorials, it is fair to expect her to prepare for the tutorial such that she becomes able to answer any highly important course-related question that will come her way.

Appeal to relatively coarse-grained foresight also helps to explain how morally ignorant wrongdoers, who nevertheless feel subjectively *certain* of their conduct's moral permissibility, could still be indirectly blameworthy for their wrongdoing. Consider the Battalion 101 officer, the American slaveholder, and the Spanish priest, who do not have right and outweighing reasons (that are explicit or consciously accessible) to refrain from their wrongdoing. There may well have been a moment in the past when these agents could have foreseen that they were taking the risk of one day treating Jewish people, African slaves, or Mayan infants "unjustly," "inhumanely," "immorally" or "unbiblically." This would have been foresight with very general content, and my claim is that it could have given them a motivating reason to take relevant precautions against future wrongdoing by inquiring into the nature of just (moral, biblical, etc.) treatment of Jews, African Slaves, or Mayan infants. In the end, of course, what actually constituted just treatment of these groups was completely different from what these historical wrongdoers eventually thought amounted to just treatment. But the point here is that they might well have foreseen the risk of unjust treatment of these groups but under a description general enough for them to have right and outweighing reasons against *that.* This reveals the possibility that morally ignorant wrongdoers may sometimes (perhaps often) start from a place in which, at a certain level of generality, they are not morally ignorant.

Let us turn now to the final condition (iv) of E-IB, the claim that the motivating reasons against the initial (benighting) act *y* include *right and outweighing motivating reasons to refrain from* performing the general type of act/omission *W* to which *x* belongs. Together with (iii), condition (iv) satisfies the *reasons-transitivity requirement* (that the normative reasons against the eventual wrongdoing must be normative reasons against the benighting act, defended in §5.6). This is secured by the fact that, for blameworthiness, the agent must have *right* (i.e., wrong-sensitive) motivating reasons to avoid foreseen wrongdoing which constitute *wrong-sensitive* motivating reasons to avoid the benighting act. *Wrong-sensitive* motivating reasons by definition map on to the normative reasons at issue in the reasons-transitivity requirement.

Now, the more general claim that there must be foresight of the moral significance of the eventual (wrong) act is plausible and widely accepted (cf. Ginet 2000, 273-5; Timpe 2011, 20-21; Vargas 2005; Fischer and Tognazzini 2009, 575). The reason for this requirement is straightforward. Suppose that you believed that by going to a series of white supremist rallies,

you would become a white supremacist yourself and would in turn promote white supremacism. But suppose (tragically) that this is what you wanted and that you were *sure* that this was the right thing to do. Then, in the absence of indirect culpability for *going to the rallies*, it hardly seems that you are indirectly blameworthy for your resulting white supremacist beliefs and actions. How then should we characterise foresight of moral significance? Well, clearly, in light of my reasons-respondence theory of responsibility, foresight of moral significance consists first and foremost in foresight of the *normative reasons against the foreseen act/omission*. But if so, then it would be most natural to apply the account of awareness of normative reasons against one's *present* conduct (which I gave last Chapter) to foresight of normative reasons against a foreseen action/omission *in the future*, under a relevant general description. That is to say, the agent must have *right and outweighing motivating reasons to avoid* the performance of a foreseen action/omission, under a general description. Other internalist tracers apply their accounts of awareness of moral significance to foresight of the moral significance of future actions/omissions in this way (e.g., Ginet 2000, 275; Timpe 2011, 20-21; Nottelmann 2007, 196-7). Note also that the same considerations that led me to my proposal about the epistemic condition on direct blameworthiness last Chapter crop up again in cases involving foresight of normative reasons against the foreseen action/omission. For example, there are cases in which you are plausibly indirectly blameworthy for some unwitting wrongdoing, because although you did not foresee it as "wrong" while performing an earlier benighting act, you still foresaw it as "unkind," "harmful," "reckless," "unjust" (to cite wrong-making features), or as "morally risky"—and these gave you right and outweighing motivating reasons to avoid risking that sort of action/omission. And to ram home the intuition behind the need for *outweighing* reasons, suppose that in every situation in which Nora foresaw the increased likelihood of having an abortion as a consequence of some benighting act/omission that she performed (e.g., not reading a book on the ethics of abortion for class), she was just as morally uncertain about the permissibility of abortion as she ended up being on the day that she actually had to make the choice about whether to abort the child. In other words, imagine that there was never any improvement in her epistemic position with respect to the (stipulated) wrongfulness of abortion. Then, it seems clear to me that she could not have been indirectly blameworthy for aborting the child, even if she had right and outweighing motivating reasons to refrain from her benighting action/omission (e.g., to read the book that she was required to read for class).

E-IB therefore appears to stand up to scrutiny. We might now ask how it applies to the capacitarian cases of *On the Rocks* and *Secret Service*, accounts of which I have promised. *On the Rocks* is the case in which the ferry pilot, Julian, gets distracted by thoughts about his romantic encounter the night before, and accidentally crashes a ferry full of passengers into rocks. *Secret Service* is the case in which a member of the president's security detail, Alvaro, gets distracted by a text message from his son while escorting the president through a big crowd, and consequently fails to prevent an attack on the president. In Chapter Six, I argued that these cases should be treated as cases of indirect blameworthiness, because, at the time of the agents' omissions (i.e., Julian's failure to avoid the rocks, and Alvaro's failure to prevent the lunge on the president) they are *not* aware of the relevant wrong-making features of these omissions. But this does not let them off the hook. They are *indirectly* blameworthy for these omissions due to culpably violating, moments before, a standing obligation to pay the proper attention to their tasks. They are indirectly blameworthy because they are directly blameworthy for letting themselves be distracted in the first place, and at *that* time (moments before) they were in a situation requiring a choice between keeping their focus and letting themselves be distracted, a situation to which they were deliberatively attuned (even if it did not occur to them at that very time as a situation requiring deliberation). Moreover, at this earlier time, they continued to have right and outweighing (decisive) motivating reasons to keep their attention on what the job required of them (i.e., on piloting the ferry and on protecting the president) relating to the harm that might otherwise be caused to the people concerned (the ferry passengers and the president).

Thus, I analyse the blameworthiness of these cases in much the same way as I analyse the blameworthiness of the two other allegedly non-tracing capacitarian cases discussed—*Hot Dog* and *Forgetting the Milk*. The difference, however, is that although Alessandra's and Randy's blameworthiness traces back to an initial act/omission (i.e., leaving the car with the dog in it, and failing to prevent distraction by philosophical musings), the reason is not that these agents do not have decisive motivating reasons to avoid their wrongful omissions (to return to the car or buy the milk) at the time of these omissions but that they do not have *deliberative attunement* at the time of these omissions (as I argued last Chapter). Plausibly when they commit the benighting act/omission, they were, however, deliberatively attuned to a choice situation, and they also foresaw, albeit implicitly, the increased risk of later omission (which they had right and outweighing motivating reasons to avoid). That is how I explain the intuition that Alessandra and Randy are blameworthy. Thus, in the end, the four apparently non-tracing cases touted in favour of capacitarianism really are *tracing* cases, but

cases which favour a weakened internalist locus of original blameworthiness not long before the failure to notice.

The weakened internalists, Dana Nelkin and Samuel Rickless (2017, 119-24), have offered uniquely sophisticated account of tracing, and account for these cases (except for *On the Rocks)* by grounding the agents' blameworthiness for their unwitting omissions in the moments of awareness to which I am appealing—namely, when Alessandra has a choice between leaving the dog in the car and taking the dog, or when Alvaro has a choice between reading the son's text message and keeping the focus on the crowd. On their so-called "Opportunity Tracing" view:

> whether an agent is morally responsible for an unwitting omission at time T2 depends entirely and solely on whether there was a prior time, T1, at which the agent had the opportunity to do something that, as she reasonably believed, would significantly raise the likelihood of avoiding later omission. (Nelkin and Rickless 2017, 120)

Such an "opportunity" is provided by a moment of control which Nelkin and Rickless take to obtain when the agent has the "interference-free ability to do the right thing for the right reasons" (p. 118). It is important on their view that this moment need not involve a "conscious decision" to take the relevant risk, as volitionists seem to require (cf. Zimmerman 1997b, 421-2; Levy 2009, 736, n. 16). That there need not be a conscious decision is required after all to account for Alessandra's and Alvaro's blameworthiness, since they do not *decide* to assume the risk of omitting to save the dog or the president. It is sufficient for Nelkin and Rickless that the agent has the ability in question and the reasonable belief that taking the relevant opportunity would significantly increase the likelihood of avoiding a later (wrongful) omission.

My view has lots in common with this view. Both are couched in a broadly reasons-responsive theory of responsibility. Both ground responsibility for omissions while lacking the opportunity (in my view, the deliberative attunement) to avoid them indirectly in a moment of reasons-responsive control when one had an opportunity to prevent them. Both do not require that the benighting action/omission is a conscious decision. Nelkin and Rickless also appear to allow that the foresight need not be of the wrongful status as such of the foreseen act but only of its wrong-making features (e.g., "disregarding others' interests later," p. 126). Finally, both views require foresight of the unwitting omission—at least, as they say, foresight "in broad strokes" (Nelkin and Rickless 2017, n. 7, 126).

Aside from having a slightly different foundation from my account, there are two issues with the *constituents* of the foresight and an issue with its content. The first issue concerning the constituent of the foresight is that they require for indirect responsibility a *reasonable* belief in an increased risk of later omission. But this seems to require too much, for the same reason that *mere true belief* (when amounting to a wrong-sensitive motivating reason to avoid the benighting act) is sometimes still not enough to excuse a wrongdoer from direct blameworthiness (§3.4.3). That is, the foresight need only consist in a true belief (or non-negligible credence), constituting one of the agent's motivating reasons. Imagine that Tim swallows an obscure stimulant despite the true belief that the stimulant has the (unlikely) side-effect of an increased risk of being aggressive (a fact that he takes to be reason against swallowing it). But imagine that Tim acquired this belief from an unreliable source on the internet, known for its misinformation—indeed a source that does not explain why or how the stimulant has that side-effect. The true belief that he forms is consequently unreasonable. But surely, if Tim were then to be violent under its influence, he could hardly be let off the hook on the grounds that his foresight of this risk was nevertheless *unreasonable*. The second issue about the constituent of the foresight is that it appears that Nelkin and Rickless require *occurrent* foresight for indirect blameworthiness (even though they remain silent on the occurrent/dispositional or implicit/explicit distinction). When discussing *Forgetting the Milk*, they hold that responsibility traces back to the moment when Randy is "[aware] of the fact that his mind is beginning to stray from its present focus (on driving to the store to purchase the milk)" (p. 121)—a moment which seems to involve occurrent awareness of one's task and indeed occurrent *meta*-awareness. But we have already given reasons for denying that the foresight need be occurrent; it can be implicit as long as it is consciously accessible through deliberative attunement (see the doctor case above).

The final issue is about the content of the foresight: although Nelkin and Rickless appear to allow that the foresight can be merely of the foreseen omission's wrong-making features, they do not offer a complete account of foresight *of moral significance*. Notice after all that their Opportunity Tracing view does not itself specify the moral significance of the foreseen omission, despite indirect responsibility being characterised as "depending solely and entirely" on the opportunity involving the foresight merely of "the omission" (with no mention of its moral significance). Charitably, though, they argue that moral responsibility is a function of being able doing the right thing *for the right reasons*, and their cases all seem to involve foresight of wrong-making features, so it would not be far cry for them to add the further condition that the agent must have the right motivating reasons to avoid or prevent

later omission. Of course, these reasons to avoid the omission must also be *outweighing*, as I have shown. That concludes my defence of E-IB.

8.4 Culpable Ignorance?

Where does this leave us regarding culpability for ignorance as a state? Blameworthiness in E-DB and E-IB is only blameworthiness for actions or omissions. But recall that volitionists demand that for a person to be blameworthy for acting wrongly in/from ignorance, the ignorance must itself be culpable; indeed, culpability for the ignorance is part of their *explanation* of blameworthiness for unwitting conduct. Others, as we saw in Chapter Six (Harman 2011; Rudy-Hiller 2017), also require that the ignorance must itself be culpable, but hold that the culpable ignorance is not what *explains* the culpability of the conduct issuing from it (rather, the culpability of both the state of ignorance and the issuing action has a common explanation). And still others (e.g., Clarke 2014) allow that unwitting wrongdoing can be culpable without requiring that the ignorance itself is culpable. Where do I stand on this issue? Note that, on my view, it makes most sense to identify the relevant ignorance as the lack of right and outweighing motivating reasons to refrain that are explicit or consciously accessible.

The first thing to say is that I agree with Harman and Rudy-Hiller: even if the relevant ignorance must be blameworthy, I do not think that culpability for the ignorance is what *explains* culpability for wrongdoing in/from it. Blameworthiness for wrongdoing, recall, is the quality of being morally at fault for wrongdoing. What matters for the act's blameworthiness is that the agent has acted in a morally faulty way in the lead up to the act, and that is a matter of responding poorly to the relevant normative reasons against it (see Chapter Five). This clearly "bypasses" the relevant ignorance that arises from one's culpable agency, and grounds the relevant explanatory relations in the agent's morally faulty exercise of her role as respondent-to-reasons at an earlier time. Now it is plausible that, when it comes to the state of ignorance—in the same way as for other mental states or attitudes such as beliefs or desires—blameworthiness for being in that state may be grounded in the morally objectionable role that the agent played in its formation or retention. If one's ignorance in the circumstances is "one's own moral fault" then it seems plausible that it is blameworthy. But how would it be the agent's own moral fault? The answer lies in our account of what can be

direct or indirect responses to reasons. In §5.6 I argued that attitudes or states of affairs at a given time *t* cannot have a moral status arising from their being, as such and at that time, our responding to reasons, since we do not at *t* have *direct* causal control over our *being* in these states. Rather, blameworthiness for ignorance is a function of responding poorly to reasons to avoid *becoming* ignorant. And, for an agent to be blameworthy for becoming ignorant, not only should the ordinary tracing conditions be met (e.g., counterfactual dependence and explanation), but also the foresight and reasons-transitivity requirements (adapted so that the foresight is of the *ignorance* itself and the reasons against *it*). Therefore, I propose an epistemic condition on blameworthiness for ignorance along the following lines (structurally equivalent with E-IB):

E-BIg: An agent who performs unwitting wrongdoing, at time *t*, satisfies the epistemic condition on blameworthiness for their state of ignorance *i* (of the act's wrong-making features and of their weight), given some earlier (wrong) benighting act/omission *y* at some earlier time *t-1, if and only if* (i) at *t-1*, the agent has right and outweighing motivating reasons against *y* that are either explicit or consciously accessible, (ii) these motivating reasons include the fact that *y* increases the likelihood of forming or sustaining ignorance *I* of the general type to which *i* belongs, and (iii) these motivating reasons include right and outweighing motivating reasons against the formation or retention of ignorance of type *I*.

Clearly the normative reasons against being in a state of ignorance envisioned here are reasons against being in a state, rather than reasons for or against *action*, but it is not clear that this makes any difference as far as responsibility and blameworthiness is concerned. It must only be the case that the normative reasons against being ignorant count as normative reasons against benighting action. It is common, anyway, to find treatments of both kinds of reasons in the same context (cf. Scanlon 1998, 17-22).

Note that E-BIg itself does not yield the *requirement* that, in order for unwitting wrongdoing to be blameworthy, the issuing ignorance must itself be blameworthy (as volitionists typically assume). In order to establish such a requirement, it would have to be the case that in *every* case of foreseeing the unwitting wrongdoing and the (normative) reasons against it (constituting right and outweighing motivating reasons for the agent), the agent also foresees the ignorance in/from which she might act, and the reasons against letting herself be ignorant in that way. Now this certainly often happens. The doctor foresees the risk of giving the patient the wrong prescription or a botched blood transfusion as a consequence

of not checking the patients details now *and thereby becoming or remaining ignorant* of the patient's details. But the question is whether foresight of the ignorance and the reasons against the ignorance must *always* accompany foresight of the unwitting act and the reasons against the act.

I think these need not always go together. Consider the following case:

Your Grandmother's Warning

Suppose that your wise and loving grandmother, Lynda, gives you a stern warning not to surround yourself with a certain group of jerks. All she says is that they will "rub off" on you. Although she does not put it this way to you, she thinks that habitually keeping company with those people will increase the risk of future wrongdoing precisely *because* it will increase the risk of your becoming morally ignorant. You never hear precisely how or why they will "rub off" on you, but you think to yourself that this must be because their company may dispose you to doing wrong things in the future. Because you respect her judgment on moral matters, you form the belief that by continuing to hang out with the group, you will risk later wrongdoing—and this gives you a right and outweighing motivating reason to avoid spending time with the group. Crucially, you do not form any belief that, by doing so, you will risk losing *a sense for what is right or wrong*. In fact, you assure yourself that you will always know the difference between right and wrong, and that surrounding yourself in their company will only make you more disposed to form a *desire* to do bad things or to be susceptible to peer pressure (but you think that you will probably remain steadfast enough to resist their influence). Nevertheless, suppose that you decide to hang out with the jerks anyway (because you want to be "cool"), and that you go on to become morally ignorant and an unwitting wrongdoer (e.g., wantonly firing employees).[165]

Are you blameworthy (indirectly) for your morally unwitting wrongdoing? I think that you are, for the following reasons. You are directly blameworthy for the (in fact, benighting) act of joining the group. At that time, you had right and outweighing motivating reasons not to join the group. You also foresaw that by doing so, you would increase the risk of later wrongdoing, which you had right and outweighing motivating reasons to avoid (that were explicit or accessible through deliberative attunement). Nevertheless, what is crucial is that

[165] This is an adaption of a case originally from Vargas (2005), which is treated as an example of tracing in Fischer and Tognazzini (2009)

you did not foresee becoming morally ignorant as a consequence of joining the group. As a matter of fact, you expressly denied this possibility, and had another explanation for how this group might influence you in the wrong direction. This case therefore reveals that foresight of future wrongdoing and recognition of it as such in consequence of a (benighting) act or omission, does not entail foresight of future moral ignorance. But if that is so, and if my story about what does explain indirect blameworthiness for unwitting conduct is correct, then it follows that blameworthiness for unwitting acts does not require blameworthiness for the ignorance in or from which the agent acts. Indirect blameworthiness for unwitting wrongdoing is *often accompanied* by blameworthiness for the ignorance, but the key point here is that it need not be. Culpable unwitting wrongdoing need not be *culpably unwitting* wrongdoing. R. Clarke (2014) is right to accept this claim. However, I have given a *tracing* explanation for why we should accept this claim (which appears unnoticed by internalist tracers). The focus on *culpable ignorance* therefore leads many theorists of the epistemic condition astray.

8.5 Conclusion: Combining the Epistemic Conditions for Direct and Indirect Blameworthiness

I have now put in place the last piece of the puzzle of my overall account of the epistemic condition on blameworthiness for actions and omissions. In the last two Chapters I have defended the following overall account of the epistemic condition, combining the epistemic condition on both direct and indirect blameworthiness for wrongdoing disjunctively:

E-B:[166] An agent satisfies the epistemic condition on blameworthiness for wrongdoing *x* in a certain choice situation C *if and only if, either*, in C, she has right and outweighing motivating reasons to avoid *x* that are explicit or accessible through deliberative attunement; *or* there was some wrongdoing *y* in some earlier choice situation D, when she had right and outweighing reasons to avoid *y* that were explicit or accessible through deliberative attunement, and these reasons included the fact that *y* would increase the risk of performing an action/omission of the relatively coarse-grained type *W* to which *x* belongs, and included right and outweighing motivating reasons against performing an action/omission of type *W*.

[166] "E" stands for "the epistemic condition" and "B" stands for "blameworthiness for wrongdoing"

In my next, concluding, Chapter, I shall consider how this account stands in relation to the volitionist's Regress Argument and its revisionist implications.

Chapter 9
Conclusion

9.1 Introduction

All my cards are now on the table. In the last two Chapters I have set out and defended my accounts of the epistemic conditions on direct and indirect blameworthiness for wrongdoing (E-DB and E-IB, combined as E-B). Clearly, a critical question for my purposes is how E-B constitutes a reply to the volitionist's Regress Argument and to its responsibility revisionist implications. Indications will have been picked up already about how my overall account constitutes a reply to the volitionist, but these need to be made explicit. I will attempt that in §9.2 and §9.3. I will then turn my attention to whether a weaker form of responsibility revisionism may be incurred by my account (in §9.3) and conclude (in §9.4).

9.2 A Response to the Regress Argument

Consider, again, the Regress Argument:

(1) An agent S is blameworthy for performing all-things-considered wrongdoing *A* only if S does *A* contrary to S's true occurrent belief that *A* is wrong all-things-considered, based on true occurrent beliefs that *A* has the features that make it wrong (i.e., *A* is "fully advertent" wrongdoing), or S does *A* in/from culpable ignorance of *A*'s all-things-considered wrongfulness (i.e., *A* is "culpably unwitting" wrongdoing).
(2) S is culpable for the ignorance in/from which S does *A* only if S's ignorance is the foreseen upshot of performing a "benighting" act or omission *B* for which S is blameworthy.
(3) S is blameworthy for performing a benighting act *B* only if *B* is either fully advertent or culpably unwitting wrongdoing.

Therefore,

(4) An agent S is blameworthy for performing all-things-considered wrongdoing *A* only if *A* is fully advertent wrongdoing, or the ignorance in/from which S does *A* is the foreseeable upshot, ultimately, of fully advertent wrongdoing (*B,* or *C,* or *D,* etc.).

Against Premises (1) and (3), I have argued that when the agent lacks an occurrent true belief in the all-things-considered wrongfulness of an action/omission (whether ordinary or benighting), an agent can still be *directly* blameworthy for it—that is, blameworthy in a way that does not depend on culpability for the lack of an occurrent true belief in the wrongfulness of the action/omission. This is because, to satisfy the epistemic condition on direct blameworthiness (E-DB), the agent at the time of acting need only have right and outweighing motivating reasons to refrain from their wrongdoing which, if not explicit (or occurrent), are at least accessible by virtue of deliberative attunement (i.e., "consciously accessible"). Indeed, there are many more cases that fit this description than there are cases of acting contrary to a true occurrent belief that the act is wrong all-things-considered (i.e., cases of "fully advertent" wrongdoing"). Wrongdoing can be directly blameworthy in the following different ways, among others: the agent has a true belief in the wrongfulness of what she does which is accessible only in virtue of deliberative attunement; the agent has no belief that what she does is wrong as such, only beliefs in wrong-making features which constitute outweighing motivating reasons for the agent against doing the act (whether explicit or consciously accessible); or the agent is uncertain about whether the action/omission is wrong or has a wrong-making feature, but this constitutes an outweighing motivating reason for them to refrain from wrongdoing (whether explicit or consciously accessible). I have argued, in other words, that directly culpable wrongdoing can be merely *partially advertent*.

Premise (2) is acceptable on my view, as long as "foreseen" is treated as synonymous with "foreseen" in the sense specified in E-IB. But an implication of what I argued in §8.4 is that Premise (2) is irrelevant, even if "ignorance" refers to the way that I define it—as the lack of right and outweighing motivating reasons to refrain from wrongdoing that are explicit or consciously accessible. There, I argued that wrongdoing in or from this state does not require that the state itself is culpable for the wrongdoing to be culpable. What matters for indirect blameworthiness for wrongdoing in or from this state (i.e., "unwitting wrongdoing") is that an action/omission of a relatively coarse-grained type *W* to which the unwitting wrongdoing would belong was the foreseen upshot of benighting wrongdoing, for which the agent was directly blameworthy (an action/omission that itself satisfied E-DB). And according to E-IB, this foresight must have involved foresight of an *increased risk* of an action/omission of type

W (in consequence of one's benighting conduct), as well as right and outweighing motivating reasons against that foreseen wrongdoing. Against volitionist analyses of this foresight condition, E-IB relaxes Zimmerman's (1986) and Ginet's (2000) requirement of *occurrent* foresight, Zimmerman's requirement of foresight of a *fine-grained* type to which the unwitting wrongdoing would belong, and the requirement that the foresight be of *wrongdoing* as such (rather than merely its wrong-making features).

9.3 A Response to Responsibility Revisionism

What follows for the two revisionists options promoted by volitionists? Recall the two options here (from §4.2 and §4.5):

Revisionism Option #1: The Rarity of Culpable Ignorance

(5) Culpable ignorance is *rarer* than many think. [partially from (2) & (3)]

Therefore,

(6) Blameworthiness for actions/omissions is rarer than many think. [from (1)-(3), & (5)]

Revisionism Option #2: The Inaccessibility of Fully Advertent Wrongdoing

(7) It is extremely difficult to ascertain whether someone has committed fully advertent wrongdoing.

Therefore,

(8) Blameworthiness for actions/omissions is extremely difficult to ascertain. [from (4) & (7)]

(9) Many think that blameworthiness for actions/omissions is relatively easy to ascertain.

Therefore,

(10) Blameworthiness for action/omissions is harder to ascertain than many think. [from (8) & (9)]

Clearly, (5) and (7) are now strictly irrelevant, given that, on E-IB, an agent may be blameworthy for acting wrongly out of ignorance without that ignorance being culpable (i.e., requiring that it was the foreseen upshot of at least partially advertent benighting wrongdoing). But we might wonder whether my account of the epistemic condition still incurs a weaker form of revisionism about responsibility and blameworthiness. To answer

this question let us substitute "culpable ignorance" as understood by volitionists using the Regress Argument, for what plays the equivalent role in my theory—being the foreseen upshot *of culpable benighting misconduct* (i.e., being "indirectly blameworthy")—and replace "acting contrary to the occurrent true belief that the act is wrong all-things-considered, based on features which make it wrong" with "acting contrary to right and outweighing motivating reasons to refrain that are either explicit or at least consciously accessible."

How, to begin with, are we to assess whether *indirectly culpable wrongdoing* is rarer than many think? It is extremely difficult to say. I should stress, however, that there are certain factors that make indirectly culpable misconduct far more likely on E-IB than what counts as culpable ignorance for the volitionist. E-IB requires only that the benighting misconduct is "partially advertent" in the sense of being contrary to right and outweighing motivating reasons to refrain that are either explicit or at least accessible through deliberative attunement. E-IB does not require *occurrent* beliefs, *de dicto* moral beliefs in wrongdoing, or *decisive* motivating reasons; the agent's motivating reasons can be purely *dispositional*, *de re*, and *non-decisive* (albeit outweighing). (In addition, benighting acts need not be the outcome of "conscious decisions.") In further support of E-IB's leniency, bring to mind the four "capacitarian" cases for which volitionists cannot account and generalise from them to other similar "failure-to-notice" cases. Julian is blameworthy for crashing the ferry into the rocks because there was a moment of time before this happens when he was deliberatively attuned to the situation as one requiring a choice between maintaining focus on the task of ferry piloting or letting one's mind drift—and *at that time*, he had right and outweighing reasons against letting his mind drift while foreseeing the risk of doing so (if only implicitly). And the fact that E-IB requires only a very general form of foresight—foresight of wrong actions and the reasons for their wrongness under a relatively coarse-grained description—seems to capture the blameworthiness in some cases of moral ignorance that volitionists struggle to capture. This is significant, given the strength of the pre-theoretic intuition in favour of the blameworthiness of morally ignorant wrongdoers. Recall my observation in §8.3 that morally ignorant wrongdoers, such as the Battalion 101 officers, American slaveholders, or Spanish priests, were not necessarily *always* morally ignorant. These agents could well have foreseen, at an earlier time, the genuine risk of future wrongdoing incurred by certain of their actions and patterns of behaviour at that earlier time, at least under a general level of description (e.g., treating Jews unjustly, Africans unfairly, Mayan infants unbiblically).

But whatever we say about whether E-IB entails that indirect blameworthiness for wrongdoing is *still* rarer than many think, it is unclear that the inference from this and E-DB should be that blameworthy conduct *in general* (whether direct or indirect) is rarer than many think. The reason is given above: that my account's conditions on *directly* blameworthy wrongdoing are much more lenient than the volitionist's conditions. To justify blameworthiness for unwitting wrongdoing, we do not always need to trace as soon as the agent lacks an occurrent belief in wrongdoing; the agent may still be directly blameworthy despite this lack (see above and Chapter Seven). In conclusion, it is not clear to me that E-IB and E-DB entail that agents are rarely blameworthy, or even less blameworthy than most think.

Of course, "many" (real-self theorists, quality of will theorists, and some other culpability externalists) will want to pronounce agents blameworthy when they are off the hook on my account, and so there is a weak sense in which my account holds that agents are less blameworthy than "many" think. But the claim that my view is still too stringent is to be expected. I never set out to capture *all* cases of alleged blameworthiness. Neither did I really set out to save common-sense intuitions about blameworthiness, even though I contended (in §4.6) that a theory's doing so would be *one* consideration counting in its favour. Attending to my initial reaction of surprise and doubt towards volitionism, I set out to question why volitionists drew the line between blameworthy and blameless wrongdoing exactly where they did, and I did so by consulting intuitions about individual cases—informed, in part, by background intuitions about blame, blameworthiness, and responsibility. This is how I arrived at my final account. Along the way, however, I found certain intuitions about the blameworthiness of cases unrecognised by volitionists to be irresistibly strong (e.g., the intuition from *Burning the House Down*), and so I had to shape my theory accordingly. I consider it a bonus, then, that my theory captures many more common-sense intuitions of blameworthiness than volitionism does. (Moreover, and in response to the quality of will theorist, I also gave some reasons in Chapter Six for thinking that the quality of will requirement is itself too stringent given cases of not acting fast enough to avoid wrongful omission.)

What about Rosen's sceptical revisionism? Is it not still the case that it is *hard to tell* whether someone has acted contrary to right and outweighing motivating reasons to refrain that are, if not explicit, at least accessible through deliberative attunement? Of course, any kind of culpability internalism will invite Rosen's claim that "the opacity… of other minds and even of one's own mind" suggests that it is "*unreasonable to repose much confidence in*

any particular positive judgment of responsibility" (2004, 308, his emphasis). And recall Rosen's point that what makes fully advertent wrongdoing difficult to determine is that it is easily confused with "ordinary weakness of will," where although you might "initially" think that the act is wrong, you end up judging that it is just as reasonable as the alternatives.

In response, and concerning judgments of one's own responsibility for wrongdoing, I have several points to make. First, whether or not something counts for you as a reason against some option is not hard to determine: just ask yourself what your reasons are, why you should act one way rather than another, or how you would justify your actions (and have justified similar actions) to others. Second, it is not as easy, but neither is it unreasonably hard, to determine the weight of your reasons. Ask yourself which way you are "leaning," where you "stand" at the moment, or what you would do if you had to make a decision right now. Alternatively, reflect on why you find yourself believing that you *ought* to act in one way rather than another, or that it would "probably be good" to avoid some act. This should reveal whether or not your reasons for an alternative are outweighing.[167] Third, sometimes your overall judgment (e.g., about your reasons being counter-balanced) can be mistaken, and further deliberation—while already deliberatively attuned—would bring out the fact that they clearly weigh in favour of an alternative, or at least would reveal a higher-order principle to which you are committed or towards which you lean, concerning what to do *when* you are morally uncertain (e.g., "do the least morally risky thing) which may thereby tip the balance of your reasons overall. Fourth, initially judging that you should not act before muddying the water by constructing a rationalisation in favour of doing the action is, I think, fertile grounds for culpability. E-DB might snag rationalised wrongdoing of this kind on the grounds that you still had right and outweighing motivating reasons to act during deliberation. It is just that something went wrong in the process, and your right and outweighing reasons wound up relegated from explicit reasons to reasons that were merely accessible reasons in virtue of your continued deliberative attunement. Nevertheless, as we have seen, this would not have been enough to acquit you. (And even if E-DB failed to snag some cases of rationalised wrongdoing, E-IB could still snag you, on the grounds that you had outweighing motivating reasons not to delay decision making, "overthink" it, or rationalise this *prima facie* wrongdoing to yourself.) Finally, since you are almost always in a state of deliberative attunement when you are in a choice situation requiring a decision, it is not hard for

[167] But even if it did not reveal whether your reasons were outweighing, it would not matter, if these judgments were nevertheless based on beliefs or credences in underlying wrong-making features; see §7.4.

motivating reasons to be accessible through deliberative attunement. Relatedly, sometimes guilt might reveal to you that you *did* in fact have outweighing motivating reasons not to commit wrongdoing at the time, which were relevantly accessible, even though they did not occur to you at the time. This would suggest that *guilt* does, in the end, track something relevant for culpability—namely, the presence of your relevantly accessible reasons at the time.

Now, a good many, if not *most*, of our judgments of responsibility and blameworthiness are made about others. But here too there are reasons why I am not so concerned about Rosen's point about the opacity of the mind. We can tell a lot about what people stand for, or know—that is, what reasons they have—from their characters, practices, habits, roles, relationships, communities, and so on, and from how they communicate or justify their actions to us. This mass of information can subsequently provide relative confidence in judgments of responsibility or of blameworthiness (i.e., in blaming). It is this mass of information that provides me with confidence to hold my friend blameworthy for his recent relapse, or to blame my old teacher for leading a double-life and lying to his wife for many years; knowing his background and theological commitments, I am sure that he would have had right and outweighing motivating reasons to avoid his duplicitous lifestyle. At any rate, my position contrasts favourably with volitionism, according to which we must also know whether these agents formed the specific verdictive judgment that they ought all-things-considered not to act as they did, that these agent's beliefs were *occurrent*, at the time of acting, *and* that they had the right beliefs about morality, or about right and wrong.

But in the end, I am not fazed if my theory recommends a certain weakened degree of revisionism. Many of the theories that we have investigated do the same—for example, the internalist theories discussed in Chapter Seven. What my theory recommends is some caution against being quick to judge or jump to conclusions about someone's blameworthiness—and I believe it recommends just the right amount of caution to provide an antidote to the widespread tendency to shift blame or point fingers.

9.4 Conclusion

We began our inquiry with reflection on the Józefów Massacre and a question about the likelihood of the Battalion 101 shooters' blameworthiness for the executions that they performed. I registered the intuition that the majority of the shooters were probably

blameworthy for the executions, and I pitted that intuition against the volitionist verdict that they were probably blameless. My account of the epistemic condition helps to vindicate this original intuition by allowing room for a variety of epistemic grounds for blameworthiness. Even if, as one of the policemen later reflected, "'[only] later did it first occur to [them] that [the massacre] had not been right'" (Browning 1992, 72), my account could have secured their culpability if at the time, and on balance, their motivating reasons (even those that did not occur to them) weighed in favour of not participating in the massacre because those reasons were sensitive to the normative reasons why doing so was wrong. I think that this could have happened, what is more, even if what the policemen wrongly mistook to be personal or prudential reasons not to participate (e.g., displeasure at the sight of dead bodies) actually tracked *moral* revulsion toward what they had to do—however suppressed its "moral" nature was in their occurrent consciousness. In this connection it must surely be admitted that the Battalion policemen had one or more of the following beliefs at the time of the choice: that the Jews are humans and that humans generally should not be killed unjustly; that whatever the Nazi utopia may require it would be cruel to kills Jews in this particular way; that it would be better to save the lives of the women and children as well as the men; that it is unlikely that these Jews deserved to be killed, even if they can somehow be identified as enemies in common with an allied force who are dropping bombs on our families back at home; that it would overall be better for everyone if the massacre did not occur; that other policemen respected for their moral wisdom have already chosen to opt out; that one cannot bear to think about how one's grandparents, family, or old Jewish friend would react to seeing one participating in such an action; that when given the chance to opt out, one should not follow an order just because everyone else is doing so; that the anti-Nazist political party to which one has some allegiance would not condone this measure; and so forth. Moreover, it is plausible that if the Battalion 101 shooters had these as beliefs (or even as non-negligible credences), they had them as *motivating reasons* (if not fully decisive reasons) against participating in the shootings, and it is plausible that these reasons would have been combined in various ways to tip the balance of their reasons overall against following orders. Furthermore, the policemen might also have had higher-order principles (*qua* motivating reasons) against the act possessed by the Battalion 101 policemen, concerning what to do when morally uncertain about what to do (when one's first-order motivating reasons do not decide either way). Quite probably, in my mind, there would have been the thought (or implicit but consciously accessible belief) that it was overall morally riskier to participate in the shootings, that it was more morally conscientious or cautious to

opt out—or that on the vast majority of views (even Nazi views) about the morality of this type of "special action," it was morally permissible to opt out of the shootings while on a good many, impermissible to participate. Did these policemen not have a "gut feeling" that there was something seriously "off" about what they had to do, a feeling that resulted in a diminution of the force of the reasons in favour of the massacre? I find it hard to believe that they did not. But any of these reasons or combinations thereof could have been epistemic grounds for the policemen's culpability, as long as overall, their wrong-sensitive motivating reasons weighed in favour of an alternative at the time and were at least consciously accessible in virtue of their deliberative attunement. But even if their reasons were not thus right and outweighing at the time of having to make choice, my hunch (reading Browning 1992 at least) is that they would have had opportunities prior to the massacre to question the correctness of Nazi ideology and on what the consequences would be of complying with the Nazis (given the battalion members' relatively non-Nazified background). This would then have created the conditions from which there could arise indirect responsibility for participating in a massacre (e.g., foresight of doing something terrible under Nazi influence). In this way, I hope that one can see that my account successfully captures the initial intuition of the Battalion 101 shooters' probable blameworthiness.

It is natural to begin inquiry into the epistemic condition on moral blameworthiness by raising the question of whether knowledge or awareness of wrongdoing is required for the wrongdoer to be blameworthy. Indeed, I consistently found this to be the best way of conveying the kind of issue that I work on to others, despite that question's subtle difference from the thesis' guiding question (about whether blameworthiness for actions or omissions depends on beliefs or credences concerning their moral significance). In this connection, the most natural interpretation of "awareness of wrongdoing" is "awareness of the fact that the action is wrong," and perhaps also "for the reasons that make it wrong." This is what explains the initial appeal of volitionism I think. But in this thesis, I have argued that, not only should the first question be varied to account for certain nuances (e.g., is the fundamental epistemic state of interest "knowledge" or "awareness"? must the awareness be "occurrent"? must it be awareness of *wrongdoing*? under what description of "wrongdoing" must it be awareness thereof? how "full" must the awareness be? at what *time* must there be the relevant awareness? etc.), volitionism itself needs to be varied in order to account for intuitions generated by what initially appear to be borderline cases (e.g., acting contrary to the belief that acting *might* be wrong) but which actually turn out to be a whole range of cases in which wrongdoers fail to meet volitionist conditions (collectively accounting, I think, for the

majority of instances of blameworthiness). That, at any rate, is what I have argued in this thesis, and it has led me to embracing the following view, that:

E-B: An agent satisfies the epistemic condition on blameworthiness for wrongdoing *x* in a certain choice situation C *if and only if, either*, in C, she has right and outweighing motivating reasons to avoid *x* that are explicit or accessible through deliberative attunement; *or* there was some wrongdoing *y* in some earlier choice situation D, when she had right and outweighing reasons to avoid *y* that were explicit or accessible through deliberative attunement, and these reasons included the fact that *y* would increase the risk of performing an action/omission of the relatively coarse-grained type *W* to which *x* belongs, and included right and outweighing motivating reasons against performing an action/omission of type *W*.

E-B is the conclusion that I have derived from intuitions generated by reflection on individual cases, but more importantly from defended accounts of the nature of *blame* (as holding morally at fault), *blameworthiness* (as being morally at fault), and *responsibility* (as consisting in direct and indirect responses to the reasons)—as well as on the basis of a consideration of *fair moral expectations* to avoid wrongdoing.

The full significance of this account will have to be left for further inquiry. In my Introduction I noted a few areas for which work on the epistemic condition (on blameworthiness for wrongdoing) has both theoretical and practical significance, including the way that it can be applied to thinking about the epistemic condition on blameworthiness for things other than actions and omissions (as well as the epistemic condition on praiseworthiness), the way that it can be applied to how we should treat others more generally, and the way that it can be applied to the more specific areas of the ethics of religious belief and criminal law. I believe that my account provides just the right foundation for beginning to think about these issues.

Bibliography

Adams, Robert. 1985. “Involuntary Sins.” *The Philosophical Review* 94 (1): 3–31.

Alexander, Larry, and K. K. Ferzan. 2009. “Against Negligence Liability.” In *Criminal Law Convserations*, edited by P. H. Robinson, S. Garvey, and F. K. Kessler, 273-94. New York: Oxford University Press.

Alston, William P. 1988. “The Deontological Conception of Epistemic Justification.” *Philosophical Perspectives* 2 (Epistemology): 257–99.

Alvarez, Maria. 2016. “Reasons for Action: Justification, Motivation, Explanation.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/archives/win2017/entries/reasons-just-vs-expl/.

Anton, Audrey L. 2015. *Moral Responsibility and Desert of Praise and Blame*. Lexington Books.

Aristotle. 2013. *Nicomachean Ethics*. Edited by David Ross. Start Publishing LLC.

Arpaly, Nomy. 2002. *Unprincipled Virtue*. Oxford: Oxford University Press.

———. 2015. “Huckleberry Finn Revisited: Inverse Akrasia and Moral Ignorance.” In *The Nature of Moral Responsibility: New Essays*, edited by Randolph Clarke, Michael McKenna, and Angela M. Smith, 603–10.

Bedau, Hugo Adam, and Erin Kelly. 2015. “Punishment.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/punishment/.

Bishop, John. 2007. *Believing by Faith: An Essay in the Epistemology and Ethics of Religious Belief*. Oxford: Oxford University Press.

Björnsson, Gunnar. 2017. "Explaining Away Epistemic Skepticism about Culpability." *Oxford Studies in Agency and Responsibility: Volume 4*, edited by David Shoemaker, 165-182. Oxford: Oxford University Press.

Bratman, Michael E. 1979. “Practical Reasoning and Weakness of the Will.” *Nous* 13 (2): 153–71.

Browning, Christopher R. 1992. *Ordinary Men: Reserve Police Battalion 101 and the Final Solution in Poland*. New York, US: HarperCollins Publishers.

Bryant, James C. 2016. “Epistemic Blame: Its Nature and Its Norms.” Baylor University. https://baylor-ir.tdl.org/bitstream/handle/2104/9833/BRYANT-DISSERTATION-2016.pdf?sequence=1.

Bykvist, Krister. 2014. “Evaluative Uncertainty and Consequentialist Environmental Ethics.” In *Environmental Ehics and Consequentialism*, edited by L. Kahn and A. Hiller. Routledge.

———. 2017. “Moral Uncertainty.” *Philosophy Compass* 12 (3): 1–8.

Capes, Justin A. 2012. “Blameworthiness Without Wrongdoing.” *Pacific Philosophical Quarterly* 93 (3): 417–37.

Clarke, Randolph. 2014. “Negligent Action and Unwitting Omission.” In *Omissions: Agency*,

Metaphysics, and Responsibility. New York, US: Oxford University Press.

———. 2017. “Blameworthiness and Unwitting Omissions.” In *The Ethics and Law of Omissions*, edited by Dana Kay Nelkin and Samuel C. Rickless. Oxford: Oxford University Press.

Clifford, William K. 1886. “The Ethics of Belief.” In *Lectures and Essays*, edited by Leslie Stephen and Frederick Pollock. London: Macmillian and Co.

Coates, D. Justin, and Philip Swenson. 2013. “Reasons-Responsiveness and Degrees of Responsibility.” *Philosophical Studies* 165 (2): 629–45.

Coates, D. Justin, and Neal A. Tognazzini. 2012. “The Nature and Ethics of Blame.” *Philosophy Compass* 7 (3): 197–207.

Code, Lorraine. 1987. *Epistemic Responsibility*. Hanover and London: Brown University Press.

Cogley, Zac. 2015. “Rolling Back the Luck Problem for Libertarianism.” *Journal of Cognition and Neuroethics* 3 (1): 121–37.

Cohen, L. Jonathan. 1992. *An Essay on Belief and Acceptance*. Oxford: Oxford University Press.

Conee, Earl, and Richard Feldman. 2004. *Evidentialism: Essays in Epistemology*. Oxford: Oxford University Press.

Corlett, J. Angelo. 2008. “Epistemic Responsibility.” *International Journal of Philosophical Studies* 16 (2): 179–200.

Curley, Edwin. 1975. “Descartes, Spinoza and the Ethics of Belief.” *Spinoza: Essays in Interpretation*, edited by Maurice Mandelbaum and Eugene Freeman, 159–89. La Salle, IL: Open Court.

Dancy, Jonathan. 2000. *Practical Reality*. Oxford: Oxford University Press.

———. 2004. *Ethics Without Principles*. Oxford: Oxford University Press.

———. 2011. “Acting in Ignorance.” *Frontiers of Philosophy in China* 6 (3): 345–57.

Daniels, Norman. 1979. “Wide Reflective Equilibrium and Theory Acceptance in Ethics.” *Journal of Philosophy* 76 (5): 256–82.

———. 2016. “Reflective Equilibrium.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Summer 2020. https://plato.stanford.edu/entries/reflective-equilibrium/.

Davidson, Donald. 1980. *Essays on Actions and Events*. Oxford: Oxford University Press.

———. 2001. “How Is Weakness of the Will Possible?” In *Essays on Actions and Events*, edited by Donald Davidson, 2nd ed., 21–42. Oxford: Oxford University Press.

Driver, Julia. 1992. “The Suberogatory.” *Australasian Journal of Philosophy* 70 (3): 286–95.

Eriksson, Lina, and Alan Hájek. 2007. “What Are Degrees of Belief?” *Studia Logica* 86 (2): 183–213.

Eshleman, Andrew. 2014. “Moral Responsibility.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2019. https://plato.stanford.edu/archives/fall2019/entries/moral-responsibility/.

Field, Claire. 2019. “Recklessness and Uncertainty: Jackson Cases and Merely Apparent Asymmetry.” *Journal of Moral Philosophy*.

Finlay, Stephen, and Mark Schroeder. 2017. “Reasons for Action: Internal vs. External.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/reasons-internal-external/.

Fischer, John Martin. 2006. “Responsiveness and Moral Responsibility.” In *My Way: Essays on Moral Responsibliity*. New York: Oxford University Press.

———. 2007. “Response to Kane, Pereboom, and Vargas.” In *Four Views on Free Will*, 184–90. Malden, MA: Blackwell.

Fischer, John Martin, and Mark Ravizza. 1993. “Introduction.” In *Perspectives on Moral Responsibility*, edited by John Martin Fischer and Mark Ravizza. Cornell University Press.

———. 1998. *Responsiblity and Control: A Theory of Moral Responsibility*. Cambridge: Cambridge University Press.

Fischer, John Martin, and Neal A Tognazzini. 2009. “The Truth about Tracing.” *Noûs* 43 (3): 531-556.

FitzPatrick, William J. 2008. “Moral Responsibility and Normative Ignorance: Answering a New Skeptical Challenge.” *Ethics* 118 (4): 589–613.

———. 2017. “Unwitting Wrongdoing, Reasonable Expectations, and Blameworthiness.” In *Responsibility: The Epistemic Condition*, edited by Philip Robichaud and Jan Willem Wieland, 29–46. Oxford: Oxford University Press.

Foot, Philippa. 1972. “Morality as a System of Hypothetical Imperatives.” *The Philosophical Review* 81 (3): 305.

Frankfurt, Harry G. 1969. “Alternate Possibilities and Moral Responsibility.” *The Journal of Philosophy* 66 (23): 829–39.

———. 1971. “Freedom of the Will and the Concept of a Person.” *The Journal of Philosophy* 68 (1): 5–20.

———. 2005. *On Bullshit*. Princeton: Princeton University Press.

Fritz, Kyle G. 2014. “Responsibility For Wongdoing Without Blameworthiness: How It Makes Sense and How It Doesn’t.” *Philosophical Quarterly* 64 (257).

Furlong, Peter. 2017. “Aquinas & the Epistemic Condition for Moral Responsibility.” *Res Philosophica* 94 (1): 43–65. https://doi.org/10.11612/resphil.1487.

Gallois, Andre. 1998. “De Re/de Dicto.” *Routledge Encyclopedia of Philosophy*. Taylor & Francis. https://www.rep.routledge.com/articles/thematic/de-re-de-dicto/v-1.

Gettier, Edmund. 1963. “Is Justified True Belief Knowledge?” *Analysis* 23 (6): 121–23.

Geyer, Jay. 2018. “Moral Uncertainty and Moral Culpability.” *Utilitas* 30 (4): 399–416.

Ginet, Carl. 2000. “The Epistemic Requirements for Moral Responsibility.” *Philosophical Perspectives* 14: 267–77.

———. 2001. “Deciding to Believe.” In *Knowledge, Truth, and Duty: Essays on Epistemic Justification, Responsibility, and Virtue*, edited by Matthias Steup. New York, US: Oxford University Press.

Goldman, Alvin. 1988. “Strong and Weak Justification.” *Philosophical Perspectives* 2: 51–69.

Guerrero, Alexander A. 2007. “Don’t Know, Don’t Kill: Moral Ignorance, Culpability, and Caution.” *Philosophical Studies* 136 (1): 59–97.

Gustafsson, Johan E., and Olle Torpman. 2014. “In Defence of My Favourite Theory.” *Pacific Philosophical Quarterly* 95 (2): 159–74.

Haji, Ishtiyaque. 1997. “An Epistemic Dimension of Blameworthiness.” *Philosophy and Phenomenological Research* 57 (3): 523–44.

———. 2010. “Incompatibilism and Prudential Obligation.” *Canadian Journal of Philosophy* 40 (3): 385–410.

Hampton, Jean. 1984. “The Moral Education Theory of Punishment.” *Philosophy and Public Affairs* 13: 208–38.

Harman, Elizabeth. 2011. “Does Moral Ignorance Exculpate?” *Ratio* 24 (4): 443–68.

———. 2015. “The Irrelevance of Moral Uncertainty.” In *Oxford Studies in Metaethics*, edited by Russ Shafer-Landau, 10:53–79. Oxford: Oxford University Press.

———. 2016. “Morally Permissible Moral Mistakes.” *Ethics* 126 (2): 366–93.

Harman, Gilbert. 1986. *Change in View*. Cambridge, MA: MIT Press.

Hartman, Robert J. 2016. “Against Luck-Free Moral Responsibility.” *Philosophical Studies* 173 (10): 2845–65.

Hauser, Marc. 2006. *Moral Minds: How Nature Designed Our Sense of Right and Wrong*. HarperCollins Publishers.

Hawthorne, James, and Luc Bovens. 1999. “The Preface, the Lottery, and the Logic of Belief.” *Mind* 108 (430): 241–64.

Hieronymi, Pamela. 2004. “The Force and Fairness of Blame.” *Philosophical Perspectives* 18. https://about.jstor.org/terms.

———. 2008. “Responsibility for Believing.” *Synthese* 161: 357–73.

Hursthouse, Rosalind. 1999. *On Virtue Ethics*. Oxford: Oxford University Press.

Husak, Douglas. 2011. “Negligence, Belief, Blame and Criminal Liability: The Special Case of Forgetting.” *Criminal Law and Philosophy* 5 (2): 199–218.

———. 2016. *Ignorance of Law: A Philosophical Inquiry*. New York, US: Oxford University Press.

Hyman, John. 2011. “Acting for Reasons: Reply to Dancy.” *Frontiers of Philosophy in China* 6 (3): 358–68.

Jackson, Frank. 1991. “Decision-Theoretic Consequentialism and the Nearest and Dearest Objection.” *Ethics* 101 (3): 461–82.

James, William. 1896. *The Will to Believe*. New York: Longmans, Green, and Co.

Jansen, Bart. 2015. “NTSB: Pilots Left Wing Controls Locked in Place in Gulfstream Crash That Killed 7.” USA Today. September 9, 2015. https://www.usatoday.com/story/news/2015/09/09/ntsb-bedford-crash-gulfstream-philadelphia-inquirer-lewis-katz/71922242/.

Joyce, Richard. 2015. “Moral Anti-Realism.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/moral-anti-realism/.

Kane, Robert. 2007. “Libertarianism.” In *Four Views on Free Will*. Malden, MA: Blackwell.

Kauppinen, Antti. 2018. “Epistemic Norms and Epistemic Accountability.” *Philosophers Imprint* 18 (8): 1-16.

Keller, Simon. 2004. “Friendship and Belief.” *Philosophical Papers* 3: 329–51.

King, Matt. 2009. “The Problem with Negligence.” *Social Theory and Practice* 35 (4).

———. 2017. “Tracing the Epistemic Condition.” In *Responsibility: The Epistemic Condition*, edited by Philip Robichaud and Jan Willem Wieland, 1–25. Oxford: Oxford University Press.

Levy, Neil. 2005. “The Good, the Bad and the Blameworthy.” *Journal of Ethics and Social Philosophy* 1 (2): 1–16.

———. 2009. “Culpable Ignorance and Moral Responsibility: A Reply to FitzPatrick.” *Ethics* 119 (4): 729–41.

———. 2011. *Hard Luck: How Luck Undermines Free Will and Moral Responsibility*. Oxford: Oxford University Press.

———. 2014. *Conciousness and Moral Responsibility*. Oxford: Oxford University Press.

———. 2016. “Culpable Ignorance: A Reply a Robichaud.” *Journal of Philosophical Research* 41: 263–71.

———. 2017. “Methodological Conservatism and the Epistemic Condition.” In *Responsibility: The Epistemic Condition*, edited by Philip Robichaud and Jan Willem Wieland, 252–65. Oxford: Oxford University Press.

Levy, Neil, and Michael McKenna. 2009. “Recent Work on Free Will and Moral Responsibility.” *Philosophy Compass* 4 (1): 96–133.

Littlejohn, Clayton. 2014. “The Unity of Reason.” In *Epistemic Norms: New Essays on Action, Belief, and Assertion*, 135–54. Oxford: Oxford University Press.

Macnamara, Coleen. 2011. “Holding Others Responsible.” *Philosophical Studies* 152 (1): 81–102.

Marilyn, Paul. 2018. “Moving from Blame to Accountability.” The Systems Thinker. 2018. https://thesystemsthinker.com/moving-from-blame-to-accountability/.

Mason, Elinor. 2015. “Moral Ignorance and Blameworthiness.” *Philosophical Studies* 172 (11): 3037–57.

McDowell, John. 1995. “Might There Be External Reasons?” In *World, Mind and Ethics: Essays on the Ethical Philosophy of Bernard Williams*, edited by J. E. J. Altham and R. Harrison, 68–85. Cambridge: Cambridge University Press.

McGeer, Victoria. 2014. “P. F. Strawson’s Consequentialism.” In *Oxford Studies in Agency and Responsibility, Volume 2: 'Freedom and Resentment' at 50*, edited by David Shoemaker and Neil Tognazzini. Oxford: Oxford University Press.

McKenna, Michael. 2012. *Conversation and Responsibility*. Oxford: Oxford University Press.

———. 2013. “Reasons-Responsiveness, Agents, and Mechanisms.” In *Oxford Studies in Agency and Responsibility Volume 1*, edited by David Shoemaker, 151-81. Oxford: Oxford University Press.

McKenna, Michael, and D. Justin Coates. 2019. “Compatibilism.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/compatibilism/.

McNamara, Paul. 2010. “Deontic Logic.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/logic-deontic/.

Mele, Alfred, and David Robb. 2003. “Bbs, Magnets and Seesaws: The Metaphysics of

Frankfurt-Style Cases." In *Moral Responsibility and Alternative Possibilities*, edited by David Widerker and Michael McKenna, 127–38. Aldershot: Ashgate Press.

Moller, D. 2011. "Abortion and Moral Risk." *Philosophy* 86 (3): 425–43.

Montmarquet, James. 1992. "Epistemic Virtue and Doxastic Responsibility." *American Philosophical Quarterly* 29 (4): 331–41.

———. 1993. *Epistemic Virtue and Doxastic Responsibility*. Lanham, MD: Rowman and Littlefield.

———. 1995. "Culpable Ignorance and Excuses." *Philosophical Studies* 80 (1): 41–49.

———. 1999. "Zimmerman on Culpable Ignorance." *Ethics* 109 (4): 842–45. https://doi.org/10.1086/233949.

Moody-Adams, Michele. 1994. "Culture, Responsibility, and Affected Ignorance." *Ethics* 104 (2): 291–309.

Murphy, Mark. 2015. "6 Words For Stopping Blame And Increasing Accountability." Forbes. 2015. https://www.forbes.com/sites/markmurphy/2015/06/12/6-words-for-stopping-blame-and-increasing-accountability/#3e5a82cc57c3.

Murray, Samuel. 2017. "Responsibility and Vigilance." *Philosophical Studies* 174 (2): 507–27.

Murray, Samuel, and Manuel Vargas. 2020. "Vigilance and Control." *Philosophical Studies* 177 (3): 825–43.

Nelkin, Dana. 2004. "The Sense of Freedom." In *Freedom and Determinism*, edited by Joseph Campbell, Michael O'Rourke, and David Shier, 105–34. Cambridge, MA: MIT Press.

———. 2011. *Making Sense of Freedom and Responsibility*. Oxford: Oxford University Press.

Nelkin, Dana Kay, and Samuel C. Rickless. 2017. "Moral Responsibility for Unwitting Omissions." In *The Ethics and Law of Omissions*, edited by Dana Kay Nelkin, and Samuel C. Rickless, 1–32. New York, US: Oxford University Press.

Nerlich, Volker. 2007. "Superior Responsibility under Article 28 ICC Statute: For What Exactly Is the Superior Held Responsible?" *Journal of International Criminal Justice* 5 (3): 665–82.

Nottelmann, Nikolaj. 2007. *Blameworthy Belief: A Study in Epistemic Deontologism*. Dordrecht: Springer Netherlands.

———. 2013. "The Deontological Conception of Epistemic Justification: A Reassessment." *Synthese* 190 (12): 2219–41.

Oshana, Marina A. L. 1997. "Ascriptions of Responsibility." *American Philosophical Quarterly* 34 (1): 71–83.

Owens, David. 2000. *Reason Without Freedom: The Problem of Epistemic Normativity*. London: Routledge.

Pappas, George. 2014. "Internalist vs. Externalist Conceptions of Epistemic Justification." In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/justep-intext/.

Peels, Rik. 2010. "What Is Ignorance?" *Philosophia* 38 (1): 57–67.

———. 2011. "Tracing Culpable Ignorance." *Logos & Episteme* 2 (4): 575–82.

———. 2014. “What Kind of Ignorance Excuses? Two Neglected Issues.” *Philosophical Quarterly* 64 (256): 478–96.

———. 2017. *Responsible Belief: A Theory in Ethics and Epistemology*. Oxford: Oxford University Press.

Pereboom, Derk. 2016. “Omissions and Different Senses of Responsibility.” In *Agency, Freedom, and Moral Responsibility*, edited by Andrei Buckareff, Carlos Moya, and Sergi Rosell, 179–91. New York, US: Palgrave-Macmillan.

Peterson, Martin. 2017. “Radical Evaluative Ignorance.” In *Perspectives on Ignorance from Moral and Social Philosophy*, edited by Rik Peels. New York, US: Routledge.

Pickard, Hanna. 2013. “Irrational Blame.” *Analysis* 73 (4): 613–26.

Plantinga, Alvin. 1988. “Positive Epistemic Status and Proper Function.” *Philosophical Perspectives* 2: 1–50.

Raffaele, Paul. 2006 “Sleeping with Cannibals.” *Smithsonian Magazine*. Accessed July 24, 2020. https://www.smithsonianmag.com/travel/sleeping-with-cannibals-128958913/.

Rawls, John. 1999 [1971]. *A Theory of Justice*. Revised Ed. Cambridge, MA: The Belknap Press of Harvard University Press.

Raz, Joseph. 2002. *Engaging Reason*. *Engaging Reason*. Oxford: Oxford University Press.

Reed, Baron. 2008. “Certainty.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/certainty/.

Robichaud, Philip. 2014. “On Culpable Ignorance and Akrasia.” *Ethics* 125 (1): 137–51.

Robinson, Darryl. 2017. “A Justification of Command Responsibility.” *Criminal Law Forum* 28 (4): 633–68.

Rosen, Gideon. 2003. “Culpability and Ignorance.” *Proceedings of the Aristotelian Society* 103: 61–84.

———. 2004. “Skepticism about Moral Responsibility.” *Philosophical Perspectives* 18: 295–313.

———. 2008. “Kleinbart the Oblivious and Other Tales of Ignorance and Responsibility.” *The Journal of Philosophy* 105 (10): 591–610.

Ross, W. D. 1939. *Foundations of Ethics*. Oxford: Oxford University Press.

Rudy-Hiller, Fernando. 2017. “A Capacitarian Account of Culpable Ignorance.” *Pacific Philosophical Quarterly* 98 (2017): 398–426.

———. 2018. “The Epistemic Condition for Moral Responsibility.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta, Fall 2018. https://plato.stanford.edu/entries/moral-responsibility-epistemic/.

———. 2019. "Give People a Break: Slips and Moral Responsibility." *Philosophical Quarterly* 69 (277): 721-740.

Ryan, Sharon. 2003. “Doxastic Compatibilism and the Ethics of Belief.” *Philosophical Studies* 114 (1–2): 47–79.

Sartorio, Carolina. 2016. *Causation and Free Will*. Oxford: Oxford University Press.

———. 2017. “Ignorance, Alternative Possibilities, and the Epistemic Conditions for Responsibility.” In *Perspectives on Ignorance from Moral and Social Philosophy*, edited by Rik Peels, 15–29. New York, US: Routledge.

Scanlon, T. M. 1998. *What We Owe To Each Other*. Cambridge, MA; London, UK: The Belknap Press of Harvard University Press.

Schlick, Moritz. 1966. “When Is a Man Responsible?” In *Free Will and Determinism*, edited by Bernard Berofsky, 54–63. New York, US: Harper & Row.

Schwitzgebel, Eric. 2002. “A Phenomenal, Dispositional Account of Belief.” *Nous* 36 (2): 249–75.

———. 2019. “Belief.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/belief/.

Sepielli, Andrew. 2009. “What to Do When You Don’t Know What to Do.” In *Oxford Studies in Metaethics: Volume 4*, edited by Russ Shafer-Landau, 5–28. Oxford: Oxford University Press.

———. 2012. “Subjective Normativity and Action Guidance.” In *Oxford Studies in Normative Ethics: Volume 2*, edited by Mark Timmons, 45–73. Oxford: Oxford University Press.

———. 2016. “What to Do When You Don’t Know What to Do When You Don’t Know What to Do...” *Nous* 48 (3): 521–44.

———. 2017. “How Moral Uncertaintism Can Be Both True and Interesting.” In *Oxford Studies in Normative Ethics: Volume 7*, edited by Mark Timmons, 1–303. Oxford: Oxford University Press.

Sher, George. 2005. *In Praise of Blame*. Oxford: Oxford University Press.

———. 2009. *Who Knew?: Responsibility Without Awareness*. Oxford: Oxford University Press.

Shoemaker, David. 2011. “Attributability, Answerability, and Accountability: Toward a Wider Theory of Moral Responsibility.” *Ethics* 121 (3): 602–32.

———. 2015. *Responsibility from the Margins*. Oxford: Oxford University Press.

———. 2017. “Response-Dependent Responsibility; or, A Funny Thing Happened on the Way to Blame.” *Philosophical Review* 126 (4): 481-527.

Singer, Peter. 1993. *Practical Ethics*. 2nd ed. Cambridge, MA: Cambridge University Press.

Slote, Michael. 2001. *Morals from Motives*. Oxford: Oxford University Press.

Smart, J. J. C. 1961. “Free-Will, Praise and Blame.” *Mind* 70 (279): 291–306.

Smith, Angela M. 2005. “Responsibility for Attitudes: Activity and Passivity in Mental Life.” *Ethics* 115 (2): 236–71.

———. 2008. “Control, Responsibility, and Moral Assessment.” *Philosophical Studies* 138 (3): 367–92.

———. 2012. “Attributability, Answerability, and Accountability: In Defense of a Unified Account.” *Ethics* 122 (3): 575–89.

———. 2017. “Unconscious Omissions, Reasonable Expectations, and Responsibility.” In *The Ethics and Law of Omissions*, edited by Dana Kay Nelkin and Samuel C. Rickless, 36–60. Oxford: Oxford University Press.

Smith, Holly M. 1983. “Culpable Ignorance.” *The Philosophical Review* 92 (4): 543–71.

———. 2011. “Non-Tracing Cases of Culpable Ignorance.” *Criminal Law and Philosophy* 5 (2): 115–46.

Smith, Michael A. 1994. *The Moral Problem*. Malden, MA: Blackwell.

Snedegar, Justin. 2018. “Reasons for and Reasons Against.” *Philosophical Studies* 175 (3): 725–43.

Sosa, Ernest. 1980. “The Raft and the Pyramid: Coherence versus Foundations in the Theory of Knowledge.” *Midwest Studies in Philosophy* 5 (1): 3–26.

Steup, Matthias. 1988. “The Deontic Conception of Epistemic Justification.” *Philosophical Studies* 53: 65–84.

———. 2011. “Belief, Voluntariness and Intentionality.” *Dialectica* 65 (4): 537–59.

Steup, Matthias, and Ram Neta. 2020. “Epistemology.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/epistemology/.

Stratton-Lake, Philip. 2001. *Kant, Duty, and Moral Worth*. London: Routledge.

Strawson, Galen. 1994. “The Impossibility of Moral Responsibility.” *Philosophical Studies* 75 (1): 5–24.

Strawson, P. F. 1993 [1962]. “Freedom and Resentment.” In *Perspectives on Moral Responsibility*, edited by John Martin Fischer and Mark Ravizza, 45–66. Cornell University Press.

Stroud, Sarah, and Larisa Svirsky. 2019. “Weakness of Will.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/weakness-will/.

Styron, William. 1979. *Sophie’s Choice*. New York, US: Random House.

Swinburne, Richard. 2012. *Mind, Brain, and Free Will*. Oxford: Oxford University Press.

Talbert, Matthew. 2008. “Blame and Responsiveness to Moral Reasons: Are Psychopaths Blameworthy?” *Pacific Philosophical Quarterly* 89: 516-35.

———. 2012. “Moral Competence, Moral Blame, and Protest.” *The Journal of Ethics* 16: 89–109.

———. 2013. “Unwitting Wrongdoers and the Role of Moral Disagreement in Blame.” In *Oxford Studies in Agency and Responsibility Volume 1*, edited by David Shoemaker. Oxford: Oxford University Press.

———. 2016. *Moral Responsibility*. Cambridge: Polity Press.

———. 2017a. “Akrasia, Awareness, and Blameworthiness.” In *Responsibility: The Epistemic Condition*, edited by Philip Robichaud and Jan Willem Wieland, 281–97. Oxford: Oxford University Press.

———. 2017b. “Omission and Attribution Error.” In *The Ethics and Law of Omissions*, edited by Dana Nelkin and Samuel C. Rickless, 17–35. Oxford: Oxford University Press.

———. 2019. “Moral Responsibility.” *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/moral-responsibility/.

Tappolet, Christine. 2004. “Values, Reasons, and Oughts.” *Philosophical Studies*, 90–117.

Timms, Michael. 2017. “Creating a Culture of Accountability, Not Blame.” Avail Leadership. 2017. https://availleadership.com/culture-of-accountability/.

Timpe, Kevin. 2011. “Tracing and the Epistemic Condition on Moral Responsibility.” *The Modern Schoolman* 88 (1–2): 5–28.

Tognazzini, Neal A., and D. Justin Coates. 2018. “Blame.” In *Stanford Encyclopedia of*

Philosophy, edited by Edward. N. Zalta. https://plato.stanford.edu/entries/blame/.

Vargas, Manuel. 2005. “The Trouble with Tracing.” *Midwest Studies in Philosophy* 29: 269–91.

———. 2013. *Building Better Beings: A Theory of Moral Responsibility*. Oxford: Oxford University Press.

Vinocour, Susan. 2020. “Criminally Insane.” In *Aeon*, edited by Sam Haselby. https://aeon.co/essays/what-can-be-done-to-rehabilitate-the-insanity-defence.

Wallace, R. Jay. 1996. *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Watson, Gary. 1975. “Free Agency.” *The Journal of Philosophy* 72 (8): 205–20.

———. 1996. “Two Faces of Responsibility.” *Philosophical Topics* 24 (2): 227–48.

———. 2001. “Reasons and Responsibility.” *Ethics* 111 (2): 289–317.

Weatherson, Brian. 2014. “Running Risks Morally.” *Philosophical Studies* 167 (1): 141–63.

———. 2019. *Normative Externalism*. Oxford: Oxford University Press.

Wieland, Jan Willem. 2017. “Introduction: The Epistemic Condition.” In *Responsibility: The Epistemic Condition*, edited by Philip Robichaud and Jan Willem Wieland, 1–45. Oxford: Oxford University Press.

Wieland, Jan Willem, and Philip Robichaud. 2017. “Blame Transfer.” In *Responsibility: The Epistemic Condition*, edited by Philip Robichaud and Jan Willem Wieland, 281–97. Oxford: Oxford University Press.

Williams, Bernard. 1985. *Ethics and the Limits of Philosophy*. London: Fontana.

Wilson, George, and Samuel Shpall. 2012. “Action.” In *Stanford Encyclopedia of Philosophy*, edited by Edward N. Zalta. https://plato.stanford.edu/entries/action/.

Wolf, Susan. 1980. “Assymetrical Freedom.” *The Journal of Philosophy* 77 (3): 151–66.

———. 1987. “Sanity and the Metaphysics of Responsibility.” In *Responsibility, Character, and the Emotions*, 46–62. Cambridge: Cambridge University Press.

———. 1990. *Freedom Within Reason*. New York, US: Oxford University Press.

Woudenberg, René Van. 2009. “Responsible Belief and Our Social Institutions.” *Philosophy* 84 (1): 47–73.

Yaffe, Gideon. 2018. “Is Akrasia Necessary for Culpability? On Douglas Husak’s Ignorance of Law.” *Criminal Law and Philosophy* 12 (2): 341–49.

Yalcin, Seth. 2007. “Epistemic Modals.” *Mind* 116 (464): 982–1026.

Zimmerman, Michael J. 1986. “Negligence and Moral Responsibility.” *Nous* 20 (2): 199–218.

———. 1988. *An Essay on Moral Responsibility*. Totowa, NJ: Rowman and Littlefield.

———. 1997a. “A Plea for Accuses.” *American Philosophical Quarterly* 34 (2): 229–43.

———. 1997b. “Moral Responsibility and Ignorance.” *Ethics* 107: 410–26.

———. 2008. *Living with Uncertainty: The Moral Significance of Ignorance*. Cambridge: Cambridge University Press.

———. 2015. “Varieties of Moral Responsibility.” In *The Nature of Moral Responsibility: New Essays*, edited by Randolphe Clarke, Michael McKenna, and Angela M. Smith, 45-

64. Oxford: Oxford University Press.

———. 2017. "Ignorance as a Moral Excuse." In *Perspectives on Ignorance from Moral and Social Philosophy*, edited by Rik Peels, 77-94. New York, US: Routledge