Situationism and Moral Responsibility

Adam Piovarchy

A thesis submitted in fulfilment of requirements for the degree of Doctor of Philosophy

THE UNIVERSITY OF SYDNEY

Department of Philosophy
Faculty of Arts and Social Sciences
2020
Statement of Originality

I certify that the intellectual content of this thesis is the product of my own work, and that all assistance received in preparing this thesis has been acknowledged.
Acknowledgements

I would like to give a special thanks to Isabelle Wentworth, Dominic Dimech, Hannah Tierney, Daphne Brandenburg and James McKeahnie for their very insightful comments on earlier drafts. I am also grateful to Manuel Vargas, David Brink, Lucy Allais, Jesse Miller, Gemma Smart, and Andrew Latham for helpful discussions. Thanks to Russell Thomson at Clear Communications for his proofreading services. I have also benefitted immensely from my audiences at a number of conferences, including many members of The University of Sydney’s philosophy department. Thanks especially to my parents for their unwavering support, patience, and willingness to help out wherever possible, so that I could complete a several-year long project with no determinate end date and questionable employment-enhancing value.

I am thankful for my opportunity to visit The University of California, San Diego, where Dana Nelkin allowed me to sit in on one of her graduate classes on responsibility. Dana is incredibly warm and kind, and went out of her way to make me feel welcomed by the department, despite my short stay. Curiously, her book was the first account of moral responsibility I ever came across, being recommended to me by someone who had no real familiarity with the moral responsibility literature. Whether because of indeterministic luck or (pre)determinism, I have enjoyed having my worldview shaped by her work in this actual-sequence of events.

I am grateful to my associate supervisor Caroline West, who has been a continued source of learning and support during my candidature. Particularly amazing is her ability to always make one feel like their contribution to any philosophical or teaching discussion is as valuable and respectable as hers, despite her intimidatingly large wealth of expertise and experience. I’ve benefitted immensely from her feedback and from her unshakeable sunny disposition.

Finally, this thesis could not have been possible without my primary supervisor, Luke Russell. Luke has been far and above what any graduate student could ever hope for in a mentor. Intellectually, he has been insightful, rigorous, charitable, open-minded, and an exemplar of what supervisors and philosophers should aspire to be. Equally as important, he has been positive, humorous, enthusiastic, encouraging, kind, and attentive to every important factor that contributes to the ups and downs of PhD life.
To be honest, some of the evidence available in 2015 suggested accepting me into a PhD program could be somewhat risky. My undergraduate marks were rather average, and half of the philosophy units I had taken were in applied ethics, none of which touched on topics like free will, moral responsibility, or character traits. My Master’s degree supervisor was (justifiably) hesitant to recommend I take further studies, because having completed that degree by distance, I didn’t know much about what academia was like, or how much philosophy I’d missed out on learning while choosing to write yet another essay on Peter Singer. Despite all this, Luke took a chance on me, without ever insinuating that any risk was being taken. He helped craft my early inchoate thoughts and writing ability, and was always ready with positive comments to balance out the constructive ones. He’s never complained about the constant stream of forms I’ve needed him to sign when applying for funds and conferences, or the numerous draft arguments I’ve rushed enthusiastically into and had him look at, only to then discard some months later. Being his modest self, if you ask him about this characterisation, he’ll probably say I’m exaggerating and my early writing samples were all fine (he’s wrong; I’ve still got them).

I could give numerous other examples of how these people have helped me, doing things to the right degree, at the right time and in the right way for the right purpose. But I think the defining feature of the virtuous academic, and the thing I am most appreciative to have experienced during my candidature, was best articulated by someone outside of the academy:

Why do we love our writing teachers so much? Why, years later, do we think of them with such gratitude? I think it’s because they come along when we need them most, when we are young and vulnerable and are tentatively approaching this craft that our culture doesn’t have much respect for, but which we are beginning to love. They have so much power. They could mock us, disregard us, use us to prop themselves up. But our teachers, if they are good, instead do something almost holy, which we never forget: they take us seriously. They accept us as new members of the guild.

George Saunders, ‘My Writing Education: A Time Line’

*The New Yorker*, October 22, 2015.
Abstract

This thesis examines whether it is appropriate to blame the subjects who act wrongly in the situationist psychology experiments for their actions, focusing on the subjects in Milgram’s ‘Obedience to Authority’ studies. Both philosophers and psychologists currently lack any convincing explanation for why subjects in these experiments behave as they do. However, one promising avenue which has not been considered is to examine subjects’ perceived reasons for action. If subjects act differently to how we expect because they do not share our assessments of their reasons for action, this would explain their behaviour.

The first part of this thesis argues that, given a number of sources of evidence from the situationist experiments, and other related experiments which philosophers have not considered, the best explanation of subjects’ wrongdoing is they had a reduced capacity to avoid wrongdoing. This entails that they are excused for their wrongdoing, and thus not blameworthy.

The second part of this thesis sees what follows if the subjects in these experiments did possess the capacity to avoid wrongdoing, but simply didn’t exercise it. Given the high rates of wrongdoing in these studies, and consistent replication of results with subjects from number of social groups, we have strong evidence that most members of the moral community would also have committed wrongdoing in these experiments. Due to the relationship between standing to blame and hypocrisy, the fact most of us would have committed wrongdoing in these experiments undermines our standing to blame. In particular, we lack the normative authority to make second-personal demands when blaming.
# Table of Contents

Introduction .................................................................................................................. 8  
  i. Rejoining the Hunt ................................................................................................. 8  
  ii. Preliminary Concerns and Caveats .................................................................... 12  
  iii. Précis .................................................................................................................. 14  

Chapter 1: A Brief History of Moral Responsibility ...................................................... 19  
  1.1 Strawson and Our Moral Practices ...................................................................... 19  
  1.2 Quality of Will .................................................................................................. 23  
  1.3 Faces of Responsibility ...................................................................................... 26  
  1.4 Why Blame? ...................................................................................................... 28  
  1.5 The Control Condition ...................................................................................... 33  
  1.6 A Primer on Capacities ...................................................................................... 38  
  1.7 The Epistemic Condition ................................................................................... 45  
  1.8 Where We Are .................................................................................................. 48  

Chapter 2: Explanations of Situationist Behaviour ....................................................... 50  
  2.1 Introduction ....................................................................................................... 50  
  2.2 Obedience Revisited .......................................................................................... 51  
  2.3 Review: The Situationist Threat to Virtue ......................................................... 59  
  2.4 Against Desire-Based Explanations ................................................................... 62  
  2.5 Against Trait-Based Explanations ....................................................................... 66  
  2.6 Against Akrasia- and Practical Wisdom-Based Explanations ......................... 68  
  2.7 Against Identity-Based Explanations .................................................................. 69  
  2.8 Against Embarrassment-Based Explanations .................................................. 72  
  2.9 Reason-based Explanations .............................................................................. 73  
  2.10 Conclusion ....................................................................................................... 75  

Chapter 3: Reasoning About Capacities ...................................................................... 76  
  3.1 Introduction ....................................................................................................... 76  
  3.2 Capacity Comparisons ....................................................................................... 77  
  3.3 The Reasonable Person Standard and Relevant Counterfactuals ..................... 79  
  3.4 Capacities and Inferences ................................................................................... 85  
  3.5 Impaired Rational Capacities 1: The Capacity to Perceive Moral Reasons ........ 90  
  3.6 Impaired Rational Capacities 2: The Capacity to Act on Moral Reasons .......... 97  
  3.7 Conclusion ........................................................................................................ 101
Introduction

i. Rejoining the Hunt

When is an agent blameworthy for wrongdoing? Our thoughts on moral responsibility are important to many areas of our lives. They affect our relationships with friends and family, judgements about strangers we meet, decisions to award medals to heroes and, most notably, who we incarcerate through our legal system. Much of what goes into answering this question will depend on certain facts. We want to know how an action came about, what the person’s motivation for performing the action was, and whether anything in their environment prevented them from doing otherwise. We also want to know what the person was capable of at the time of action. Children and the mentally disabled are often excused for their actions precisely because they are unable to perform certain actions or recognise certain considerations. But even once all these facts are known, questions can still linger about what exactly it is that makes someone responsible.

Answering this question requires a philosophical account of moral responsibility. We have strong folk intuitions regarding what makes agents praiseworthy or blameworthy. The fact that a wrongdoer ‘should have known better’ or ‘could have done otherwise’ often makes us feel licensed in treating them negatively and demanding an apology. Such assessments rely on a series of background assumptions regarding what sorts of things agents are capable of doing, how we can expect people to behave, and how difficult it is to do the right thing in a particular setting. These assessments also rely on assumptions about what it is we do when we blame, and under what circumstances we are entitled to blame.

From the 1960s onwards, major challenges to our folk assumptions about human behaviour began to emerge. A woman named Kitty Genovese was famously murdered in a square surrounded by apartment towers, and yet nobody called the police after hearing her

---

1 ‘Joining the Hunt’ is the title of Chapter 1 of Doris’s (2002) *Lack of Character*. 
cries for help.² Stanley Milgram, a psychologist at Yale, demonstrated that 65% of people will electrocute a stranger into heart failure when ordered to by a scientist (Milgram, 1963, 1974). Phillip Zimbardo, a psychologist at Stanford, studied the behaviour of statistically normal university students who had been assigned to the roles of prisoners or guards in a mock prison (Haney, Banks, & Zimbardo, 1973, hereafter referred to as ‘Stanford Prison Experiment’). The study was shut down after only six days because the ‘guards’ began treating the ‘prisoners’ in ways that bordered on torture.

These cases all defied both commonsense and expert predictions on how people will behave in given situations. They are only the most famous examples of a large raft of research into what factors influence our moral behaviour. The situationist studies are psychological experiments which demonstrate that seemingly irrelevant environmental conditions (hereafter ‘situationist factors’) can greatly influence whether people perform morally wrong actions. These actions strike us as being ‘out of character’ and conflict with our folk predictions of how people will act, the reasons on which those people will act, and what kinds of factors actually influence decision-making.

For example, while we often that think prosocial behaviour is primarily the product of how caring or charitable someone is, Isen and Leven (1972, hereafter referred to as ‘Dime’) examined how often people who had just exited a phonebooth would help someone who dropped a stack of papers. In the control condition, only 4% of subjects helped the confederate. In contrast, 87.5% of subjects in the experimental condition, who found a dime in the phonebooth, provided assistance. Though picking up papers is a relatively minor form of altruistic behaviour, the size of the change, and the ease with which this was achieved, is very startling.

Another group of situationist experiments show that helping behaviour can easily be inhibited, even in situations where the appropriate action seems obvious. Darley and Batson (1973, hereafter referred to as ‘Samaritan’) asked seminarians to give a speech on the Christian

² The Genovese case is treated as a canonical starting point in the psychological literature on bystander effects, so I have followed suit citing it here. However, as has been noted before by Alfano (2013, p. 41), the details of this case are disputed, and there is good reason to question the most commonly offered characterisation of events. However, there are numerous other cases with the same relevant features i.e. a large number of people failing to help despite being able to do so with ease. For the original story in The New York Times, see <Author Unknown> (1964). For reasons to be sceptical of the standard story, see Gallo (2015). For relevantly similar, but better-documented cases, see Wines (2011), Harmon (1998) and Livingston (2010).
parable of The Good Samaritan. Upon arriving, the seminarians were told their scheduled talk had been moved elsewhere. On their way to the second location, the seminarians passed by someone who appeared to be unconscious in the middle of the day. It was found that the chances of any seminarian stopping to help fell dramatically if they were told they were running slightly late (10%), compared to if they were told they were early (63%).

It has also been found in a series of related experiments (hereafter collectively referred to as ‘Bystander’) that an agent’s likelihood of helping a stranger can be reduced significantly by the presence of other people. For example, when by themselves, 70% of subjects will check on someone in another room who is heard to cry out in pain after a large crashing sound. But if unresponsive confederates are present, only 7% will help (Latané & Rodin, 1969). Though such results are sometimes interpreted as showing that people are simply more selfish than we expect, self-interested behaviour is also easily affected by situational factors. Most of us intuitively think that if we saw smoke billowing into the room we were in and heard the fire alarm go off, we would take this as a sure sign there is a fire, that we are in danger and that it would be in our own self-interest to leave the room. And indeed, Latané and Darley (1970, hereafter referred to as ‘Fire’) did find that 75% of subjects in a room by themselves left within four minutes of smoke beginning to billow in. However, when in the presence of two confederates who ignored the smoke, 90% of subjects remained in the room. Had the smoke been coming from a real building fire, these subjects would have remained in the room long enough to die of asphyxiation.3

More worryingly, not only can situationist factors decrease helping behaviour, they can also play a dramatic role in encouraging people to actively harm others too. Milgram’s (1974) famous ‘Obedience to Authority’ experiments (hereafter referred to as ‘Obedience’) had subjects administer shocks to a stranger as part of what was purportedly a memory task. The subjects or ‘teachers’ would administer increasing shocks to a ‘learner’ (actually a confederate who wasn’t receiving shocks) every time the learner got an answer wrong. It was observed that even when the learner demanded to be let out, acted as if they’d had a heart attack, and then ceased responding altogether, 65% of subjects continued to administer the dangerous shocks when directed by the experimenter.

Even today, most people are surprised at how easily and quickly subjects’ behaviour is influenced by the seemingly irrelevant environmental factors in these experiments. Many, like Milgram, appreciate these studies’ relevance for understanding how mass atrocities can be committed by ordinary agents. Only some entertain the idea that they too would have acted wrongly in these experiments, and even fewer consider the idea that perhaps they also could be culpable for a moral atrocity in the right setting.

Psychologists examining these experiments have tended to focus on questions regarding the validity or generalisability of the results. One of the rare attempts to instead understand their relation to moral responsibility comes from Zimbardo, in his expert testimony provided at the trial of soldiers involved in the Abu Ghraib prison scandal (Zimbardo, 2007). Drawing from his work in the Stanford Prison Experiment, Zimbardo argued the actions of the soldiers were a product of the environment they were in. This environment was in turn a product of the systems put in place by the military and the United States Government, and so these institutions were morally responsible. However, he also claimed that “the situational approach is not excuseology” [sic] (Court Martial of Ivan L. Frederick, II, 2004, p. 576) and that the soldiers were “guilty … but not responsible” (ibid., p. 553–554). His attempt to “reconcile [these explanations] with common-sense notions of personal responsibility” has been described as “a fundamental weakness” by other social psychologists (Mastroianni, 2013, p. 58) and needs to be clarified and developed in more detail.

Because Zimbardo isn’t a philosopher, it can be tempting to dismiss his claims as the result of confusion about certain concepts and what it means to be morally responsible. But I believe he is rightly drawing our attention to an area that deserves philosophical analysis. That these experiments conflict with our folk predictions of how people will act and what motivates them to act is important, because it suggests many of our folk assumptions about what can reasonably be expected of people may be mistaken. In particular, the fact that most people we consider to be good-natured could easily be led to commit wrongdoing suggests

---

4 Though Milgram would often overstate the causes of this, arguing we have an “inborn tendency to obey” and that the situation can put us into an “agentic state” in which we are totally passive to others’ orders (Milgram, 1974, pp. 123–124). Such an account did not explain how those who did disobey managed to do so (Perry, 2013a, p. 254) and his theorising is recognised to be “the weakest part of the book” (Blass, 2004, p. 216).

5 E.g. Griggs (2014) presents a summary of criticisms on the external and internal validity of the Stanford Prison Experiment and whether it ought to be included in university psychology textbooks, and Perry (2013a) argues Obedience is not an externally valid experiment.
that we may regularly misjudge how difficult it is to avoid wrongdoing. And, as noted earlier, assumptions like this form the background from which we judge whether agents are blameworthy, and whether we are licensed to respond to them in certain ways.

ii. Preliminary Concerns and Caveats

Given this thesis examines experiments from social psychology, some readers may have certain worries regarding the evidence upon which I’m basing my analysis, and these should be acknowledged. Many readers of this thesis are probably familiar with what has been dubbed the ‘replication crisis’. For those unfamiliar, this refers to an ongoing phenomenon in science where many researchers have been unable to replicate the results of earlier well-known studies (Earp & Trafimow, 2015). For example, one very famous study claimed to have found that when subjects were primed with words associated with the elderly such as ‘retirement’ or ‘wrinkled’, those subjects would then walk slower down a hallway (Bargh, Chen, & Burrows, 1996). This study initially appeared to be replicated by other researchers (Cesario, Plaks, & Higgins, 2006) and was a central case study for examination by Doris in Lack of Character. Unfortunately, later and more carefully run experiments failed to replicate this result (Doyen, Klein, Pichon, & Cleeremans, 2012), as have a large number of experiments purporting to show social behaviour can be substantially influenced by priming (e.g. Harris, Coburn, Rohrer, & Pashler, 2013; Pashler, Coburn, & Harris, 2012).6

There are multiple reasons for why so many positive results have been thought to be valid, only to later be found erroneous. These include p-hacking, journal bias towards publishing positive results, the file-drawer problem, and perverse incentives (Goldacre, 2014; Ioannidis, 2005, 2012). Social psychology has served as the face of this crisis and received much derision in popular media (e.g. Dominus, 2017; Engber, 2016). This is not completely unwarranted, as one team of researchers famously reported being able to replicate only 39 of 100 studies from top psychology journals (Open Science Collaboration, 2015). Additionally, of those studies that did replicate, effect sizes were consistently much smaller.

Because this replication crisis is still ongoing, some readers may worry that the studies I cite are not valid, or that we should adopt a general scepticism towards the results of social

---

6 Studies showing priming leads to faster mental associations or responses are well-replicated. It is only the studies examining effects on behaviour (e.g. walking) or our judgements (e.g. liking someone more) that are failing to replicate.
psychology experiments. While the replication crisis is a very important problem that we ought to keep in mind, I think readers should not be too wary of my claims in this thesis for a few reasons.

First, failures to replicate earlier results are (troublingly) occurring in many areas of science (e.g. Freedman, Cockburn, & Simcoe, 2015; Prinz, Schlange, & Asadullah, 2011), so scepticism of social psychology results alone is certainly unwarranted. Second, several notable failures to replicate results or evidence of scientific malpractice occurred while I was writing this thesis (e.g. Open Science Collaboration, 2015; Perry, 2018), leading to me to remove studies I had originally intended to use as evidence for some claims (e.g. Sherif, Harvey, White, Hood, & Sherif, 1961; Vohs & Schooler, 2008). The fact that I’m aware of the replication crisis, have acknowledged its existence here, am familiar with its underlying causes and have identified some studies which are unreliable should improve readers’ confidence that the studies I have retained have not been included in an epistemically negligent manner. I don’t have the space to engage in a literature review of each kind of study I refer to, but I have made every effort to ensure each remaining study I do refer to is reasonably well-supported. Third, the experiment which takes up much of my analysis, Obedience, is very well-replicated using subjects from a variety of cultures, professions and time periods (Burger, 2009; Doliński et al., 2017; Edwards, Franks, Friedgood, Lobban, & Mackay, 1969; Shanab & Yahya, 1978), so this should justify some further confidence in my arguments. I concede there is always the possibility I have missed some research with conflicting findings, or that future research will demonstrate that these studies are not valid, but this possibility is unavoidable. Readers will have to assume, given the evidence currently available, that the empirical studies cited are valid and accurate representations of agent behaviour.

Another concern stems from the fact that much of my argument relies on making inferences about subjects’ mental states which are not directly observable and which should perhaps be investigated through empirical means. While this isn’t necessarily a worry when doing conceptual analysis of beliefs, desires, attitudes, and deliberative processes, readers may worry about how confident we can be when making inferences about particular subjects’ beliefs, desires and attitudes during experiments, some of which were carried out decades

---

7 Doris (2002) notes that replications of Isen & Levin’s (1972) Dime study have had mixed results (p. 30, fn. 4), but he chose to include it because it is highly representative of an established pattern of results. Given this study has been referenced by many philosophers writing on the situationist experiments, I will do the same.
ago. While this is a problem, it is one present in nearly any attempt to explain someone else’s behaviour, or any attempt to assess an agent’s mental states at the time of their wrongdoing. I will follow Doris’s (2002) invocation of Winston Churchill’s comment on democracy being the worst form of government: “Experimental psychology is perhaps the worst available method for understanding human life. Except, I hasten to add, for all the other methods” (p. 8). This is not to deride those other methods, such as intuition pumps or a priori reasoning. Rather, I mean that insofar as we’re trying to explain a behaviour or think about a particular subject’s mental states, we can only make the best of what tools we have available. The situationist experiments themselves are strong evidence that our intuitions about other people’s behaviour are not always reliable, so it isn’t clear the staunch sceptic of moral psychology or social psychology would be better equipped to answer my questions. My goal in this thesis is to argue for what claims we presently have the most reason to believe, given our evidence, and to examine what implications the results of these experiments have for our normative theorising.

iii. Précis
This thesis asks: Are agents who act wrongly in situationist-like scenarios morally responsible for their actions, and what kinds of responses are we justified in having towards them? In order to answer this question, a number of concepts need to be unpacked, and a number of related sub-questions need to be addressed: What is it exactly that the situationist experiments demonstrate? What does it mean to say that an agent’s choices were ‘influenced’ by a situation? Why do some agents act wrongly, while others act rightly in these situations? And what does it mean to be morally responsible for an action?

Chapter 1 is a review of theories of moral responsibility, and introduces a number of important concepts and distinctions. Taking Strawson’s (1962) Freedom and Resentment as a canonical starting point, I introduce readers to the reactive attitudes. These are sentiments which play a crucial role in our moral responsibility practices, and which are taken by many philosophers to be reliable, though not infallible, markers of when an agent is morally responsible. I identify that in asking whether subjects who act wrongly in the situationist experiments are morally responsible for their actions, I am interested in whether those subjects are deserving of blame. I adopt Nelkin’s (2011) ‘rational abilities’ account, which argues that in order to be morally responsible for their actions, agents require the capacity
and opportunity to do the right thing for the right reasons. This account ties nicely with my investigation, because it isn’t clear whether subjects in the situationist experiments failed to avoid wrongdoing because they simply didn’t exercise their capacity to avoid wrongdoing, or because they had a reduced capacity to avoid wrongdoing.

In Chapter 2, I examine a range of previously offered explanations for why so many subjects act wrongly in the situationist experiments, focusing on Obedience. Most of these explanations come from philosophers and psychologists trying to refute the argument that these experiments show that global character traits do not exist. These explanations implicitly assume that the subjects in these experiments had the capacity to do the right thing for the right reasons, but failed to exercise it. The explanations considered include (1) the subjects experienced very strong desires to obey the experimenter which outweighed their desire to disobey, (2) the subjects acted from morally bad character traits which competed with their good character traits (e.g. selfishness rather than compassion), (3) the subjects were akratic, weak-willed, or lacked practical wisdom, (4) the subjects identified with the experimenters and their goals, and (5) the subjects’ fear of embarrassment inhibited them from disobeying. I argue that each of these proposed explanations contain important gaps. To understand why so many subjects commit wrongdoing, we ought to consider their perceived reasons for acting as they do, and their capacity to recognise moral reasons.

In Chapter 3, I consider two common means by which philosophers assess when an agent lacks a particular capacity in a particular situation. The first is through use of the reasonable person standard, which looks at how most reasonable people would behave in that situation. The second is by appealing to counterfactuals, which investigates how that agent would have acted in relevantly similar settings. I argue that both of these methods face problems. We cannot use the reasonable person standard without antecedently knowing which reference class of people to apply it to, and the most commonly proposed counterfactuals make some contentious assumptions about what kinds of factors do or do not impede capacities. I argue that instead of trying to apply a particular test to decide whether the subjects in the situationist experiments had a reduced capacity to avoid wrongdoing, we should think of the claim that subjects had a reduced capacity to avoid wrongdoing as a hypothesis which can be more or less plausible. This hypothesis should be assessed against the competing hypothesis that subjects had this capacity, but simply didn’t exercise it. I argue that if a particular factor causes agents to regularly fail at ϕ-ing in a range of circumstances when we have reason to
believe they are trying their best to $\phi$, this is good evidence that that factor is reducing their capacity to $\phi$.

I then consider two examples of agents who have a reduced capacity to perceive or act on moral reasons. The first type of agents are addicts, who I consider as a case study in the different ways that agents can have more or less of the capacity to perceive moral reasons. The second are agents who experience goal neglect. This acts as a case study in how agents can be prevented from exercising a certain capacity despite successfully exercising that capacity in many very similar settings.

In Chapter 4, I turn to examining the capacities of the subjects who act wrongly in the situationist experiments. I argue that the best explanation of subject behaviour in Dime is that finding a dime puts subjects into a good mood, broadening their attention. I argue that the subjects who failed to help in this experiment are blameworthy, because there is no evidence to suggest that without finding a dime, their attentional capacities were so low as to warrant being excused.

I then argue that subjects in Bystander are fully excused for their actions. By looking at a number of studies examining which factors cause subjects to help or not help, I argue that the best explanation of subjects’ failure to help is that they did not interpret the confederate as someone who was in need of help due to the ambiguity of the situation.

The subjects in Samaritan form a heterogeneous group. Based on the transcripts of subjects’ reactions to hearing about the confederate, it seems that some subjects did not notice the unconscious confederate at all, or interpreted him as not being in need of aid. These subjects are excused, similar to the subjects in Bystander. Some subjects did notice the confederate, but failed to help. I argue that some subjects plausibly experienced something like goal neglect, which could excuse them. However, some subjects may have simply experienced weakness of will, as evidenced by the fact that they plausibly could have turned around after initially walking by. These subjects will be somewhat blameworthy.

Finally, I consider Obedience, and argue that most subjects who obey in this experiment are excused for their actions. A number of factors in Obedience made what was happening far more ambiguous and confusing to subjects than we expect, which made it difficult for them to realise that disobeying was something they could do. Subjects who did disobey seemed to do so by triggering a gestalt switch in their construal of the situation, which then enabled them to form an intention to stop administering the shocks.
In Chapter 5, I consider an objection to my argument that the obedient subjects in Milgram’s experiments are excused for their actions. The objection is that the learner in this situation seems entitled to blame the obedient subjects, and excusing the obedient subjects seems unfair to victims. In response, I argue that excused wrongdoers can incur duties of reconciliation towards those they have wronged. Failure to take up these duties displays a lack of regard for the victims, and can make the wrongdoer deserving of blame. By adopting Radzik’s (2009) account of reconciliation, I argue that excused wrongdoers also ought to demonstrate a commitment to morality, make reparations for any harms they have caused, and apologise to their victims in such a way that withdraws any insult that was expressed in performing the wrong. On this account, excused wrongs can express an insult because wrongs have a social meaning, even if the wrongdoer did not meet the conditions necessary for responsible agency at the time of action. More importantly, if the excused wrongdoer is aware they have the power to withdraw the insult, but fails to do so, this is something which itself expresses a lack of regard for the victim and can make the wrongdoer deserving of blame. How the excused wrongdoer reacts to the victim after the original wrong is itself something that has a social meaning.

In Chapter 6, I consider whether we are entitled to blame the obedient subjects even if they are, in fact, blameworthy for their actions. This chapter breaks from previous chapters by examining what follows if we take the obedient subjects to be morally responsible and thus blameworthy for their actions, in virtue of having had the capacity to avoid wrongdoing. I investigate whether we in the moral community have the standing to blame the obedient subjects, given we likely would have acted the same way they did.

I adopt and develop Todd’s (2019) account of standing to blame, which argues that agents lack the standing to blame a wrongdoer for \(\phi\)-ing when they are not sufficiently committed to the kinds of values which would condemn \(\phi\)-ing. This account entails that many agents who would commit a particular wrong in certain settings also lack the standing to blame, even if they haven’t in fact acted wrongly. This is because the fact they would commit that wrong can show they are not sufficiently committed to the relevant values. I argue that, since Obedience has been very well-replicated using subjects from a number of cultures and time periods, at least 80% of the general population would also commit wrongdoing in Obedience. Given we are stipulating that all these agents would have the capacity to disobey, and given the egregious nature of the wrong, this is diagnostic of a lack of sufficient commitment to
the relevant values, in the same way that a single murder can be evidence of a lack of respect for human life. At least 80% of us thus lack the standing to blame the obedient subjects.

In Chapter 7, I examine why agents who are not sufficiently committed to the kinds of values which would condemn ϕ-ing lack the standing to blame agents who ϕ. I argue that existing uses of the term ‘standing’ are very vague, and attempts to understand the term by thinking about entitlements or appropriateness fail to explain why hypocrisy could ever undermine standing. I argue that we should instead think of standing to blame as a status which grants agents the ability to exercise a normative power. This is analogous to the way in which agents simply cannot consent, promise, or forgive on behalf of most other agents without first meeting certain conditions. On my account, the claim that hypocrites lack the standing to blame is slightly misleading. While hypocrites can blame, they cannot exercise the normative power agents normally exercise in blaming, which is the addressing of valid demands. I argue that agents must have a certain authority in order to make legitimate demands of others. To explain when agents have or lack this authority, I adopt Darwall’s (2006) account of second-personal authority, which argues that demands must be made from the perspective of a free and rational agents amongst equals to be legitimate. However, hypocrites’ prior wrongdoing shows that they do not accept others’ demands of them, violating this requirement.
Chapter 1: A Brief History of Moral Responsibility

1.1 Strawson and Our Moral Practices

Historically, philosophers working in the analytic tradition thought that in order to determine whether agents were morally responsible for their actions, we first had to work out which conditions people need to meet to qualify as agents who act. This meant focusing on questions regarding the metaphysics of agency and free will, determinism, compatibilism, sourcehood and the ability to do otherwise.\(^8\) The canonical shift from thinking about moral responsibility in this way came from Strawson’s (1962) *Freedom and Resentment*. Strawson argued that we don’t need to first settle whether determinism was true before being able to decide whether people could be morally responsible for their actions. Because our moral practices are an inescapable part of our human life, the solution to understanding the nature of responsibility is to instead focus on these practices rather than antecedent ‘panicky metaphysics’ (p. 79). Basically, even if determinism was true and incompatible with free will, this would make no difference to our actual practices of holding one another responsible.

Strawson argued that moral “responsibility needs to be understood in the context of a central range of sentiments that we are subject to in our dealings with each other” (Wallace, 1996, p. 7) which he called ‘reactive attitudes’. Examples of reactive attitudes include resentment, gratitude, indignation, and guilt. These reactive attitudes are not merely emotions we passively happen to feel in response to others’ actions, or tools we deploy to regulate the behaviour of wrongdoers.\(^9\) Rather, they form a necessary part of our relationships and interactions with others. They are the means by which we hold one another morally responsible for our actions.

---

\(^8\) See Levy and McKenna (2009) for a review of recent research in these areas.

\(^9\) Strawson’s argument was so influential in part because it offered a way to reconcile a debate between ‘optimists’, who thought that agents do meet the conditions required for moral responsibility because our practices are justified by their ability to regulate behaviour in socially desirable ways, and ‘pessimists’, who thought that we don’t meet the condition required for moral responsibility because we are not the ultimate source of our actions. Focusing on our actual practices reconciled these positions because it seems clear that we don’t blame people merely to regulate behaviour, but also that in blaming and praising we aren’t responding to facts regarding libertarian free will (since we don’t know if such a thing exists). See McGeer (2014) for a summary and analysis of how Strawson’s approach has particular implications for thinking about normative demands and agential capacities.
Strawson argued that moral responsibility should not be understood as a set of conditions that an agent needs to meet before we can hold them responsible. Rather, our practices themselves are to be understood as constitutive of moral responsibility itself. Watson summarises this neatly: “It is not that we hold people responsible because they are responsible; rather, the idea (our idea) that we are responsible is to be understood by the practice, which itself is not a matter of holding some propositions to be true, but of expressing our concerns and demands about our treatment of one another” (2004, p. 222). On this account, to be morally responsible for an action just is to meet the conditions under which one would be responded to with the reactive attitudes by one’s moral community.

While many philosophers recognise Strawson’s insight and the value of his approach, they also accept his specific account requires revision, as our actual practices are sometimes grossly mistaken. Consider a prejudiced community that ‘holds responsible’ and blames people with severe mental disabilities and egalitarians: two groups of people who fail to adhere to the norms of the community. Fischer and Ravizza (1993) point out that under Strawson’s account, the egalitarians and mentally disabled agents are, in fact, blameworthy for their actions. While most of us do intuitively think that egalitarians are morally responsible for their words and beliefs, we disagree with the content of this racist community’s assessments and reactive attitudes. Egalitarians are to be praised, not blamed or resented. And with regard to the severely mentally disabled, we are strongly inclined to believe they are not morally responsible for their actions at all. Strawson’s emphasis on our actual moral responsibility practices is at odds with our intuitions regarding not just when an agent is blameworthy, but what it means to be morally responsible for an action. In short, we seem to think there is a difference between being held morally responsible for an action, and actually being morally responsible for an action.

Instead of focusing on the tendency of an act to evoke our reactive attitudes when determining responsibility, the most popular modification to Strawson’s approach has been to focus on the appropriateness of our reactive attitudes when holding someone responsible (e.g. Fischer & Ravizza, 1998; Nelkin, 2016a; Wallace, 1996). This appropriateness condition
has, in turn, been understood in two main ways. The first approach, in line with our intuitions about the prejudiced community described above, takes facts about whether agents are morally responsible to be independent of whether they are actually held responsible, or would be held responsible according to our current moral practices. This approach argues that, since our practices can be mistaken, any account of moral responsibility requires external justification (e.g., Fischer & Ravizza, 1998; Levy, 2011b; Nelkin, 2016a). I will call this a ‘response-independent’ approach to moral responsibility. On this approach, to determine whether an agent is blameworthy, we need to appeal to considerations independent of our actual practices. Examples of plausible considerations include whether it would be fair to blame the agent, whether the agent had libertarian free will, or whether the agent’s actions express a lack of regard. Philosophers working in this tradition do not discount entirely our actual tendency to deploy the reactive attitudes, however. They take our reactive attitudes to be reliable, but fallible, epistemic markers of moral responsibility. We can appeal to our reactive attitudes and moral practices as defeasible evidence when trying to assess whether particular kinds of agents are morally responsible for their actions.

The other group of philosophers resist this appeal to considerations which are independent of our moral practices, because this threatens a return to the original debates on the metaphysics of agency. Those original debates have been perceived to be in a “dialectical stalemate” (Fischer, 1994, p. 83), and dissatisfaction with this approach was what motivated appealing to our actual moral practices in the first place. As a result, the second way of understanding this appropriateness condition tries to retain something from Strawson’s approach, while avoiding saying that egalitarians who are blamed by racist communities are,
in fact, blameworthy. Call this a ‘response-dependent’ approach to moral responsibility. This approach sees our moral practices and reactive attitudes as somewhat partially constitutive of what it is to be morally responsible. But how to understand this ‘partially constitutive’ condition has until recently been quite unclear.\(^{11}\)

The strongest defence of this approach comes from Shoemaker (2017), who makes an analogy with the funny. Shoemaker points out that if a community were to find genocide funny, this would not thereby make genocide funny. Genocide is not worthy of laughter or amusement. However, it would be a mistake to conclude from this fact that what is amusement-worthy is determined by some Platonic form of amusement, or a set of facts about what is ‘really’ amusing. There simply are not any properties that funny things all share, other than their tendency to evoke laughter in us. The funny \textit{just is} what (properly trained) human sensibilities generally find amusing, and this is compatible with there being wide cultural variation regarding what sorts of things count as funny. Likewise, Shoemaker argues that ‘the blameworthy’ should be understood to just be the kinds of things that properly trained human sensibilities find worthy of blame.

In favour of this approach, Shoemaker argues that response-independent approaches to moral responsibility face many objections. Their main problem is that they end up being too revisionary, either blaming agents we usually don’t think are intuitively blameworthy, or failing to blame agents who we are strongly inclined to blame. Though I agree with Shoemaker that a response-dependent account of moral responsibility can avoid many of these counterexamples, in this thesis I will adopt a response-independent approach for a few reasons. First, not many philosophers follow Shoemaker in endorsing a response-dependent account of moral responsibility. Second, it is too soon to tell whether those who do take such an approach agree with Shoemaker’s analogy to the funny, as previous presentations of this ‘partially constitutive’ condition have been rather unclear.\(^{12}\) Third, if one were to adopt an account like Shoemaker’s, our investigation would reduce “to a matter of investigating the fittingness conditions of anger” (p. 521) which is rather unrelated to the kinds of questions I am interested in.

\(^{11}\) Todd (2016) considers attempts to explain what this condition amounts to by examining previous characterisations from Watson (2004, 2014), Kane (2005), Brink and Nelkin (2013), Coates and Tognazzini (2013a), Tognazzini (2013) and Wallace (1996), and argues all of them fail once spelled out more clearly. Shoemaker’s (2017) notable defence is a reply to Todd’s argument.

\(^{12}\) Ibid.
The main upshot of adopting a response-independent account is that advocates of such accounts are more willing to be somewhat revisionary regarding which kinds of agents count as morally responsible for their actions. On this approach, our reactive attitudes are thought of as epistemic markers, which reliably, but not infallibly, respond to facts about agents. Advocates of these accounts can still accept that, all else being equal, the less revisionary a theory of moral responsibility the better. In instances where we observe a conflict between our considered moral practices and the judgements of culpability delivered by a theory, that theory should provide some explanation for why we ought to revise our practices.

Assessing whether people in the situationist experiments are morally responsible for their actions will thus depend on whether they are appropriate targets of the reactive attitudes. This requires investigating what exactly we target with our reactive attitudes, and how these targets must be connected to the agent in question for the agent to be blameworthy. In the situationist experiments, there are a variety of psychological features that may connect to the subjects’ actions in some way and which could thus be worthy of resentment. Subjects may recognise moral reasons but fail to act on them, express certain judgements about the interests of other people, or display a lack of good will. Identifying what we target with the reactive attitudes and why we blame will allow us to get clearer about whether these subjects meet the conditions required for moral responsibility.

1.2 Quality of Will

When we judge someone to be morally responsible, or we hold someone morally responsible, we take them to be morally responsible for something. It initially seems like agents are responsible for their actions, but this isn’t quite right. Two agents can perform the same action, yet one agent may be excused for doing so. Something thus needs to connect an agent and their actions in order to make them a fitting target of our reactive attitudes.

Strawson (1962) argued that when we direct reactive attitudes such as resentment or gratitude towards other agents, these are in response to the agents’ quality of will. As evidence of this, Strawson identified two common instances where our initial judgement that someone is blameworthy disappears. If I tread hard on your foot, you feel some anger towards me, which is, all else being equal, appropriate. However, if I tell you that I’ve got an excuse (someone pushed me) or a justification (there was a dangerous insect about to sting you), your anger subsides, because you recognise I did not have any deficient quality of will.
towards you. There was no intended harm or disrespect displayed in my action and, in the case of the dangerous insect, my action actually displayed good will. As a result, I am not morally blameworthy for my action. This difference is nicely put by Watson (2004): “In general, an excuse shows that one was not to blame, whereas a justification shows that one was not to blame” (emphasis in original, p. 224). In addition to being excused or justified, Strawson argued there was a third way agents could be inappropriate targets of resentment for their actions. This is by being exempt from our moral responsibility practices, due to being incapable of understanding those practices sufficiently to take part in them. Exempt agents include very young children and the severely mentally disabled.

Recall we are considering what makes the deployment of a reactive attitude appropriate, and thus what grounds moral responsibility. Strawson’s emphasis on quality of will has inspired a number of related accounts. Watson (1975, 1987) argues that quality of will can’t be what we target when we blame or praise, because we excuse people who say rude things under great stress, even though they display deficient quality of will. We also exempt children when they manifest ill will towards us. Watson argues that the reason we refrain from blaming these agents is not because they didn’t display a bad quality of will, but because the explanation of their rudeness (being stressed, or only five years old) does not disclose anything worth resenting about themselves qua moral agent. What matters for our assessments is the agent’s valuational system. This is because their valuational system discloses something important to us about them as an agent, namely how they see others. This approach in turn inspired a number of similar accounts, with philosophers locating what makes someone blameworthy for an action in the way that action connects to something about how the agent is constituted. Such accounts were originally dubbed ‘Real Self’ accounts.

Arpaly and Schroeder (1999), for instance, argue for a ‘whole self’ view, in which an agent is blameworthy or praiseworthy for an action to the extent that the desires and beliefs underlying that action are integrated within her overall personality. This account is developed in response to inverse akrasia cases, where an agent acts against their conscious judgements about what they should do, but thereby ends up doing the right thing for the right reasons. These cases are argued to pose a problem for Watson’s account, because the agent is acting against their consciously endorsed judgements and what they value, yet they seem to be praiseworthy for doing so.

McKenna (2012) and Shoemaker (2015) build on Strawson’s original account, and argue that ‘quality of will’ is best understood as the value or worth of one’s regard for another
person. ‘Will’ does not designate some faculty or ability as in the phrase ‘free will’, but the regard one shows others in light of relevant moral considerations. Similarly, ‘quality’ refers not only to value in the sense of a will having a positive or negative valence (e.g. good will or ill will), but includes the worth with which someone acts towards another person. To make this clear, imagine a case in which a father decides to go drinking at the bar instead of to his daughter’s recital. Despite not actually showing her any negative or ill will, his actions are appropriate targets of the reactive attitudes due to his lack of regard for her.13

Talbert (2008, 2009b, 2012, 2013) endorses a position called ‘Attributionism’. This is something like an updated quality of will account, as it maintains agents are blameworthy if their actions express an objectionable evaluative judgement about what moral reasons are present. If I back up my car over you, but couldn’t have known you were there, then my actions can’t be said to express any objectionable judgement about you. In contrast, if I know you’re behind my car, and back it up anyway because I simply don’t care about your well-being, then this does express an objectionable evaluative judgement, for which I can be blamed. The reason Talbert calls this position Attributionism is because, in order for my actions to express an objectionable evaluative judgement, the action needs to be connected (i.e. attributable) to me in a particular way. If I am aware of my actions, but my backing up the car is caused by a literally irresistible desire, or coercion, or an epileptic fit, then my actions are not attributable to me in the right way. I am thus not blameworthy. The central objection to this account of moral responsibility is that some of its assessments seem unfair, as some wrongs can be attributable to agents despite the agent lacking moral knowledge of what they are doing, or a certain kind of control over their actions. If an agent is manipulated by neuroscientists to commit a wrong action, Talbert argues so long as they still qualify as an agent who can make assessments about reasons and guide their actions by those assessments, that agent is blameworthy (Talbert, 2009a).

13 In response to Watson’s objection that children can express ill will, McKenna (2012) argues children aren’t considered responsible even though they can show a lack of regard in their actions because they do not understand the meaning of their actions, and cannot engage in moral address. Their wrong actions thus do not express disregard or lack of regard in the same way as adults (p. 75–77). It should be noted that McKenna argues for a ‘conversational’ account of responsibility, which isn’t strictly Attributionist in nature. According to McKenna, our actions are bearers of meaning, and this meaning is a function of the quality of will with which we act, but agents also need certain capacities to understand the meaning of their actions. The main objections to this account are that it seems hard to point to any moral conversation when blaming agents who are dead, or when we only privately blame (Driver, 2016; Vargas, 2016).
For consistency with the nomenclature in the current literature, I will follow Levy (2008) and loosely refer to all these approaches that treat quality of will, degree of integration, evaluative judgements, or an agent’s valuational system as central to moral responsibility as ‘Attributionist’ approaches. This is because they all identify moral responsibility with whether an act is attributable to an agent in a certain way. This is in contrast to Control accounts (examined below in §1.5), which identify moral responsibility with whether an agent had a certain kind of control over their actions.

1.3 Faces of Responsibility
Philosophers generally agree that our moral responsibility assessments target more than just something in an agent’s actions. Watson (1996) made a now commonplace distinction between two ‘faces’ of responsibility: responsibility as accountability and responsibility as attributability. Accountability responsibility is what folk usage of the term ‘moral responsibility’ and the discussion thus far picks out. To say an agent is accountability responsible is to say they are connected to their actions in such a way as to make them an appropriate target of praise and blame, depending on the moral character of their actions. If an accountability responsible agent commits a wrong which is severe, we are licensed to respond to them with sanctions and punishments.

Watson argued that we often also target people with reactive attitudes for things they did not choose, when no ill will or regard is displayed, and even in cases in which they haven’t actually acted at all. We can direct reactive attitudes such as resentment towards people’s character or judgements, even if we would think it inappropriate to hold them accountable for their character or judgements in virtue of their not having performed any action. In contrast to accountability responsibility, the attributability face of responsibility constitutes “appraisal of the individual as an adopter of ends” (Watson, 1996, p. 229). Paradigmatic examples include a psychopath who never harms anyone but may fantasise for hours about doing so, or the mother who has thoughts about drowning her screaming baby (in contrast to a mother who never has such thoughts enter her mind at all). We may feel some initial temptation to blame these agents and hold them accountable, but there are no actions for which to hold them to account. The psychopath and the stressed mother haven’t harmed anyone, or displayed any ill will or lack of regard in their actions which we could call on them to either justify or apologise for, because they haven’t acted at all. Nevertheless, we feel some
reactive attitudes are appropriate, even if it would be inappropriate to blame them. The harmless psychopath is cruel and contemptible, and differs in an important way from the stranger next to him who has also never harmed anyone but doesn’t fantasise about doing so. While blame may be too strong a response to the fantasising mother (no doubt some pity is warranted), she too differs in an important way to the mother who has no such thoughts.

The attributability face of responsibility helps us make sense of our reactions towards agents who commit egregious wrongdoing, but who also seem to be excused for this. Watson (2004) uses the example of Robert Harris, a serial killer who brutally tortured and raped his victims, and whom we initially feel is a fitting target of the reactive attitudes for his actions. Harris definitely expressed an ill will, lack of regard, and objectionable evaluative judgements towards his victims and so, at first glance, appears to be blameworthy. However, Harris had a horrific childhood, which can make blaming him seem unfair. He suffered years of abuse from his parents, and it is not implausible to think this history acts as an excusing condition. Watson’s point was that, even if we think Harris ought to be completely excused for his actions, this would not make responsibility assessments unfitting altogether. It is a fact that Harris’s character is cruel and sadistic, and it is appropriate for us to adopt some negative attitude which reflects this. The distinction between the attributability and accountability faces of responsibility helps make sense of our conflicting intuitions.

It is worth noting that Watson took his Real Self account to be concerned primarily with identifying the conditions under which an agent was attributability responsible for an action, as he took this kind of moral responsibility to be deeper and more important than accountability responsibility. However, most philosophers have taken the conditions required for accountability responsibility to be the central question worth investigating. While other Attributionists generally agree with Watson that attributability isn’t shallow, they diverge from his account in arguing (roughly) that attributability is sufficient for accountability. That is, if a wrong action is attributable to you in the right way, then you are blameworthy for that action. On these accounts, Harris is, in fact, a blameworthy agent.

---

14 Wolf (1990) notably derided the attributability face as “shallow”, arguing it doesn’t add much other than to say someone is “a bad act-maker”. Watson (1996) replied that emphasis on blameworthiness alone leaves out much of the interpersonal manner by which we hold people morally responsible, and that responding to a wrongdoer’s attributability responsibility captures something important even if we don’t engage in blame.
In this thesis, I follow the majority of philosophers in taking accountability responsibility to be the central concern when investigating whether wrongdoers are morally responsible for their actions. For reasons I will detail in §1.4–1.5, I reject Attributionist accounts of moral responsibility, and thus the claim that attributability is sufficient for accountability. However, it is important to understand these accounts because they have been very influential in how the literature has progressed, and because arguments in support of them will play an important role in later chapters of this thesis.\footnote{Other Attributionists include Smith (2005, 2015) (though she rejects this label) who takes an agent to be responsible when it would be \emph{intelligible} to ask them to \emph{answer} to us for their judgements, and Scanlon (1998, 2008), who believes that we blame people for failures of self-governance that “have a particular kind of [moral] significance for that agent’s relations with other people” (p. 276). The main objections to Smith’s account are that plenty of lapses she takes to be blameworthy simply do not express an agent’s evaluative judgements (Levy, 2011a), that blaming agents presupposes there is no justified answer they can give for their actions, and that demanding an answer does not always amount to blame (Nelkin, 2015). The main objections to Scanlon are that there seem to be clear cases where we blame without modifying our relationships with people, cases where we modify our relationships on the basis of culpable wrongdoing without blaming (Wolf, 2011), and that it seems to miss the crucial affective component we experience when blaming (Watson, 2002).}

\subsection*{1.4 Why Blame?}

Since there are numerous candidates for the conditions under which agents are accountability blameworthy, it is important to consider what justifies our blaming responses. One important distinction in discussions of moral responsibility is the difference between it being in principle appropriate for us to blame someone, and for us being justified in blaming someone, all-things-considered. If openly blaming your culpably unfaithful spouse would cause them to commit suicide, we might typically describe this as a case where expressing blame would be inappropriate. But this sense of ‘inappropriateness’ differs in an important way to the inappropriateness at issue when your spouse has done nothing wrong. In the first case, your spouse is an appropriate target of blame, but you would not be justified in blaming them. In the second case, your spouse is not an appropriate target of blame, and you are not justified in blaming them. If blaming your spouse was the only way to prevent some terrible outcome from occurring however, we would describe this as case where blame is all-things-considered justified, but not appropriate.
This distinction is important, because we often blame for a number of reasons. We want
to deter wrongdoers, uphold moral norms, and facilitate apologies. Most philosophers accept
that these ‘forward-looking’ considerations are important, and that there can be cases where
we ought to blame, even if blame isn’t entirely appropriate. However, they treat these
considerations to be conceptually independent of ‘backward-looking’ considerations, which
are primarily concerned with the way in which agents are connected to their wrongdoing.

To bring out the difference, consider our practice of blaming children when they commit
wrongs. Though they aren’t full agents, we often engage in blame-like behaviours to teach
them and train them to become moral agents. In cases where blaming behaviours would fail
to serve these purposes however, say because the child couldn’t understand what they were
doing was wrong, blame often loses its point. In contrast, we often blame adult wrongdoers
even when doing so will serve no benefit e.g. the wrongdoer is a repeat offender who just
doesn’t care about being blamed. Blame remains appropriate. Philosophers investigating the
conditions agents must meet to be blameworthy usually have this kind of appropriateness in
mind, though there are exceptions. For example, Fricker (2016) argues that in paradigmatic
cases of blaming, blame functions as a proleptic mechanism, which has the aim of inspiring
wrongdoers to feel remorse for what they have done.

One interesting account which gives an important role to forward-looking considerations
comes from Vargas (2012). Vargas argues that forward-looking considerations shouldn’t be
thought of as contingent consequentialist considerations which we should treat as
conceptually distinct from backward-looking considerations. Instead, we should think of
moral responsibility at the second-order level of a practice, with the aim of this practice
being to cultivate moral considerations-sensitive agency. However, while this forward-
looking goal justifies our practice overall, the appropriateness conditions for any particular
instance of blaming within those practices are still backward-looking. By analogy, basketball
has rules against travelling with the ball to prevent players from gaining an unfair advantage
in certain circumstances. This is a forward-looking consideration, in that we want to prevent
certain kinds of behaviour. However, we all agree that players who travel are appropriately
penalised, even in cases where they travel without gaining an advantage. Thus an agent can
be fairly penalised, with this penalty being appropriate in a backward-looking way because of how the player was connected to their act of travelling.16

Another promising account which justifies blame through its agency-cultivating ability comes from McGeer (2013, 2015). McGeer argues that blame’s function isn’t to merely regulate or change a wrongdoer’s behaviour. Rather, it calls on them to exercise their capacities as moral agents. By analogy, trying to convince you of P doesn’t merely attempt to manipulate you into believing P like hypnosis does, instead, it engages with you qua reasoning agent while respecting you qua believer. Blame acts as a costly signal which communicates that we don’t interpret the wrongdoer’s actions as being acceptable, drawing them into an exchange which makes them and the victim try to find a shared sense of the wrong and how the wrongdoer ought to behave now. In doing so, blame helps sustain our moral capacities when we commit wrongdoing, because caring about our moral interactions with others requires caring about whether we live up to the shared moral standards, particularly the standards of those who educated us about right and wrong during our moral development.

In this manner, moral responsibility practices ‘scaffold’ our agency (McGeer & Pettit, 2015b). Though in principle, such scaffolding and agency-cultivation could be achieved without deploying blame and its ‘opprobrium’, given our evolutionary history blame is a uniquely powerful form of communication suited to the task.17

Though forward-looking accounts are important and worth investigating, I believe that when we ask, ‘Is this person blameworthy?’ we are primarily interested in the backward-looking sense. We are interested in whether this person is connected to their action in such a way that licenses blame, independent of the future effects of our blaming. Vargas’s and McGeer’s accounts don’t justify blaming by appeal to forward-looking considerations alone, but thinking about our blaming practices in this broader manner introduces meta-ethical questions which are outside the scope of this thesis. In this thesis, I will focus exclusively on blameworthiness in this backward-looking sense.

---

16 The main objection to Vargas’s account concerns whether he succeeds in locating questions of justification at the second-order level of practices, instead of at the level of particular blaming responses to deserving wrongdoers as most philosophers have. See McGeer (2015) and Levy (2014b). See Vargas (2015) for replies.

17 The main objection to this account is that it seems to have difficulty justifying blaming culpable wrongdoers who are dead, dying, absent, incorrigible, or who already self-blame, since these are the kind of people whose agency won’t be cultivated through being blamed (Vargas, 2015).
Our question remains what justifies blaming agents in this backward-looking sense. It is worth noting that many philosophers’ accounts of what makes blame appropriate are tightly connected to what they take holding an agent morally responsible to amount to. For example, Smith (2015) maintains that in blaming a wrongdoer, we are calling on her to answer to us and give us some justification for her actions. Agents are blameworthy then, when it is intelligible to call on them to answer to us. Scanlon (1998) maintains that blame consists of a judgement of culpability and a modification of one’s relationship. Since agents regularly modify their relationships and expectations with others, he is quite permissive regarding which kinds of agents count as blameworthy. Talbert (2012), Hieronymi (2001) and Smith (2013) (again) argue that the point of blame is to protest wrongdoing, and that this act of protest is something victims are entitled to do in virtue of having been wronged. Even if not expressed, private blame can communicate to the victim that they deserve better.

Another group of philosophers interprets the backward-looking question in a very different way to the accounts surveyed thus far. For these philosophers, to ask ‘Is this agent an appropriate target of blame?’ is to ask ‘Does this agent deserve to be blamed?’ Pereboom (2014) takes the key sense of desert in this question to be ‘basic desert’, where a wrongdoer deserves to be blamed “just because she has performed the action, given an understanding of its moral status” (p. 2). This desert is basic in that it is not justified by further (e.g. consequentialist or contractualist) considerations.

Of course, examining the concept of desert is no less complicated than examining the concept of moral responsibility. However, we can say some things about desert. First, to say that a wrongdoer is a deserving target of blame is not simply to say that it would be not unfair to blame them. A wrongdoer could reasonably complain our blame was distributively unfair if we failed to blame a second wrongdoer who was equally blameworthy, without this thereby showing the first wrongdoer wasn’t an appropriate target of blame.\footnote{It would also be not unfair if both wrongdoers were to get less blame than they deserve.} We might think that undeserved blame is unfair in an intrapersonal sense, however. Second, to say that a wrongdoer is a deserving target of blame is not to say they are a deserving target of the reactive attitudes, although this gets us close. If a wrongdoer relished the thought of being blamed, or would be helped considerably by being blamed, they would deserve not to be blamed. Following Nelkin (2013a, 2016a), I propose that to say a wrongdoer is an
appropriate target of blame is to say they deserve to be sanctioned, with blame and reactive attitudes constituting the means by which we ordinarily sanction.\textsuperscript{19}

There is a further question concerning what a basic desert account entails for members of the moral community. If an agent deserves to be blamed, does this mean that it would be intrinsically better for them to be blamed by us, or that we have a duty to blame, or something else?\textsuperscript{20} Nothing substantive in this thesis turns on this question, but I’m inclined to follow Nelkin (2016a) in thinking of desert claims as regarding a special kind of fittingness, because we are concerned with the relationship between a wrongdoer and ways of treating them. However, we can’t be interested in mere fittingness, because mere fittingness doesn’t itself provide moral reasons for action. That it would be fitting on one dimension to laugh at a funny racist joke itself does not provide any \textit{pro tanto} moral reason to respond by laughing. Likewise, the fact it would be fitting with Japanese norms to take off my shoes in my home does not give me any reason to do so. Instead, fittingness only gives us \textit{pro tanto} reasons to do these things in certain contexts (laughing when a joke isn’t wrong, taking off shoes when in a Japanese house). Nelkin argues that to say an agent deserves to be blamed is to say we in the moral community have a conditional \textit{pro tanto} reason to sanction them.\textsuperscript{21} She leaves open in what kinds of circumstances we might lack such a reason, but plausible answers include when the wrong was committed a long time ago and the wrongdoer has atoned, when the wrongdoer has been forgiven by the victim, or when the wrongdoer has

\textsuperscript{19} One recent related proposal is that wrongdoers deserve to feel guilt, which is intrinsically unpleasant, and blame is the means by which we induce this (Carlsson, 2017; Clarke, 2016). One problem with this explanation regards cases where wrongdoers already self-blame and feel too much guilt, but blame still seems appropriate (Tierney, 2019).

\textsuperscript{20} McKenna (2012) and Clarke (2013) support the axiological thesis that all else being equal, a world where a culpable wrongdoer is blamed (\textit{qua} sanction) contains more intrinsic value than one where that wrongdoer is not blamed, while Scanlon (1998) and Nelkin (2014b) reject this.

\textsuperscript{21} This account also overlaps significantly with Feinberg’s (1970) claim that to say an agent deserves X is to say “that he ought to get X in the \textit{pro tanto} sense … that a person’s desert of X is always a reason for giving X to him, but not always a conclusive reason … that considerations irrelevant to his desert can have overriding cogency in establishing how he ought to be treated” (p. 60).
been blamed enough by the moral community and more blame would be disproportionate to the wrong done.\textsuperscript{22}

1.5 The Control Condition

Thinking of the appropriateness of blame in terms of whether blame is deserved poses a problem for Attributionist accounts of moral responsibility. It seems to many philosophers that, if an agent is deserving of blame, they must have had the ability to avoid being subject to that blame. This in turn implies that the agent must have had some level of control over their actions, such as the ability to do otherwise. In contrast to Attributionists, who locate responsibility in the way that an agent is connected to her actions, ‘Control theorists’ locate responsibility in whether an agent has a certain kind of control over their actions. This seems like an intuitively appealing account, but treating control as a necessary condition for blameworthiness generates a few worries.

One worry is that if determinism is true, it looks like none of us have the capacity to avoid wrongdoing, or possess control over any of our actions. Pereboom (2014) is a sceptic about moral responsibility in the sense required for basic desert for precisely this reason. Another worry is that, even if compatibilism is true, certain kinds of unfree or determined actions still seem to be excused, and we need some way to separate these kinds of actions from more ordinary actions we are responsible for, even if those ordinary actions were also determined.

A large literature has sprung up regarding a number of issues, so let me briefly identify some key battlegrounds. One area of debate concerns manipulation arguments (Kearns, 2012; McKenna, 2008; Mele, 2006b; Pereboom, 2014; Sartorio, 2016; Sripada, 2012). These arguments introduce thought experiments in which an agent seems to be determined to act in certain ways (e.g. due to nefarious neuroscientists). The cases are described in such a way that the agent intuitively seems to be unfree and thus excused for their wrongdoing. However, these cases are also described in such a way that there seems to be no freedom relevant, or morally relevant, difference to actions performed in a deterministic universe.

\textsuperscript{22} If one is worried that this ‘in certain circumstances’ clause risks not doing any work since it could plausibly just mean ‘when there are no countervailing considerations’ (which we’ve already marked out with our appropriate/justified distinction), it seems plausible that actions can be all-things-considered unjustified in different ways. If you are deciding whether to blame your suicidal spouse, you have a reason to blame, but this reason is outweighed by other pro tanto reasons against blaming. But in cases where someone has been blamed enough, or the wrong was forgiven, you plausibly instead lose your reason to blame.
Another debate concerns whether the ability to do otherwise is necessary for moral responsibility. Some philosophers have argued for The Principle of Alternate Possibilities (PAP), which holds that agents must be able to do otherwise in order to count as blameworthy for performing an action (Copp, 1997, 2008; Widerker, 1991, 2000). This is typically thought to be a corollary of the ‘Ought Implies Can’ principle, as we often think that agents can’t be required to \( \phi \) if they are unable to \( \phi \). Likewise, if an agent can’t \( \phi \), it seems unfair to blame them for not \( \phi \)-ing. The most natural reading of PAP however, is challenged by Frankfurt-style cases (FSCs). These are cases where a wrongdoer lacks the opportunity to do otherwise due to the presence of a counterfactual intervener (Frankfurt, 1969). Were the agent to try to avoid wrongdoing, someone else (typically a nefarious neuroscientist) would intervene and force the agent to perform the action. Crucially, the agent does not try to avoid wrongdoing, and the counterfactual intervener never intervenes. Some philosophers believe FSCs show that PAP is false.

Frankfurt-style cases have also inspired a large literature with much back-and-forth between participants (e.g. Franklin, 2015b; Ginet, 2002; Kane, 1996; McKenna, 2003; Mele & Robb, 1998; Robinson, 2014). These debates generally centre on modifications to PAP or adoption of PAP-like principles, alternate senses of ‘ability’, and increasingly sophisticated FSCs. One of the most influential arguments for the claim that agents in FSCs maintain a kind of responsibility-grounding control over their actions comes from Fischer and Ravizza (1998). Fischer and Ravizza argue that, while agents in FSCs lack ‘regulative control’ in that they lack access to genuine alternative possibilities, they still have ‘guidance control’ in the actual sequence of events. By analogy, a learner driver taking a left turn maintains control over the car’s turning, even if it is true that, had they tried to turn right, the instructor would have engaged the second steering wheel and prevented them from doing so.23

One useful and influential concept from their account is ‘reasons-responsiveness’. In order to have guidance control, Fischer and Ravizza argue that an agent’s action must issue from a reasons-responsive mechanism that the agent has taken ownership of. Reasons-responsiveness is made up of two components: moderate reasons-receptivity and weak reasons-reactivity. An agent has moderate reasons-receptivity if, holding fixed the operation

---

23 Note that Fischer and Ravizza are concerned with identifying the capacities that agents need to be morally responsible, arguing that agents must meet further conditions in order to be blameworthy (e.g. Fischer, 2012, p. 215 fn. 16).
of the mechanism that issues in the action, the agent would recognise reasons, some of which are moral, in such a way that gives rise to an understandable pattern across nearby possible worlds. An agent has weak reasons-reactivity if, when given sufficient reason to act, the mechanism would operate and the agent would act in at least one possible world.\(^{24}\)

Even if Control accounts succeed in identifying a kind of compatibilist control sufficient to ground moral responsibility, a common objection from Attributionists is that Control accounts excuse many kinds of agents who we typically treat as blameworthy. Upon philosophical and psychological investigation, it seems that many kinds of people have less control over their actions than we typically take them to have, even if we’re compatibilists. One philosopher who not only bites this bullet but enthusiastically chews it is Levy. Across many provocative articles, Levy argues that we cannot be blameworthy for moral ignorance (2003, 2009), moral blindness (2007), lapses in judgement or memory (2011a), implicit attitudes at odds with our explicit attitudes (2014d), or actions which are wrong in virtue of facts which are not personally available to us (2013) (excluding tracing cases, explained below). Levy’s argument regarding most of these cases is that wrongdoers simply could not do the right thing for the right reasons if they are not aware of the relevant facts at the time of wrongdoing, could not rationally decide to do the right thing, or could not guide their actions by the relevant moral reasons. These agents are thus not deserving of blame.

Unfortunately, Levy (2011b) is also a sceptic of moral responsibility in the basic desert sense altogether because he thinks luck affects our agency and every decision we make, and this is incompatible with basic desert. In this thesis however, I want to assume that at least some agents are morally responsible in the basic desert sense. Though this may beg the question against Levy and other basic desert sceptics, I think this is necessary insofar as my project is worth carrying out at all.\(^{25}\) If subjects who act wrongly in the situationist experiments are excused for their actions because everyone is, there wouldn’t be anything about these experiments worth investigating for theorising about moral responsibility and

\(^{24}\) Objections to their account regard what counts as a mechanism, whether the account can handle manipulation cases, whether the weak reasons-reactivity condition is too weak, and whether appealing to possible worlds is the best way to think about an agent’s moral responsibility. See e.g. McKenna (2000); Mele (2000); Todd and Tognazzini (2008); Watson (2001). For replies see Fischer (2006, 2012).

\(^{25}\) Against Levy’s luck argument, I am sympathetic to Franklin’s (2015a) objection that agent-causal libertarians are able to offer contrastive explanations of behaviour, and so agents can thus be directly responsible for akratic action. See Levy (2015b) for a reply.
blameworthiness. Insofar as I think that at least some agents are deserving of basic desert, I am committed to there being at least some agents who maintain a kind of control over their actions sufficient to ground blaming and praising responses.

One Control account which maintains some agents are morally responsible in the basic desert sense comes from Dana Nelkin.26 Nelkin (2011; also with Brink, 2013) provides a very simple but powerful account of what conditions agents must meet in order to be morally responsible for their actions. According to this ‘rational abilities’ view, to be morally responsible, agents must have the capacity to do the right thing for the right reasons.27 This capacity can depend on other more basic capacities such as perceptual, emotional, and physical capacities. Agents must also have the opportunity to exercise their capacity to do the right thing for the right reasons, in that they must not be subject to duress or coercion.

Nelkin argues that, in order to be a deserving target of blame for wrongdoing, agents must have had a fair opportunity to avoid wrongdoing. One interesting aspect of Nelkin’s account is that it is asymmetrical regarding whether agents must have had a fair opportunity to do otherwise to be deserving of blame or praise. Agents don’t need a fair opportunity to do otherwise in order to be praiseworthy for performing a right action. Even if one simply

---

26 In what follows I treat ‘abilities’ and ‘capacities’ as synonymous, as most philosophers do.

27 It should be noted that Vargas and McGeer also endorse a ‘capacitarian’ account of moral responsibility, in that agents must have had the capacity to respond to moral reasons in order to be blameworthy. These seem to overlap with Nelkin’s account regarding which kinds of agents are accountability responsible. But the considerations Vargas takes to justify blaming lead him to adopt a broader sense of when an agent has the capacity to do the right thing for the right reasons than Nelkin, one that appeals to whether the agent would succeed in doing the right thing in a suitable proportion of nearby possible worlds. In contrast, Nelkin places a much greater emphasis on deliberating and acting for reasons in the actual sequence, which means that agents whose wrongdoing is the product of, say, non-traceable forgetting are excused, even if it is true that they would have succeeded in remembering in a suitable proportion of nearby possible worlds. It isn’t clear whether McGeer’s (2018) account of capacities is more like Nelkin’s or Vargas’s. On the one hand, she argues that answering whether an agent could have tried or intended to do the right thing is not needed for a particular disposition to count as an intelligent capacity, and that failures to exercise our intelligent capacities are appropriate targets of blame in virtue of those capacities’ sensitivity to be developed by blame (as is our overall agency). It seems like our capacities to remember or notice are the kinds of things which can be developed in this way. On the other hand, her arguments focus on cases where agents are aware of the relevant reasons, but simply choose not to act on them. There seems to be an important difference between agents who fail to help us because they don’t notice us, and agents who fail to help us because they choose not to help us, even if in both cases the failure is the product of an unexercised intelligent capacity. Capacities are discussed further in the next section.
‘cannot do otherwise’ or ‘had no choice’ to do the right thing, they will still be praiseworthy if they acted for the right reasons.

I am sympathetic to Nelkin’s account, and will adopt it as a starting point for my investigation into the moral responsibility of subjects in the situationist experiments. I adopt this particular account for a number of reasons. First, the account is relatively simple, but coheres with our intuitions about a wide variety of cases. Second, it respects our concern for basic desert. Third, Nelkin (2013a, 2016a) has shown that while identifying accountability responsibility with basic desert, her account can also give an important role to moral demands and the reactive attitudes. Many philosophers have argued that accountability responsibility must give a central role to moral demands and the reactive attitudes as these play an important role in our actual moral practices. Nelkin’s account thus respects central features of our actual moral practices, and is compatible with many claims by other philosophers that, on first glance, may appear to support competing accounts. Fourth, the account has existing replies to some of the challenges that have arisen in the complex debates over free will. Most notably, Nelkin (2011) argues her account can fend off objections from compatibilists who argue PAP-like principles are false due to FSCs, incompatibilists who argue compatibilism is false due to manipulation cases, and libertarians who argue compatibilists can’t secure the kind of control necessary for basic desert.

Adopting this account improves the strength of my argument. This chapter has reviewed previous philosophical work on moral responsibility. While I’ve aimed to give readers an idea of how key ideas and arguments have developed over time, I’ve also tried to argue for (or at least point to considerations strongly favouring) the conclusion that we should think about moral responsibility and blameworthiness in a particular way. Due to the large size of the free will literature, however, in this section I’ve only been able to gesture towards key debates. Because Nelkin’s account does engage with that literature, adopting it means my claims about the responsibility of the subjects in situationist experiments have a stronger foundation than if I were to argue for a new account of moral responsibility altogether. Choosing this account as my starting point allows me to ‘free-ride’ (with permission) off Nelkin’s previous contributions and move on to examining the subjects in the situationist experiments.

The rational abilities account also seems like a natural starting point for my thesis, because it is subjects’ decision-making capacities in the situationist experiment that are of interest to us. However, the radical difference between how participants act in these situations and how
they act in normal situations suggests that the experimental environments could be negatively affecting their decision-making capacities.

1.6 A Primer on Capacities

Nelkin’s (2011, 2016b) rational abilities account argues agents only possess control in the relevant sense required for accountability responsibility if they had a fair opportunity to avoid wrongdoing. In order to have a fair opportunity to avoid wrongdoing, agents must have the capacity to do the right thing for the right reasons, the opportunity to exercise this capacity, and not be subject to unreasonable costs. While the basic structure of the account is intuitive enough, Nelkin mostly defends it while engaging in debates about deterministic universes or Frankfurt-style cases, though she clearly intends for it to generalise to more ordinary contexts where we set aside the truth or falsity of determinism (e.g. Brink & Nelkin, 2013; Nelkin & Rickless, 2017). This means she doesn’t spend much time detailing what a capacity is, or how we know when an agent has the capacity to do the right thing for the right reasons (although she provides many instructive examples). This isn’t a criticism; many philosophers have made substantive arguments in these areas while relying on an intuitive sense of capacity, or showing that a purported counterexample doesn’t demonstrate an agent lacks a particular capacity. For example, it is possible to show that wrongdoers in Frankfurt-style cases who don’t trigger the counterfactual intervener are responsible for their wrongdoing, because they retain the interference-free capacity to act on moral reasons and avoid wrongdoing. In these settings, ‘avoiding wrongdoing’ is understood as initiating the action which triggers the counterfactual intervener (Nelkin, 2011).

As I explain below, I also will not be providing a comprehensive account of what capacities are, or under what conditions an agent possesses a particular capacity. Instead, to make my argument, I will be examining what kinds of evidence should lead us to think that an agent possesses or lacks a particular capacity. However, examining existing philosophical work on capacities is worthwhile, because capacities are understood in different ways by different philosophers in different contexts, and talk about capacities can carry different implicit assumptions depending on our interests.

Talk about capacities is very context-sensitive, and failure to recognise this can lead to seemingly inconsistent answers (Kratzer, 1981; Lewis, 1976). Do I have the capacity to speak French? In one sense, no, as I don’t know how. In another sense, yes, as I can enrol in
lessons. Because of such context-sensitivity, questions about capacities are typically answered with regard to certain facts. There is very rarely talk of the capacity to \( \phi \). Can I play the piano? In one sense, yes, given I have the training. In another sense, no, given I am currently on an airplane and no piano is to be found. Do I have the capacity to choose to go to work tomorrow? In one sense, yes, given I’ve done it many times before. In another sense, I may lack this capacity, if determinism is true and it’s determined that I will stay home.

To help clarify talk about capacities, philosophers distinguish between general capacities and specific capacities (Clarke, 2015; Jaster, 2019; Maier, 2010; Mele, 2006b). An agent has a *general* capacity to \( \phi \) if they would normally succeed at \( \phi \)-ing in a wide range of contexts. Kobe Bryant has the general capacity to play basketball because he reliably dribbles and sinks the ball, and can successfully do so in a wide variety of circumstances. When we ask if an agent has a *specific* capacity, however, our focus is restricted to particular circumstances, holding particular facts fixed. If Bryant is tied to chair, he retains his general capacity to play basketball because he knows how to play and would succeed in playing basketball in most normal circumstances. But he lacks the specific capacity to play basketball in this situation because he cannot presently play basketball, no matter how hard he tries.

Agents can possess a general capacity to \( \phi \) while lacking the specific capacity to \( \phi \) in two ways: due to features internal to the agent, or due to external features of their situation. While I have the general capacity to play piano, I will lack the specific capacity to play piano if I have broken hand (an internal factor), or if I am on a plane with no piano available (an external factor). Agents who lack a specific capacity due to external factors are said to lack the *opportunity* to exercise their capacities (sometimes described as lacking the wide ability). Agents who lack a specific capacity due to internal factors are said to lack the narrow capacity to act (Clarke, 2015; Vihvelin, 2013).

It is also useful to distinguish between agential and non-agential capacities (Jaster, 2019; Löwenstein, 2017; van Inwagen, 1986). Agential capacities are capacities we can actively exercise at will, such as talking, walking, or playing an instrument. Non-agential capacities, by contrast, cannot be directly exercised at will. The capacity to digest food, to hear something nearby, and to unintentionally read street signs, are not capacities we can exercise at will. We can, however, do things which trigger these capacities, such as eating food or going somewhere where there are street signs.

These terms help us clarify some talk about capacities, and explain away some *prima facie* inconsistencies in our everyday capacity ascriptions. But when we begin thinking about
capacities for the purposes of assessing an agent’s moral responsibility, our use of the term ‘capacity’ may diverge from how capacities are discussed in other philosophical contexts. These contexts include understanding agency (Mayr, 2011), explaining action (Small, 2017; Smith, 2003), reconciling talk of capacities with talk about determinism and free action (Clarke, 2009; Franklin, 2011; Vihvelin, 2004), and understanding the semantics and truth conditions of counterfactuals (Lewis, 1973). These are legitimate questions, but they are beyond the scope of this thesis. So long as we are clear regarding what our interests are, and how we are using our terms, our use of the term ‘capacity’ does not need to map perfectly onto how the term is used in these contexts. Let me therefore specify some considerations for how we should understand the capacity to do the right thing for the right reasons.

The first consideration is that talk of the capacity to do the right thing for the right reasons must be understood more narrowly than talk of specific capacities in other contexts. Suppose I fail to give aid to someone. To assess whether I am blameworthy, we need to assess whether I had the capacity to see that person, recognise they were in need, realise that their being in need gives me a moral reason to assist, run to them, recall my first aid training, form an intention to give first aid, and translate that intention into action to, in fact, provide first aid. If we maintain that doing the right thing in this situation is ‘giving first aid’, and I lack any of these specific capacities, then I will lack the capacity to do the right thing for the right reasons. If I lack any of these capacities, I may still have other duties e.g. to call for help, but this would also change what constitutes doing the right thing.

Some philosophers resist talking about so many specific capacities because it seems to multiply capacities beyond necessity (D. Locke, 1973; Small, 2017). For example, instead of having one capacity to walk that I exercise in different contexts, we have to say I have an indeterminably large number of capacities to walk to an indeterminably large number of places. They argue that it is far simpler to say I have one capacity, which is to walk, that I can exercise in a number of different ways and places. But whether I have the specific capacity to walk down a particular street matters greatly if I am being blamed for failing to help someone on that street. The fact that I have a general ability to walk is much less relevant. For our purposes, we need to index talk about capacities to a particular situation and context. The fact an agent typically has a general or specific capacity to \( \phi \) in many other contexts doesn’t guarantee they have the specific capacity to \( \phi \) in this particular context.

Another consideration to keep in mind is that an agent can possess a specific ability to \( \phi \), have ample opportunity to \( \phi \), try to \( \phi \), and yet sometimes still fail to \( \phi \) (Austin, 1956; Jaster,
2019; Vihvelin, 2013). To put it another way, a single instance of an agent trying to \( \phi \) the best they can, and failing to \( \phi \), does not necessarily entail they lacked the specific capacity to \( \phi \) in that situation. This is important to recognise, because many philosophers working on moral responsibility do think that the fact an agent tried to do the right thing, but failed, is often exculpatory. But if we’re assessing an agent’s culpability by assessing whether they had the capacity to do the right thing for the right reasons, we may erroneously categorise agents who tried to do the right thing, but failed, as blameworthy in virtue of the fact that they had the capacity to succeed in doing the right thing.

To show what I mean, suppose Bryant must make a three-pointer in order to save a local hospital. Suppose he tries his best, but misses. Would we blame him for this, since he had the specific capacity to shoot a three-pointer? Surely not. But if we interpret ‘doing the right thing for the right reasons’ as scoring-a-three-pointer because this is the relevant specific capacity in other philosophical contexts, and we then ask whether Bryant had the capacity to do the right thing for the right reasons, we will mistakenly conclude that he is blameworthy.

This case shows it is important to keep in mind the context-sensitivity at play when talking about capacities, and how it can be difficult to identify which capacity must be exercised to count as ‘doing the right thing’ in any given setting. In this example, doing the right thing can’t simply be shooting-a-three-pointer. Bryant failed to shoot a three-pointer, but he isn’t blameworthy. Perhaps we want to say that the right thing to do in this situation is to try to shoot a three-pointer (or to intend, or choose, or decide, and then act on this intention or choice or decision). But suppose Bryant tries half-heartedly. He tries to shoot, but also tries to ensure the cameraman catches his good side, distracting him. He could ignore the camera and devote his attention entirely to getting the basket, but doesn’t. In this case he has tried to get a three-pointer, but he would still be blameworthy. Trying to do the right thing itself thus isn’t sufficient to count as doing the right thing for the right reasons.

One might be tempted to say he should try his hardest. But this causes problems in cases were trying too hard can reduce one’s chance of success, as sometimes happens when athletes ‘choke’ in sport. I think we should say Bryant needs to try his best, but here’s one prima facie worry with this proposal. Suppose I am running a marathon, and the further I run, the more money is raised for a hospital. Suppose I could run the entire race, but that doing so would leave me in a large amount of pain for several months. At a certain point, the cost of continuing running will simply be too high to be reasonably expected of me. I stop running. I didn’t try my best, but perhaps I tried sufficiently.
We can explain why I’m not blameworthy despite not trying my best in this case by introducing the concept of costs. When a mugger says to me, ‘Your wallet or your life’, I have the capacity to refuse to give my wallet (I can say ‘no’), I have the opportunity to exercise this capacity (nothing prevents me from physically speaking), but there is a sense in which I cannot really say ‘no’. I could not be reasonably expected to say ‘no’, and I could not be blameworthy for failing to say ‘no’. This is because the costs of saying ‘no’ are so high that I can’t be considered morally responsible for failing to retain my wallet. Thus, when we are trying to assess an agent’s capacity to φ for the purposes of assessing their responsibility for not φ-ing, we need to also ask whether φ-ing is something that they could have done without incurring unreasonable costs.28

Suppose we say that, in any given situation, the right thing is understood as trying one’s best to perform the right action (where the right action is one that doesn’t cause the agent to incur unreasonable costs). So far so good. But here’s a big worry: in many cases, trying itself looks a lot like an action that agents have the capacity to perform.

At this point, philosophers often attempt to analyse what a capacity is, or what it means to have a capacity, in one of two ways. The first is by analysing capacities in terms of conditionals, e.g. S has a capacity to φ if and only if S would φ, if S intended to φ (Aune, 1963; Hume, 1748 (1999); Moore, 1912; Sosa, 2015; Vihvelin, 2013). The second is to analyse capacities in terms of possible worlds, e.g. S has a capacity to φ if and only if there is a set of relevant worlds in which the agent φ’s, where the set of the relevant worlds will vary across ascriber context (Fischer & Ravizza, 1998; Kratzer, 1981; Lewis, 1976; Vetter, 2015). These analyses then have to grapple with questions surrounding masks (Fara, 2005), finks, deviant

---

28 Another related term which can account for why failures to complete marathons are not blameworthy is that of ‘difficulty’. Some things are very difficult to do and, plausibly, the more difficult something is, the less of a capacity we have to do it. While I do reference difficulty in my upcoming arguments, I didn’t want to make my argument exclusively in terms of difficulty because we currently lack any consensus regarding what it means for something to be difficult. Nelkin (2016b) argues that to say something is difficult is to say it involves effort or costs. Below I argue the subjects in the situationist experiments are unable to interpreting their situation in a particular manner, and while interpreting situations is in some sense ‘difficult’, it isn’t clear whether this difficulty is analogous to what we mean when we talk about physical effort, or even the difficulty we with overcoming weakness of will. Von Kriegstein (2019) makes the point that not all effortful things are difficult. Moving heavy furniture isn’t difficult, it’s just hard. He proposes instead that, to say something is difficult is to say it is has a low probability of success. This definition matches ordinary usage of the term, but isn’t especially helpful for us, because we’ve been trying to analyse what difficulty is precisely in order to work out when those failures are excused (or at least, less blameworthy).
causal chains (Wilson & Shpall, 2008), impeded intentions (Lehrer, 1968), and how to specify which possible worlds are relevant (Jaster, 2019).

While such analyses can help us get even clearer about when agents have a certain capacity, it is particularly difficult to account for the fact that trying (or intending) itself often looks like a capacity, and references to trying or intending in one’s account of what a capacity is risks creating a well-known regress (Jaster, 2019; Small, 2017; cf. Moore, 1912). For example, suppose we take an agent to have a capacity to $\phi$ if and only if they would succeed in $\phi$-ing were they to try to $\phi$. To assess whether an agent has the capacity to try to $\phi$, we then need to know whether they have the capacity to try to try to $\phi$, and so on.

For our purposes, we don’t need to settle the question of whether capacities should be understood as dispositions cashed out in terms of conditionals or facts about possible

---

29 For this reason, Jaster (2019) argues that intending (or choosing or trying or deciding) cannot be an agential action. Instead, intendings are non-agential capacities which are triggered by certain factors (though she is silent on which factors, and what it means for them to be ‘triggered’). Analysing capacities in terms of possible worlds generates what is, in some way, an opposite problem: how to distinguish capacities from other restricted possibilities. The possible worlds approach not only fails to account for the negative affective component usually present when we say a wrongdoer could have done otherwise, it also fails to identify any difference between my capacity to walk, and my ‘capacity’ to fall off a cliff I am standing on that is, unbeknownst to me, unstable. McGeer and Pettit (2015b) put these worries nicely: “The two failures of the standard reason-responsive approach may be put as follows: first, that it reduces the ascription of a specific reason-responsive capacity to a modal report; and, second, that it reduces it to a modal report” (p. 176).

30 McGeer (2018) raises a second related worry, which she calls ‘the hard problem of responsibility’. This regards how the mere fact that someone could have done otherwise can make them blameworthy. If we think of capacities like dispositions, then, when we try to explain why an agent with the general capacity to do the right thing did the wrong thing instead, at some point our explanation bottoms out and the agent seems merely ‘glitchy’ or unlucky, making consolation more apt than blame. Her answer is that an agent’s ‘intelligent capacities’ are metaphysically and morally different to ordinary dispositions, due to the way they are developed and sustained by agents and feedback from the environment. This connects with McGeer’s agency-cultivation model of blame, because blame is particularly apt for providing the requisite kind of feedback: it draws the wrongdoer’s attention to the fact that they have done something wrong and that this failure is sourced in their own agency, while insisting that their capacity to do better is a persisting feature of them and calling upon him to in fact do better in the future. McGeer and Pettit (2015b) think that the fact that blame would play this role to culpable wrongdoers means that blame does amount to a kind of desert. I think the rational abilities account isn’t threatened by this challenge, because it takes our deliberations to play a genuine explanatory role in accounting for why we perform one action rather than another (or why our disposition was exercised rather than not exercised), and our deliberating manifests a commitment to our being free and responsible agents (see Nelkin, 2011, Ch. 6–7). Matters are delicate here and for brevity I will simply have to put aside whether the rational abilities account can handle the hard problem of responsibility.
worlds, nor do we need to grapple with regress problems. This is because in practice, we don’t make capacity assessments by first investigating the truth of certain counterfactuals or by looking at nearby possible worlds. It isn’t clear how one even could look into nearby possible worlds to assess someone’s capacities. Recall we are trying to assess whether agents in situationist settings are morally responsible for their wrongdoing. We are doing this by assessing whether they had the capacity to avoid wrongdoing. It is clear that if we found out that a mind-control machine was present in all the experiments, or that the subjects were physically restrained from doing the right thing, then we would reasonably conclude they lacked the specific capacity to do the right thing for the right reasons. Whether these factors would act as evidence of incapacity in the relevant sense doesn’t depend on which metaphysical account of capacities we endorse. What matters for the rational abilities account is the capacity to act on moral reasons, and we can assess whether subjects had this capacity without first settling what a capacity is.

To see how we can we can assess whether an agent possessed a capacity without looking into nearby possible worlds, consider an example from Nelkin and Rickless (2017). When we want to know whether Randy is responsible for forgetting the milk on the way home due to his mind wandering, it seems clear that he lacked the capacity to remember the milk if we hold fixed his mind having wandered away from the milk. But even if he never decided to forget the milk before his mind wandered, he retained the capacity to engage in certain actions which would have prevented his forgetting (e.g. turning off the distracting radio, repeating ‘milk’ to himself over and over). As his mind began to wander, he would have been aware that he might forget, and that this could be avoided by engaging in those mental actions. But by deliberating and then not engaging in those actions, he knowingly accepted the risk that he would forget, and this is why his wife can blame him for not getting the milk.

Here then, is how to understand talk of capacities for the purposes of assessing an agent’s culpability under the rational abilities account. To be blameworthy for wrongdoing, an agent must have had the specific capacity to do the right thing for the right reasons, which is made up by the capacity to recognise moral reasons and the capacity to act on those reasons. These capacities are, in turn, constituted by a number of physical, mental and emotional sub-capacities. Although we do not have a precise account of what it means to have a capacity or the conditions under which an agent can be said to have a capacity, we have a sufficiently intuitive notion of what kinds of things impede the exercise of capacities. Ignorance, duress and physical restraints, among other things, can remove an agent’s capacity and/or
opportunity to do the right thing for the right reasons. Ordinary lack of motivation and determinism do not impede an agent’s capacity or opportunity to avoid wrongdoing in the sense required for accountability responsibility. Whether an agent possesses or lacks a capacity is to be settled by making inferences based upon observable evidence.

In addition to having the specific capacity to do the right thing for the right reasons, agents must also have the opportunity to exercise this capacity, and not incur prohibitive costs in doing so, before they can be considered blameworthy. Opportunities, costs, and capacities are all scalar notions. According to Nelkin, an agent’s capacities, opportunities and costs together constitute the agent’s quality of opportunity. All else being equal, the greater an agent’s quality of opportunity to avoid wrongdoing, the more blameworthy they will be for not avoiding wrongdoing. However, we can set aside the opportunity and costs components of the account, because subjects in the situationist experiments clearly have the opportunity to exercise any capacities they do have, and would not incur any unreasonable costs in doing the right thing. No external factors are present in the experiments which prevent the subjects from exercising their capacities, and subjects are not threatened, coerced, or under duress. To assess the culpability of subjects who act wrongly, we only need to assess whether they possessed the capacity to do the right thing for the right reasons.

1.7 The Epistemic Condition

Just as there is debate over when agents are excused for wrongdoing due to a lack of control over their actions, there is also debate over when agents are excused due to a lack of knowledge regarding their actions. To be sure, it can be difficult to neatly separate the two. If I don’t know I’m putting arsenic in my tea, I don’t freely put arsenic in my tea in any meaningful sense, though I might freely act when I put arsenic in my tea. But philosophers dating back to Aristotle (Nicomachean Ethics, 1999, 30[1110a]) have found it useful to separate questions about an agent’s control over their actions from questions about their knowledge or beliefs regarding those actions.

Philosophers have investigated to what extent agents who are morally ignorant due to culture (Calhoun, 1989; FitzPatrick, 2008; Rosen, 2004) or moral blindness (e.g. due to psychopathy; see Levy, 2007; Nelkin, 2015; Nichols & Vargas, 2008; Talbert, 2008) can be blameworthy for wrongdoing. Control accounts typically require that agents be ‘normatively competent’ in order to be morally responsible for their actions. Agents must be capable of
recognising, evaluating and acting on moral reasons, because ignorance of the moral features of one’s actions makes it impossible to do the right thing for the right reasons.

Until recently, relatively little analysis had been given to more common forms of ignorant wrongdoing (cf. Sher, 2009). It has proven particularly difficult to make any definite assessments regarding the culpability of agents whose fault consists in a certain consideration failing to occur to them. The problem is that, if the agent had the capacity to eliminate their ignorance and become aware of that consideration, but didn’t exercise that capacity because they didn’t know that they should exercise that capacity, then we’re faced with another case of ignorance for which we first need to evaluate their culpability. If this higher-order ignorance does excuse them, then unwitting wrongdoing is always excused. If it doesn’t, then without some explanation of why, a regress threatens (Levy, 2009; Rosen, 2004; cf. Robichaud, 2011).

To demonstrate how difficult it is to show that agents can be blameworthy for failing to notice a certain consideration, let me consider several recent attempts. Björnsson (2017a) argues that agents are blameworthy for ignorance when that ignorance is a product of their quality of will falling below a standard that could be properly demanded of them. However, what can be properly demanded of an agent depends on that agent’s capacity to care, and Björnsson doesn’t examine what capacities are, what sense of capacity is relevant, or when we should think that an agent lacks a capacity. This encounters the regress problem, because it is difficult to see how an agent could have cared more (or exercised their capacity to care more) if they did not know that they should care more.

Attributionists avoid this problem when it comes to agents who are ignorant of moral facts, since these agents’ actions do express a denial that certain facts give them moral reasons to do otherwise (Talbert, 2008). But depending on the specific account, Attributionists may encounter problems grounding culpability in cases of circumstantial ignorance. For example, if an Attributionist maintains that agents are only blameworthy for acts or omissions that express an objectionable judgement, then it is hard to see how most circumstantial unwitting wrongdoings can express something objectionable. It is hard to see how a failure to notice that someone is behind my car, or a failure to realise that this blood is the wrong type for a transfusion, expresses an objectionable judgement about the agent that my driving or transfusion then harms. In contrast, if the Attributionist maintains that agents are blameworthy for circumstantial ignorance produced by moral vices, this could make ignorance attributable to the agent and thus plausibly ground culpability for some ignorant wrongdoings.

Based on previous work (Björnsson, 2017b) and how I describe the difficulty faced by the obedient subjects in §4.3, Björnsson would say that the obedient subjects lack the skill to recognise they can disobey, and they are appropriate targets of ‘skill blame’, but not moral blame.
FitzPatrick (2017) argues that agents are blameworthy for ignorance when they could be “reasonably expected to know better” (p. 30). But when trying to show what counts as reasonable expectations with examples, he claims that ignorance resulting from failures in the “normal functioning” (p. 38) of our capacities, such as forgetting due to a failure of memory, are excused. This is a problem, because he doesn’t consider the fact that many other failures to realise that an action or omission is wrong also result from failures in the ‘normal functioning’ of other capacities we possess. In short, he provides no way to distinguish blameworthy failures to exercise a capacity from excused failures to exercise a capacity, beyond saying that failures which we ‘expect’ not to occur are blameworthy.

Clarke (2014) argues that agents are blameworthy for ignorance when that ignorance “falls below a cognitive standard that applies to her, given her cognitive and volitional abilities and the situation she is in” (p. 303). This uses a sense of capacity that is too broad. Clarke does not explain how the mere fact that it was possible for a consideration to have occurred to an agent thereby shows that the agent is deserving of blame in the actual sequence of events, in which this consideration did not occur. Once we try to explain how unwitting wrongdoers could have succeeded in doing the right thing, in the actual sequence, the agents seem to lack the specific capacity to do the right thing for the right reasons, even though they had the general capacity to do the right thing for the right reasons (Levy, 2017).

In each case, these Control accounts fail to show how the fact that an ignorant agent ‘should have known better’ can make the agent blameworthy. They either fail to avoid the regress problem, blame agents who are intuitively excused for their ignorant wrongdoing, have trouble explaining why some agents are culpable for ignorant wrongdoing while others are excused, or rely on a notion of capacities which seems to locate blameworthiness in the mere fact that a certain state of affairs (e.g. a thought occurring to the agent) was possible.

To reduce the number of agents who count as excused for their ignorant wrongdoing, Control accounts place a lot of importance on ‘benighting acts’ (H. Smith, 1983) or what is more commonly called ‘tracing’. This occurs when we trace an agent’s ignorance back to an earlier blameworthy act. For example, if I don’t know I’m offending you with my jokes because I’m blind drunk, but willingly and knowingly chose to get drunk, aware of the risks, then I will be blameworthy for offending you because we can trace my ignorance back to an earlier action for which I am responsible. One problem with such appeals is that it is often very difficult to see what consequences of our actions are likely. One might cultivate a jerky
disposition in high school to seem cooler, without seeing that these actions will lead one to be a jerky boss thirty years down the track (Vargas, 2005).

Let me stipulate that in this thesis, I am concerned only with direct responsibility. That is, whether subjects in the situationist experiments are morally responsible for their actions depends on whether they had a fair opportunity to avoid wrongdoing during the experiment. I am not interested in whether they could have taken steps to cultivate a different disposition earlier in life, and might be morally responsible for their actions in virtue of this. Philosophers writing on moral responsibility are generally concerned only with direct responsibility because, even though everyone agrees agents can be indirectly responsible for wrongs in virtue of some earlier action, we still need an account of direct responsibility in order to identify which kinds of earlier actions can ground culpability.

1.8 Where We Are

I will now summarise how the philosophical literature on moral responsibility thus far bears on my analysis of subjects who act wrongly in the situationist experiments. This thesis takes a Strawsonian-inspired approach to moral responsibility, in that to be morally responsible for an action is to be an appropriate candidate of the reactive attitudes for that action. Though I take our reactive attitudes and moral practices to be reliable markers of responsibility, unlike Strawson I take them to be defeasible, as facts about blameworthiness are ultimately independent of our responses to blameworthiness. All else being equal, the greater consistency with our reactive attitudes that an account has, the more reason we have to think that account is true. If an account of moral responsibility delivers a judgement that an agent is blameworthy (or excused) which conflicts with our moral practices, that account owes some explanation for why we should think our considered reactive attitudes are mistaken.

When evaluating whether subjects in situationist experiments are morally responsible for their actions, I am concerned only with accountability responsibility in a backward-looking sense. This is the sense of moral responsibility concerned with praise, blame and sanctions which may include punishment. I take the question of whether an agent is an appropriate target of blame to really be about whether an agent is a deserving target of blame, where blame is typically taken to be a kind of sanction. I am not concerned with whether we would be justified in blaming wrongdoers in these experiments, all-things-considered. I am also only
interested in whether these subjects are directly responsible for their actions, not indirectly responsible in virtue of steps they could have taken prior to the experiments.

An appealing account of moral responsibility that is compatible with all of these conditions, which I will adopt, is Nelkin’s rational abilities account. On this account, in order to be blameworthy, agents must have had the capacity and opportunity to avoid wrongdoing. Since part of what makes the situationist experiments so interesting is that all the subjects do have the opportunity to exercise any capacities they possess (i.e. no subjects experience duress or coercion), our primary investigation regards subjects’ capacities.

Though I am not an Attributionist, I believe that the kinds of considerations Attributionists appeal to (namely how an agent is constituted and concern for the victims’ entitlement to blame perpetrators) has two important implications for our account of moral responsibility, even if one is a Control theorist. First, I believe that the way an agent is constituted can have implications for whether they have the standing to blame other wrongdoers. Second, I believe that our concern for victims shows that excused wrongdoers can still incur new duties of reparation. The issues of our duties towards victims, and standing to blame, will be taken up in Chapter 5 and Chapter 6 respectively.

In the next chapter, I will begin examining causal explanations of why so many subjects act wrongly in the situationist experiments. This will then allow me to investigate the extent to which these subjects had the capacity to do the right thing for the right reasons.
Chapter 2: Explanations of Situationist Behaviour

2.1 Introduction

Having reviewed the existing philosophical literature on moral responsibility, I will now look at the philosophical and psychological literature on the situationist experiments. Existing analyses of these experiments need to be clarified in two ways. First, we need a clearer account of which environmental features count as situational factors, or what makes something a situationist factor. Second, we need a clearer understanding of what causes the subjects in these experiments to act as they do.

What counts as a situationist factor and how exactly they affect us is not something that philosophers have been able to articulate (McKenna & Warmke, 2017; Miller, 2014). Most philosophers only say that they “influence” us through “subtle forces”, with such language being used by Doris (2002, p. 644), Ross and Nisbett (1991) and Sabini and Silver (2005, p. 548). This lack of clarity matters because, without any concrete explanation for why subjects act as they do, it is difficult to tell whether their actions are the product of a reduced capacity to avoid wrongdoing, or simply the product of a failure to exercise the capacities they have. This, in turn, makes us unable to tell whether the subjects who act wrongly are deserving targets of blame.

This chapter begins by introducing a thought experiment of sorts in §2.2. I say ‘of sorts’ because I believe this thought experiment is not only conceptually possible, but empirically plausible. I argue that different results could be obtained in Obedience by only making slight changes to the experimental design. This matters, because any causal explanation for why so many people obey in Milgram’s Obedience ought to generalise to this modified experiment too. This thought experiment will help pump the intuition that existing causal explanations of subject behaviour are inadequate. It is also useful for getting an accurate sense of how easily ordinary agents can be induced to act wrongly in some settings.

In §2.3, I summarise Doris’s and Harman’s arguments that we lack empirical evidence for thinking global character traits exist. I then outline the responses to Doris and Harman, which try to explain why subjects behave as they do in such a way that does not undermine the existence of global character traits. I then also consider some leading explanations for subjects’ behaviour in situationist experiments from psychologists. These explanations for why subjects fail to do the right thing include:
(1) Subjects experience a competing desire.
(2) Subjects act from a competing trait.
(3) Subjects experience akrasia or lack practical wisdom.
(4) Subjects identify with the experimenter.
(5) Subjects are inhibited by a desire to avoid embarrassment.

In §2.4–2.8 I argue that all of these accounts suffer from important gaps, and we currently lack a sufficient explanation for why situationist factors cause so many subjects to commit wrongdoing.

2.2 Obedience Revisited

When psychologists discuss Obedience, it is often portrayed as an example of what humans are capable of only at their absolute worst, and only in highly specific settings (e.g. Darley, 1995; Perry, 2013a, p. 9). Perry is particularly critical of any claims that the actions of the obedient subjects can be used to make inferences about the ease with which people can be made to cause harm in other settings. Unlike other settings, the subjects are in an unfamiliar situation, they have a certain justifiable level of trust in the experimenter, and are told that the shocks will not cause any damage. The experimental set-up is carefully designed with a multitude of factors present that all contribute to the subjects’ obedience. When the subjects had free rein over the level of shocks they administer and no direction from the experimenter, 98% of subjects chose not to administer the highest shock (Perry, 2013a, p. 307).

This characterisation has some intuitive appeal, but I believe it is unduly influenced by framing. From our perspective, safe in our homes and with the benefit of time and hindsight, being thrust into a room and told to electrocute a stranger by someone in a grey coat is a command many people believe they would be able to resist. Prior to the experiment, psychologists predicted only 1% of people would actually deliver the highest available shock (Blass, 1999b, p. 44). It is initially reasonable to conclude that the 65% figure seen in Obedience is a maximum. We see it as the number of people who, in maximally influential conditions, will electrocute a stranger. In support of this interpretation, we might also point out 35% do disobey, that this rate of disobedience increased greatly in many of the experiment’s variations, and many of those who obeyed did so with great anguish. Many
obedient teachers did the bare minimum asked of them, they looked for other ways to avoid shocking the learner by trying to help him cheat, and clearly wanted to stop.\footnote{33}

Though most people are surprised to learn that 65\% of people did not disobey, reflection on these additional facts usually makes the 65\% figure much less jarring. The fact that 35\% of people did disobey, that this figure was easily increased, and that the obedient subjects were in great distress show that the average person is not a mindless drone who will do anything they are told.

Once these facts are known, most people’s focus usually moves to accounting for why 65\% of subjects do obey even when the learner has stopped responding. This is one of the most famous quotes from (an observer of the experiments quoted in) Milgram’s 1963 paper:

> I observed a mature and initially poised businessman enter the laboratory smiling and confident. Within twenty minutes he was reduced to a twitching stuttering wreck, who was rapidly approaching nervous collapse. He constantly pulled on his ear lobe and twisted his hands. At one point he pushed his fist into his forehead and muttered “oh God let’s stop it”. And yet he continued to respond to every word of the experimenter, and obeyed to the end.


Most psychologists and philosophers have focused on the phrase “and yet he continued to respond”, taking this to be the most important part of the experiment worth explaining. This is reflected in the fact that much of the philosophical and psychological literature talks about weakness of will, the internal struggle the man felt, his behaviour, and why he continued despite his clear judgement that doing so was wrong. While such analyses are worthwhile, I believe they can miss something important. Here is the phrase that I think is far more interesting when analysing \textit{Obedience}:

> Within \textit{twenty minutes} … [emphasis added]

\footnote{33 Though there are some exceptions. In Condition 4, subjects had to force the learner’s hand onto a shock plate. Participant Paul Batta had “no difficulty in forcing the learner’s hand down on the shock plate, displaying what seems like “total indifference” towards him, punctuated by a moment of annoyance when he scolds him for refusing to answer” (Milgram 1974, p. 46).}
Twenty minutes, two people and a room with an electrocuting machine is all it takes to turn 65% of us into candidates for manslaughter. When framed this way, the fact that 35% disobeyed seems of little consolation, as an unsettling question arises: what if the room had been even more conducive to producing obedience?

A further unsettling fact is that even of those in the ‘disobedient’ category, most still went far beyond what we would consider an acceptable shock. Only half of the 35% who did disobey stopped when the learner began complaining about chest pains, which seems to be a natural defining step on the slippery slope between 15V and 450V. Furthermore, in order to qualify as ‘disobedient’, subjects did not have to outright refuse to continue. They merely had to voice any kind of concern or question four times in a row. At this point, the experiment was terminated by the experimenter.34

Milgram’s study may have been extreme compared to normal life, and it was intentionally designed to elicit a high level of obedience. But it was also heavily limited compared to what it could have been. Milgram had finite funds and resources. His materials consisted of little more than wires, a box with switches, a recording, and there were only two helpers present in most of the iterations. The setting differs considerably to other cases in which a large proportion of normal people act badly, such as in cases of brainwashing, intense cultural pressures or wartime situations. Unlike such cases, in which an agent’s deep desires and values are changed over an extended period of time by forces beyond the agent’s control, the subjects in Obedience were only recently in the ‘setting’ of their normal lives and culture. As a

34 There appears to be some confusion regarding the accuracy of this point. Milgram originally reported that the experimenter used a “firm but not impolite” (1964, p. 374) voice and that the subject was deemed disobedient if they questioned the experimenter four times. Gina Perry’s The Shock Machine gives an alternative interpretation of Obedience based off Milgram’s original transcripts and recordings. Perry’s account of the transcripts and recordings reveals that the ‘experimenter’, an actor and biology teacher named Jack Williams, initially followed this script. But as the trials went over months, Williams himself went down a slippery slope of straying further from the script, saying other things to encourage obedience and requiring more and more refusals to qualify as disobedient (p. 115–119). Perry notes he was particularly aggressive in Condition 20 with female subjects, arguing it qualified as coercion: ‘Williams’s behaviour implied they wouldn’t be leaving the lab unless they got to the end of the shock board’ (p. 116). Perry uses this as evidence to argue Obedience has low external validity, arguing that it only shows people are obedient when they are coerced. I would agree with her interpretation, were it not for the numerous replications in which this coercion was not present. These replications were far more rigorous in terminating the experiment whenever subjects “expressed either verbally or nonverbally a reluctance to continue the study” four times (Burger, 2009, p. 7). For reviews of earlier replications which also properly followed this script, see Blass (1999a); (Doliński et al., 2017) and Miller (1986, Chapter 4).
result, we would not expect subjects’ deeply held beliefs and desires to have been altered in any way that would make them alien or artificial. We take it to be the case that they qualified as agents who could be held responsible for their behaviour prior to the study, and the lack of obvious constraints on their behaviour suggests they were still responsible agents for the duration of the study. Milgram began the study with ideas but no scientific evidence regarding which variables would help or hinder the obedience rate, and he certainly did not have access to later studies demonstrating which other environmental features can increase the ease with which people will shock strangers.

There are two facts that philosophers and psychologists do not adequately emphasise when discussing *Obedience*, but which are important to note. The first is that the authority figure in *Obedience* is not the only factor causing subjects to obey. Milgram’s own iterations show that a number of factors contributed to the standard 65% obedience rate. First, the institutional setting of Yale helped to foster a sense of legitimacy to the task. Second, that there was only one other person to look to for guidance limited the possibility of taking cues to disobey from others (though as *Bystander* shows, the presence of other people could also foster compliance too). Reducing the harm to a mere flick of a switch likely also helped: when subjects had to force the learner’s hand onto a shock plate, obedience fell to 30% (Milgram, 1974, p. 35). The authority figure’s calm tone, saying they would take responsibility, was also reassuring that shocking was an acceptable action. Zimbardo (2007, pp. 273–275) lists ten features of the environment which he believes all likely contributed to the final rate of 65%, including providing more palatable descriptions of the actions (“continuing the task” as opposed to “electrocuting someone”), getting subjects to agree to follow the experiment ahead of time, and creating opportunities for the diffusion of responsibility. One of the most important features is the gradual scale between 15 and 450V which serves as a psychological slippery slope. Starting with a “small, seemingly insignificant first step” and then having gradually increasing steps meant that “no new level of harm seemed like a noticeable difference from the prior level” (p. 274). Evidence that this factor was particularly important comes from the fact that many real-world instances of seemingly ‘ordinary men’ developing into mass killers and torturers have occurred in settings where the demands to cause harm gradually increased over time (Austin, 2016; Browning, 2001; Phillips, 2012).

The second fact worth noting is there are other experiments that involve electrocuting strangers, but which are less famous than *Obedience*. The basic set-up involves having a
subject choose the level at which to shock as a form of punishment for incorrect responses to a given task, though of course no-one is actually shocked in any of them. It has been found that anonymity increases the shocks people deem acceptable, as does being present in a group (Zimbardo, 1969). Subjects administer greater shocks and increase the rate faster over iterated trials when they are with two other subjects, compared to being by themselves. Notably, it was also found that merely ‘overhearing’ (actually a deliberately included variable) the experimenter refer to the soon-to-be shocked learners as “animals” dramatically increased the rate and level of shocks administered. Conversely, hearing the experimenter refer to the learners as “nice” resulted in decreased levels of shocking (Bandura, Underwood, & Fromson, 1975). What is relevant about these experiments is that subjects’ decisions are affected by factors which were not present in the original Obedience experiments. That is, there exist other situationist factors which can increase subjects’ willingness to harm others.

These two facts lead to an unsettling conclusion. Since a number of situationist factors influence the degree to which subjects obey, and since the number of situationist features in the experimental set-up could have been easily increased, this means that the obedience rate could have been pushed even higher with the addition of extra situationist factors. The Obedience studies should not be considered a worst-case scenario or typical obedience rate to any authority figure, but merely one iteration of many possible scenarios with many possible obedience rates. The 65% figure is not a maximum possible obedience rate, it is merely the highest obedience rate demonstrated so far.

When we look at the experiments this way, we can see that there are a number of other potential ways the experiment could have been conducted with even more situationist features at play. First, we could habituate the subjects. Having them shock people up to, say, 150V on one day (at which point the learner would ‘finish’ their assigned memory task) followed by 200V the next day, would make the overall situation, obeying the experimenter, and administering shocks much less foreign. We could make the ‘slope’ between 15V and 450V even more slippery, by removing the clearly delineating point where the learner says he doesn’t want to be here anymore. Half of those who disobeyed in the actual experiment chose to stop here. His cries could be more ambiguous at lower voltages and slower in escalation, though he would still end up complaining about his heart and wanting to be let out. We could have the teacher first observe other people in other rooms shocking other people, helping foster a sense that shocking people is socially acceptable and the subject’s
own situation not so unusual. We could have two authority figures who agree that the teacher ‘has no choice’, thereby creating even more psychological pressure.

We could combine all these differences with what we now know from other psychology experiments. The teachers could overhear the learner being referred to as an ‘animal’. They could be told the experiment features what is called a ‘Resilience and Determination Game’, and that each participant has been selected because they are believed to be able to overcome irrational feelings and distraction when carrying out necessary tasks. Participants could be given a dark lab coat of their own, to match that of the authority figure, and feel like they too have to go along with what their newfound ‘peers’ are signalling is the correct course of action. The rooms could have images of weapons and pictures that portray other people in a negative light. We could choose a learner who is of a visible out-group to the teacher, to help diminish identification with their pain. We could have the teacher believe that the learner has acted rudely or is racist, to help the subject rationalise the shocks as being an

35 Gino, Ayal, and Ariely (2009), amongst others, have demonstrated that our willingness to cheat increases dramatically if we observe other members of our in-group doing so. Diener, Dineen, and Endresen (1975) demonstrated that people will much more readily perform aggressive acts if they observe an aggressive act first.

36 Diener et al. (1975) demonstrated that people will much more readily perform aggressive actions when the actions have been ‘verbally sanitised’. Gambino (1973) provides an extensive analysis of how this is used in public discourse, e.g. where bombing missions are referred to as “servicing the target”. Liberman, Samuels, and Ross (2004) demonstrated that merely changing the name of a game to be played between two players has dramatic differences in whether players choose to work together (observed more when the title was “The Community Game”) or act selfishly (which was higher when the title was “The Wall St Game”). This was despite the game having the exact same rules, goals and means of winning.

37 Frank and Gilovich (1988) demonstrated that dark clothing results in an increase in aggressiveness. Asch (1952) famously demonstrated how the number of people in a group can result in conformity even when other’s judgements are clearly incorrect.

38 Berkowitz and Lepage (1967) found that simply having a weapon in the same room increases the ease and extent to which people will perform aggressive actions. It was later found that a picture of a weapon can prime aggression-related thoughts and increase the ease with which people will display aggressive behaviours (Turner, Layton, & Simons, 1975). As mentioned in Chapter 1, numerous studies purporting to show effects of priming have failed to be replicated, but the ‘Weapons Effect’ seems to successfully replicate according to the current evidence. See Benjamin and Bushman (2016).

39 Leyens et al. (2003) summarises the evidence for and manner in which we perceive out-groups to share less of the essential qualities that make us human. Solomon, Greenberg, and Pyszczynsksi (2000) give a summary of research showing how being reminded of one’s morality increases out-group bias against people with different worldviews.
outcome of living in a ‘Just World’\textsuperscript{40} The teacher could also be guaranteed anonymity, shown to increase the level of harm people will cause.\textsuperscript{41}

None of these potential factors are especially elaborate, complex or coercive. Having two people in coats, with a modified experiment title, and making the learner’s cries more gradual does not change the moral permissibility of electrocuting a stranger to the point of having a heart attack. Nor can these changes in environment be easily rationalised to explain subjects’ change in behaviour. Nevertheless, it seems that making these changes would lift the 65\% obedience rate much higher. We know that these situational features help increase obedience rates when they are present in these other experiments. It seems reasonable to infer that they would still function in this way when transferred to other settings such as Obedience’s specific set-up. Furthermore, it also does not seem a stretch to say that the effects of these factors are additive. That is, if situationist factors A and B are each known to independently increase the chances of behaviour X, having both present should increase the chances of observing X further still.

This characterisation is falsifiable. It could be the case that there is a ‘ceiling’ of disobedience at about 65\%, and that the addition of extra situationist factors has no effect. However, this possibility seems unlikely. Remember that the removal of some factors (e.g. university setting, the unanimity of experimenters) resulted in dramatic reductions in obedience rate. It would thus be an incredible coincidence if the standard Obedience experiment (Condition 5) happened to also correspond perfectly to the total proportion of

\textsuperscript{40} Lerner (1980) has conducted numerous experiments demonstrating we are more willing to assign blame to victims and praise to beneficiaries the more we have a Belief in a Just World (BJW), even in cases of pure chance (e.g. winning the lottery). Of particular relevance is Lerner and Simmons (1966) who demonstrated that people observing someone being shocked in an Obedience-like situation will devalue the learner’s attractiveness and have less compassion the more the victim is seen to be shocked. This seems similar to Mills’ (1958) studies on cognitive dissonance and rationalisation, in which people’s views on cheating are more lenient after they’ve been induced to cheat on a test. It is likely that when playing a more active role in a shocking experiment such as Obedience, BJW and cognitive dissonance make it easier to justify shocking the learner or to rationalise not having disobeyed earlier. For a more contemporary summary and review of the BJW literature see Hafer and Begue (2005).

\textsuperscript{41} Silke (2003) shows that fighters in Northern Ireland who are anonymous and deindividuated are more violent in battle. Watson (1973) demonstrated that tribes are much more violent in battle when fighting in masks. Burnham (2003) found that people act more selfishly in dictator games when they cannot be personally identified to the recipients, and also act more selfishly when the recipients are not identified to them. Zimbardo (1969) demonstrated that subjects who are deindividuated and anonymised are more willing to electrocute strangers at higher voltages, even with no authority figure present.
people who are susceptible to obeying, such that the addition of extra situationist factors had no effect on the obedience rate. Additionally, there are already indicators that higher rates can be achieved in similar experiments, such as the 77% obedience rate observed when the experiment involved actually electrocuting a real howling puppy (Sheridan & King, 1972).

One of the variations in *Obedience* saw the obedience rate increase to 72% when there were two other ‘teachers’ (actually confederates) who “obey[ed] the experimenter’s instructions and mutter[ed] their disapproval if the third teacher … hesitates or refuses” (Perry, 2013a, p. 306). A more distressing, and less well-known, study involved nurses receiving a phone call while at work from someone claiming to be a doctor they’d never met. This ‘doctor’ ordered them to give a patient 20 milligrams of a drug called Astrogen, despite the drug not being on the list of medications to be given that day and the bottle clearly saying that the 10 milligrams was the maximum safe dose. Though 83% of nurses said they would never do such a thing, 95% actually began to administer it to a patient, being stopped before they entered the room (Hofling, Brotzman, Dalrymple, Graves, & Pierce, 1966).

There are four reasons I take this plausible thought experiment to be philosophically important, and for going into as much detail as I have. First, this thought experiment will be relevant to Chapter 6, where I argue that most members of the moral community lack the standing to blame the obedient subjects because we would have acted in the same way. Second, philosophers working on moral responsibility ought to have an accurate sense of how ordinary agents act in various settings. When discussing agency, blameworthiness and moral psychology, it is important for our intuitions to be empirically informed about what things people are or are not responsive to when making ethical decisions. In particular, it is important to recognise that there are a class of factors which can have large effects on behaviour that aren’t obviously morally or prudentially relevant to agent’s decisions.

Third, having a more accurate understanding of these studies has some epistemic value, both to philosophy and psychology. As noted earlier, many psychologists regularly downplay the importance and validity of *Obedience*. Milgram’s explanations were not well-supported by his evidence, and are rightfully criticised for this. But being able to identify which claims are supported by his results ensures that we don’t throw out the baby with the bathwater. These studies and my preceding analysis teach us something about the people around us that we probably would not have arrived at by introspection. There is some value in understanding people’s dispositions and moral psychologies, and gaining a clearer understanding of a study that is so famous and influential.
Fourth, realising that we could achieve such a high rate of obedience, even without coercing subjects, pumps the intuition that the obedient subjects’ rational capacities were impeded in some way. At the very least, many people would reduce their confidence in the claim that these subjects were fully morally responsible for their actions. Some readers may not share this intuition. That is fine; my argument in the remainder of this chapter does not rely on this intuition. However, any proposed explanation for why subjects behave in *Obedience* ought to generalise to this modified experiment too, and it is worth keeping this in mind as I examine some causal explanations of subject behaviour that have been offered.

In the remainder of this chapter, I will argue that existing causal explanations contain important gaps. These explanations were provided in debates over whether the situationist studies act as evidence that people lack global character traits, so let me briefly summarise this debate.

### 2.3 Review: The Situationist Threat to Virtue

Doris (1998, 2002) and Harman (1999, 2000) were the first to examine the implications of the situationist psychological experiments for philosophical conceptions of virtue and character traits. Both philosophers observed that moral behaviour seems to be influenced far more by external situational factors than internal character traits, and both argued that we are therefore unjustified in believing that certain kinds of character traits exist. The most sustained, detailed and influential argument is found in Doris’s book *Lack of Character* (2002), where he argues that globalist virtue-based ethical theories lack empirical support. These are theories which take agents to have character traits with the following properties:

- **Consistency.** Character and personality traits are reliably manifested in trait-relevant behavior across a diversity of trait-relevant eliciting conditions that may vary widely in their conduciveness to the manifestation of the trait in question.
- **Stability.** Character and personality traits are reliably manifested in trait-relevant behaviours over iterated trials of similar trait-relevant eliciting conditions.
- **Evaluative integration.** In a given character or personality the occurrence of a trait with a particular evaluative valence is probabilistically related to the occurrence of other traits with similar evaluative valences.

Doris (2002, p. 22)
Doris notes that virtue theory proponents often point to three features in support of the existence of global traits. First, we observe what appear to be behavioural manifestations of global traits. We can readily point to instances where someone appears to act in accordance with a virtue or vice e.g. by performing brave or cowardly actions. Second, global traits seem to strongly cohere with folk intuitions regarding predictive and explanatory power. When we see someone act bravely or cowardly, we think that global traits provide an explanation of these actions and allow us to make predictions regarding future behaviour. Third, we have a rich language to describe these first two points. We can say that Sally acted bravely (plausible observation of a global trait manifesting), Sally is a brave person (intuitive ascription of a global trait), Sally saved that person because she is brave (explanatory power), and Sally will save that person because she is brave (predictive power).

Doris argues that our beliefs that people act consistently, and that people have evaluatively integrated behaviours, are mistaken. We fail to observe any consistency as situationist features are far more reliable predictors of a person’s behaviour than individual dispositional differences. This is primarily shown in studies such as *Samaritan* and *Bystander*. We expect that whether or not someone stops to give aid will be a product of whether they are a caring person, but we discover that situational factors such as the number of people nearby are a far greater determinant of behaviour.

Doris argues that evaluative integration is shown to be false by referring to a second group of experiments within the situationist literature. I have not yet mentioned these studies because they are not the focus of this thesis. To summarise, these studies show that even when people do act according to the expected features of the situation, whichever way they act is a very poor predictor of how they will act in other similar circumstances. Any given dimension of personality that we observe in one situation (such as extraversion, honesty, curiosity) has only a low correlation ($r = 0.1–0.3$) with being observed in similar situations in which we would expect it to be elicited (Ross & Nisbett, 1991). As a result, people lack cross-situational stability. For example, observing someone being honest on an exam is a very poor predictor of whether they will be honest when presented with the opportunity to steal coins, despite our expectation that honest people would refrain from both cheating on exams and stealing coins (Hartshorne, May, & Maller, 1929).

Due to this lack of consistency and evaluative integration, global trait theories rest on a “foundation of sand” (Alfano, 2013, p. 82) and we have no empirical evidence to support
thinking of them as accurately describing existing features of our psychology. Doris concedes it could be possible that virtue theories are still true as a normative theory of what actions are right or wrong, and as a theory of which traits would be admirable or deplorable. But given that empirical and intuitive plausibility are regularly touted as strengths over other competing ethical theories by proponents such as Brandt (1988) and Macintyre (1984), conceding these supposed strengths were actually absent would still be a major loss for many proponents of virtue-based accounts of morality.

Doris and Harman quickly received numerous responses that fell into three, somewhat overlapping camps. The first response has been to argue that, although the situationist critique is correct, virtues and traits can be salvaged. We may need to reconceptualise how we think of character traits, but they are still very real and present in subjects.

The second type of response has been to argue that these experiments pose no threat to virtue or globalist trait theories, for Doris and Harman have mischaracterised the accounts of virtue they are attacking. Furthermore, even if global trait theories are descriptively inaccurate, virtue theories are prescriptively sound. Alternatively, virtue itself is expected to be rare and trait theories provide a viable guide for ethical reasoning.

While Harman (2009) thinks that virtue trait theories should be abandoned altogether due to a lack of evidence, Doris argues the Stability criterion is empirically defensible. How people act in a certain situation is a very good indicator of how they will act in an almost identical situation, and so we can still argue for the existence of highly specific localised traits. Such a local traits account would still be quite revisionary and unsatisfying to globalist trait proponents, however.

Merritt (2000) argues that, although Doris and Harman’s critique is problematic for Aristotelian virtue ethics, a Humean account of virtue ethics would be able to survive. This is because even if one’s possession of virtues is sustained only by social settings, which could include situationist factors, according to Hume, these virtues are no less genuine than if they were sustained by inner factors as Aristotelian accounts require. Prinz (2009) and Miller (2014) too argue for the existence of traits that are different to the traditional Aristotelian notion. For Prinz, things such as ‘being a fan of jazz’ or ‘being a republican voter’ meet Doris’s criteria for trait-hood, but these are far from the traditional Aristotelian conception. Miller agrees with Doris and Harman that the situationist studies do not support the existence of ‘traditional virtues or vices’ (p. 195) but that we still have ‘mixed traits’ which are neither innately virtuous nor vicious.

For arguments that Doris and Harman attack the wrong conception of virtue, see Annas (2005), Besser-Jones (2008), DePaul (2000), Swanton (2003), Webber (2006). Doris (2005) replies he never claimed to be attacking any particular conception of character, only to be attacking specific notions of character that are common.

The third, and most pertinent to this thesis, has been to question Doris’s interpretation of the studies, or their relevance to virtue theories. These kinds of responses generally take the following approach: they try to provide an explanation for why subjects act as they do, in such a way that is compatible with subjects having global character traits. These explanations are often run together but, in general, there are three main types. These include:

1. Subjects have global traits which are overridden by other desires (Annas, 2003; Athanassoulis, 2000).

2. The obedient subjects are merely acting from a competing trait or virtue (Brink, 2013; Solomon, 2003; Swanton, 2003; Webber, 2007).

3. The obedient subjects (and most people) are akratic, weak-willed (Badhwar, 2009; Wielenberg, 2006; Winter & Tauer, 2006), experience “gaps” between their moral concerns and their dispositions to act (Besser-Jones, 2008) or lack the practical wisdom required to do the right thing.

There are also two explanations from psychologists on why so many subjects in Milgram’s experiment obey which are worth considering. These include:

4. Subjects obey because they identify with the experimenter (Haslam & Reicher, 2012; Haslam, Reicher, & Birney, 2014).

5. Subjects obey because they are inhibited from disobeying by a desire to avoid embarrassment (Sabini & Silver, 2005).

I will now consider each of these explanations and argue that they are all incomplete. In particular, they cannot explain the relative variation in subject behaviour that occurs between conditions. This matters because, if we do not know why subjects act as they do, then it becomes hard to identify whether they are deserving targets of blame.

2.4 Against Desire-Based Explanations

The first proposed explanation says we observe a dramatic relative change in obedient subjects’ behaviour because they experienced a desire to obey the experimenter, which they then acted on. People do not ordinarily experience desires to obey experimenters because they are not ordinarily in settings with experimenters giving orders. Furthermore, subjects (and we) tend to underestimate the strength of the desire to obey that is experienced, which is why psychologists predicted only 1% of subjects would obey the experimenter. Although subjects still had a desire not to harm the learner, this desire conflicted with their desire to
obey the experimenter, and it is the latter desire that was stronger. Once the desire to obey the experimenter is posited as the explanation for the obedient behaviour, the subjects’ decision to obey is understandable. Furthermore, the subjects’ decision to obey does not look like it manifests any kind of incapacity or irrationality on the subjects’ part. The subjects were simply doing what they most strongly desired to do.

Such a desire-based explanation can be applied to Bystander. We can say that subjects who do not help simply experienced a desire to not help, or a desire to conform to what everyone else was doing. This desire renders subjects’ behaviour intelligible in such a way that doesn’t suggest any impairment of capacities. That so many subjects acted in accordance with this desire only strikes us as odd because we underestimate the strength of the desire to follow everyone else that we would feel in that situation.

I don’t think this is a convincing explanation of subject behaviour in all of the experiments. If we apply such reasoning to Fire, in which subjects remained in a room that was filling with smoke, this would entail that subjects had a desire to conform with others that was so strong it outweighed their desire for self-preservation. Had the smoke been coming from a real building fire, this decision would have resulted in their own death. That subjects simply chose to act on this desire instead of their desire for self-preservation not only seems unlikely, it would entail subjects’ reasoning fell below the minimum standards of rationality we take most agents to have. This would give us reason to question subjects’ decision-making capacities.

Desire-based explanations also appear unable to account for the influence of factors such as coins, smells and sounds. Having a desire to help others that is conditional on having found a coin or been exposed to a pleasant smell is very odd. Although we can observe that finding a coin may lead to a desire to help others, the responses to Doris do not offer any detailed explanation for how and why this occurs. The smell of food can make me desire food, and an offer of money can make me desire to do things to get that money, but the coins and smells in these experiments produce desires that have nothing to do with coins or smells. Furthermore, the fact that there is a coin or pleasant smell does not seem to enter subjects’ conscious deliberations. Without some further explanatory story, the purported

---

46 Interviews with survivors from the T1 building of the 2001 World Trade Center attacks found that, on average, workers waited 5 whole minutes after the first plane hit before deciding to leave their floors, with some people waiting up to 45 minutes. See Averill et al. (2005).
desires seem disconnected from the agent’s goals and deliberative processes. Subjects in *Dime* likely wouldn’t have been more likely to give help if they’d found a worthless silver disc of similar size and weight to a dime. It is reasonable to want an explanation for this seemingly unusual pattern of desire formation, before we conclude that subjects’ deliberative processes were not impeded in some way.\(^47\)

Reflection on studies such as *Dime* and *Bystander* highlight what is distinctive about the actions of the subjects in the situationist experiments. The challenge I am addressing is not merely that of explaining why many people obey when ordered to by an authority figure, or why many people fail to act in the morally right way that we expect, or why they fail to behave consistently across situations.\(^48\) We must also explain the dramatic relative variation in behaviour that occurs as a result of seemingly innocuous changes between conditions. For example, we ought to be able to explain why the gradual increase in the shocks is so effective at leading people to administer shocks at 450V. It seems likely that, had the participant demanded to be let out after the second shock, and complained of heart trouble after the third, that many more subjects would have successfully disobeyed. We also ought to be able to explain why the obedience rate drops considerably when the experimenter gives the orders from outside of the room.

One way to salvage a desire-based explanation is to take the relative variation in observed behaviour as evidence that subjects’ desires are quite specific. It could be that subjects have non-instrumental desires to obey experimenters that are contingent on certain other conditions being met. This could occur in the same way that, say, my inconsistency in buying ice cream when given the opportunity is explained by my having a desire to eat ice cream that is contingent on the weather being hot and there being enough money in my wallet. I don’t think this reply is promising. When we specify the kinds of desires that subjects are purportedly experiencing, they seem very odd. We are forced to say that subjects have a

---

\(^{47}\) Doris (2002, p. 30) and others suggest that finding the coin puts subjects into a happier mood, and this is what leads to them deciding to help. I’ll return to this explanation in Chapter 4.

\(^{48}\) Arpaly (2005) is one of the few philosophers who correctly, albeit briefly, identifies the relationship between particular situationist factors and subjects’ behaviour as cause for concern: “But one implication of Doris’s data strikes me as very disturbing to moral psychologists, not only to virtue ethicists but also to Kantians, and to ordinary moral talk … Doris’s data threaten to wreck this kind of scale … We cannot always assume that a hundred-dollar bonus would influence our behavior more than a dime found in the phone booth … not only are situations more important than we thought, but one cannot even predict which feature of one’s situation will effect [sic] us more than others” (pp. 646–647).
desire to obey experimenters who are within a certain proximity but not experimenters who are further away, or a desire to obey experimenters whose commands increase gradually over time but not experimenters whose commands increase sharply. Such desires seem arbitrary and irrational. The factors which my desire (e.g. to obey the experimenter) is conditional upon or sensitive to (e.g. his orders being gradual) appear to lack any reasonable connection to each other, or any other desires.

Additionally, if subjects were merely acting in accordance with a desire to obey the experimenter, we would expect subjects to act in the same manner in similar situations when similar desire-relevant features are present. For example, we might think that subjects obey because they have a desire to avoid the embarrassment that comes with disobeying, and so they act on this desire. But as Sabini and Silver (2005, p. 559) note, were we to give the subjects a choice between electrocuting-a-stranger or five-minutes-of-embarrassment, almost no-one would choose the former. Similarly, even if subjects experience a desire to avoid confrontation or to obey an experimenter, it is unlikely they would obey if such an experimenter were to suddenly approach them in the street and order them to begin administering shocks to the learner. Even saying subjects experience ‘a desire to obey’ is somewhat misleading, because this implies they have a desire to obey whatever order that is given. But Sabini and Silver point out that “if the experimenter were, in the middle of the experiment, to insist that the subjects stop and shine his shoes, [subjects] would not do so” (p. 550). The content of this supposed desire thus stands in need of further explanation too.

Arguments that subjects obey the experimenter because they have a desire to obey thus contains an important explanatory gap. To summarise: although it makes sense to claim that situationist factors could somehow lead to an increased desire to obey, a supplementary story is needed regarding how this occurs, given that the effect of many situationist factors on subjects’ decisions seems to be unconscious. We need an explanation for why subjects’ desires are sensitive to specific environmental factors in the experiment, since many of these environmental factors seem irrelevant to those desires and to the agent’s other beliefs and goals. We also need a more specific account of the content of the ‘desire to obey’ that subjects experience in Milgram’s study, given that the obedient subjects would plausibly disobey many other types of orders.
2.5 Against Trait-Based Explanations

This difficulty with explaining the relative variation in subject behaviour also poses a challenge for trait-based explanations. Solomon (2003, 2005), Brink (2013) and Webber (2007) argue that individual situationist factors do change which traits people act from, but these traits were always present in the subjects. Similarly, Swanton (2003) explains the behaviour of the obedient subjects by saying, “where virtues such as trust and fidelity appeared to conflict acutely with another – non-maleficence – many subjects gave insufficient weight to the latter” (p. 31). Some situationist factors are clearly relevant to which traits people will act from. For example, in Bystander, the number of people who are nearby will be relevant to those people who have the trait of conformity. As a result, the observed behaviour in this setting can be explained by appealing to traits even though subject behaviour changes dramatically, conflicts with subjects’ expressed desires to help, and seems difficult to rationalise. But again, providing a full causal explanation of how situationist factors change which traits subjects act from in other studies is much harder. In explaining subject behaviour, we’re interested in why this particular change in environment causes subjects to act on that particular trait, instead of another. 49

Consider Obedience. Even if subjects are acting from the trait of obedience which “virtually all of the subjects had been brought up with and practiced every day since kindergarten” (Solomon, 2005, p. 653) or “acting from a [competing] disposition … to cooperativeness” (Kamtekar, 2004, p. 473), a supplementary story is still needed. When we look at the obedience rate between various experimental conditions, we are left wondering why a slipperier slope of escalating demands increases the number of subjects who display the trait of obedience. Further explanation is also needed to explain why factors such as coins and smells (or lack thereof) are sufficient to make people suddenly choose (or fail) to display their trait of helpfulness. Adding a coin or changing the title of a task doesn’t suddenly produce a new situation with new morally relevant reasons which a different trait should respond to. If traits are dispositions to act, or to feel and judge in response to settings, then it needs to be explained why subjects’ traits respond so dramatically to these factors in particular.

49 If talking about traits in this manner appears too mechanistic because people also use their practical wisdom in deliberating, as has been argued previously (Annas, 2005, p. 637; Kamtekar, 2004, p. 460; Webber, 2006, pp. 208–209), a similar objection is made below regarding subjects’ level of practical wisdom.
It also seems worrying that which traits guide subjects’ behaviour in this experiment doesn’t correspond to the degree to which we’d expect certain groups of people to possess those traits. *Obedience* has been replicated on people with different genders, ages, careers, and from different cultures, and none of these things seem to predict whether a particular subject will disobey, despite there being character trait differences between these groups of people. It seems particularly relevant that, as mentioned earlier, Hofling et al. (1966) found that 95% of nurses in their study – who are presumably more compassionate than the average person – were induced to obey harmful orders in their workplace.

It is possible that subjects’ traits are particularly fine-grained. Instead of the global trait of helpfulness, someone could theoretically have the local trait of ‘helpfulness-when-a-coin-is-found’. Doris (2002) thinks such traits are empirically defensible. Such traits would allow us to explain why individual situationist factors influence subject behaviour. However, a few considerations count against taking this strategy. The first is that this approach gives up the idea that traits are broad-based, which many virtue ethicists are committed to. Which account of traits we endorse is not simply a matter of preference. We rightfully care a lot more about the non-existence of global traits than the non-existence of local traits, because our moral standards make us more concerned with behaviour across global-trait eliciting settings.50 We not only expect that people help in situations where help is needed, we demand that they be helpful in such situations. The fact that help is needed is important to us in ways that the fact a coin was present is not.

A greater worry with adopting a local trait explanation concerns whether these traits actually play any explanatory role. Doris acknowledges that, while fine-grained traits might be great at predicting behaviour, it is reasonable to question their explanatory power: “I deduced the existence of the trait … then turned around and invoked the trait to explain the behavior. My approach reflects a credible predictive strategy – again, past behavior is a pretty good guide to future behavior – but it is not an enlightening explanatory strategy” (p. 66). His response to this problem is to argue that worries about fine-grained traits being uninformative also afflict globalist trait accounts, given we also ascribe global traits to people

---

50 Doris (2002, p. 80) describes this concern in more detail: “Talk of consistency or inconsistency simpliciter is meaningless… The real question concerns what regularities, or failures of regularity, should interest us … [moral inconsistencies] aren’t easy to ignore, because they emerge in reference to socially shared standards for interpersonal conduct … Without compelling argument to the effect that the standards expressed in traditional moral trait names are misbegotten, inconsistency with regard to them is important for moral psychology.”
based on observations of their past behaviour. The reason this is not thought to be a problem for global trait theories is that we can always provide explanations for those traits in other ways. Doris argues we can do this for local traits too: “A more enlightening explanation appeals to motives, goals, values, attitudes, strategies, or whatever else it is that forms, for a given individual, the psychological context of the trait … But the point is that there is nothing about local trait explanations inimical to adducing such contexts” (p. 66).

This response may allow that some local traits have explanatory power, but I don’t think it can generalise to the kinds of local traits needed to explain the subject behaviour I am concerned with. This is because, if we start to cash out these traits in terms of ‘motives, goals, values’ etc., we run into problems similar to those faced by desire-based explanations. Recall these local traits will be things like helpfulness-when-a-coin-is-found or obedience-when-an-authoritative-figure-is-within-a-certain-proximity. Ascribing to subjects the goal or values of helping-when-a-coin-is-found or valuing obeying-when-an-authoritative-figure-is-within-a-certain-proximity simply doesn’t seem plausible. Such goals, values or attitudes seem too gerrymandered, and fail to track recognisable patterns of reasons.

Without some further story about how local trait-based explanations of subject behaviour can be explanatorily powerful, trait-based explanations are insufficient. We still need more details about how specific situationist factors are relevant to subjects’ decision-making, plausibly with reference to subjects’ motives, goals, or attitudes. No such account has yet been provided.

2.6 Against Akrasia- and Practical Wisdom-Based Explanations

A popular explanation of the obedient subjects’ behaviour is to say that they experience weakness of will or akrasia, understood as the common phenomenon where agents act against their better judgement.51 According to this explanation, subjects judge they should not obey, but obey because they experience “psychological forces” or feel the “pressures of the situation” (Wielenberg, 2006, p. 474; Winter & Tauer, 2006, p. 79, respectively).

---

51 This characterisation is often thought to be too broad (Holton, 1999; Watson, 1977) because there are a variety of ways in which agents can act against their better judgements. The phrase ‘weakness of will’ can refer to a variety of phenomena e.g. intentionally acting against one’s considered judgement, ‘losing one’s nerve’, and changing one’s better judgement in the face of temptation. I recognise my characterisation here fails to spell out these nuances, but it reflects how philosophers responding to Doris have used these terms. Additionally, that there are different phenomena within this category doesn’t affect my objection.
Attempts to explain subject behaviour as the product of akrasia induced by ‘psychological forces’ still contain important gaps. In many cases of akrasia, when agents act against their better judgement, they do so by acting on some other desire. If my better judgement is that I should follow my diet and I then act akratically, my desire to eat unhealthy food usually features strongly in an explanation of my behaviour. If I am too weak-willed to deliver a speech in front of a crowd, my behaviour can be explained by saying I experienced a desire to avoid being in a situation in which I risk embarrassment. If philosophers think subjects obey because they are weak-willed in the presence of a stronger desire, these philosophers’ explanation run into the problems with desire-based accounts raised earlier.

These philosophers might try to get around this problem by saying some akratic actions can occur without the presence of any corresponding desire, and that the obedient subjects’ behaviour in Obedience is one such example. Nevertheless, saying these subjects are akratic is still insufficient as a causal explanation, because we are trying to explain the relative variation that occurs between conditions. Philosophers who adopt an akratic explanation are yet to provide such an explanation, other than gesturing to vague ‘psychological forces’. If we want to explain why factor A caused an agent to φ instead of ψ, saying ‘they judged ψ-ing to be the right action but φ-ed instead’ doesn’t answer our question.

Arguments that subjects’ behaviour can be explained by saying subjects simply haven’t cultivated the required virtue for that situation, or do not possess the level of practical wisdom required to do the right thing, face a similar problem (e.g. Annas, 2003; Athanassoulis, 2000; Reed, 2015; Sosa, 2009; Winter & Tauer, 2006). Pointing to a lack of practical wisdom only initially appears to explain subject behaviour because we have a number of implicit assumptions about subjects’ desires and traits which normally allow us to fill in the explanatory gaps. To show what I mean, imagine a case in which we observe the subjects singing to the experimenter in Obedience while administering the shocks. This would strike us being very odd, precisely because we don’t have any background assumptions to fill in the details. Simply saying subjects lacked the practical wisdom to disobey in this setting would not function as an adequate explanation.

2.7 Against Identity-Based Explanations

Finally, let’s consider how contemporary psychologists explain why so many subjects obey. Haslam and Reicher argue in a number of articles (Haslam & Reicher, 2012; Haslam,
Reicher, & Birney, 2016; Reicher & Haslam, 2017) that the subjects in Milgram’s experiments are not obedient subordinates so much as they are ‘engaged followers’. They argue that “people’s willingness to accede to the requests of others is predicated upon social identification with them, and an associated sense that they are legitimate representatives of shared group goals, values, and aspirations” (Haslam, Reicher, Millard, & McDonald, 2015, p. 60). Their evidence for this is that, when looking at the obedience rate across conditions, a very strong correlation exists between the obedience rate and the extent to which we would expect subjects to ‘share’ the perspective of the experimenter (Haslam et al., 2014). For example, when orders were delivered by phone, the obedience rate was only 20.5%, plausibly because subjects found it much harder to identify with the experimenter’s aims.

Construing the obedient subjects’ behaviour as ‘engaged followship’ seems mistaken to me for a few reasons. First, the account equivocates between motivating and explanatory reasons for action. In many places, Haslam and Reicher take identification with the experimenter to be subjects’ reason for action. For example:

At the 150-volt point they become aware of a competing social identity (as moral citizens in the world) … The destructive behavior within the Milgram paradigm is a reflection not of simple obedience, but of active identification with the experimenter and his mission.


But as Sandis (2015) argues, things like biases, framing effects and identification with a group are not peoples’ motivating reasons for action, because these are not the reasons agents act on. If we asked the obedient subjects, ‘Why did you obey?’ they would not report, ‘Because I identify with the experimenter’. Instead, any purported identification with the experimenter only causes subjects to perceive certain reasons and not perceive others, or to perceive certain reasons as being stronger or more choice-worthy than others. Identification with the experimenter or learner may contribute to subjects acting as they do, but it is not their reason for acting.

Second, the account has excessive latitude regarding potential targets of identification, which makes the account nearly unfalsifiable. For example, recall that the obedience rate dropped when the experiment was conducted in a warehouse, purportedly by a private
company rather than a university. Rather than taking this to be evidence against the claim that subjects’ obedience is explained by their identification with the experimenter, Haslam and Reicher explain this by saying subjects had less identification “with science” (p. 319). Similarly, recall that the obedience rate dropped as a result of seeing other teachers refuse to continue. This can’t be accounted for in terms of identification with the learner or experimenter, since the learner and experimenter behave in the same way in this variation. But instead of taking this to be evidence against their theory, Haslam and Reicher say that the drop in the obedience rate was caused by subjects having greater identification with “the general community” (ibid.).

Third, ‘identification’ seems to be too strong a term given the anxiety subjects felt, the fact that many subjects emphasised the correct answer to the learner (which would presumably invalidate the ‘results’ and thus thwart the experimenter’s goals), and how easily this ‘identification’ with the experimenter is lost. Haslam and Reicher use the term ‘identification’ to connect their explanation to a broader account of social identity theory, which explains peoples’ motivations in terms of an internalised sense of themselves as members of an in-group. They could thus say they are using the term semi-stipulatively. But this understanding of ‘identification’ risks not doing any real explanatory work, because we are left without an answer to why certain factors cause subjects to change which party they identify with, or how these ‘internalised’ senses of self can be gained or lost so quickly.

For example, suppose we want to know why many subjects disobeyed when the learner mentioned heart problems. Haslam and Reicher would say at this point that the subjects identified with the learner more than the experimenter. But then we’d want to know why they identified more strongly with the learner here, but not earlier. Subjects surely don’t see people-who-have-heart-problems as some special in-group. To answer this, we have to point to something in subjects’ subjective construal of their situation, or the salience of certain considerations at that point. But once we introduce these factors, we no longer need to reference identification with the experimenter and identification with the learner in order to have a comprehensive explanation of why subjects obey or disobey. We can just say that, at that point, or in a particular condition, the reasons favouring disobeying became much more obvious to particular learners, and this is why they disobeyed.
2.8 Against Embarrassment-Based Explanations

Psychologists Sabini and Silver (2005) are critical of the explanations given by Doris, Harman, and Ross and Nisbett. They complain that describing subjects’ as being influenced by ‘subtle forces’ is “vague and misleading” (p. 560). Sabini and Silver believe there is a common thread that runs through many of the experiments, where giving aid or leaving a smoke-filled room carries with it a risk of embarrassment, and this risk of embarrassment explains their failure to act. At first glance, their account looks like a desire-based account. People generally desire to not be embarrassed, and so it could be that people value avoiding embarrassment so highly that they ignore the moral requirements of each situation. But Sabini and Silver foresee and reject this interpretation. As noted earlier, were subjects offered the choice between electrocuting someone or suffering five minutes of embarrassment, they would overwhelmingly choose the latter. Rather, according to Sabini and Silver, subjects are confused by other people’s apparent interpretations of the situation. Subjects are “thrown into doubt about their view of the moral world by the differing opinion held by the seemingly normal experimenter” (p. 554). The experimenter and the bystanders seem to be acting as if they see no need for concern, and this confuses the subjects who take there to be great cause for concern. Then, when subjects are confused by other people’s apparent interpretations of situations, fear of embarrassment “inhibits them from acting, without their articulating to themselves that they really have nothing to fear but embarrassment” (p. 559).

Sabini and Silver’s account ultimately suffers from the same criticism they level at Doris: vagueness. Explaining why an agent didn’t help by saying ‘her behaviour was influenced by subtle forces – the presence of four others’ is criticised for being too vague. But their replacement, ‘her moral understanding of the world was confused by the presence of four others, and her fear of embarrassment then inhibited her from helping’, is hardly any better. If philosophers must “go beyond metaphor and to say what those subtle situational forces are … in terms of [the agent’s] particular beliefs and desires” (p. 560) then so do Sabini and Silver. They need to specify in what way people’s “understanding” can be “confused” by other’s actions, and second, how exactly subjects are “inhibited” by a fear of embarrassment.

One problem regards the metaphorical language when accounting for how the experimenter affects subjects’ perceptions of the situation. According to Sabini and Silver, “the experimenter shows the subjects a vision of the moral order’ which confuses subjects’ “moral world” the result of which was that “the subjects … lost their moral rudders” (p.
To be charitable, these explanations with this metaphorical language may be sufficient for Sabini and Silver’s purposes. They are only aiming to show how the results of the situationist psychology experiments are compatible with the existence of global character traits, and their language may be specific enough to achieve this (cf. Doris, 2010). But for our purposes the explanation is lacking. We need to know what effect individual situationist factors have on subjects’ decision-making, and simply observing that subjects’ moral world is confused is not helpful.

A second problem is that, even if we accept their metaphorical language, several details seem difficult to reconcile with it. For example, our understanding of the ‘moral world’ at least partly includes descriptive facts such as whether an action is causing harm or not. Milgram’s subjects seem to understand perfectly well that they are causing harm, that they are being ordered to cause harm, that the learner has requested the experiment stop and that the experimenter wishes the study to continue. It seems that, if asked, subjects would be able to report all of these things, indicating these are things they are not confused about at all. Interpreting claims about ‘the moral world’ to refer to purely normative claims such as ‘causing harm is wrong’ also seems inaccurate. The fact that causing harm is wrong seems to be the main reason subjects are so distressed at what they are doing, and this wrongness is an important part of why they try to disobey.

Finally, it is hard to make sense of the claim that, after becoming confused, a desire to avoid embarrassment ‘inhibits’ subjects from disobeying. Because this supposed desire is not something subjects consciously consider, it is unclear what exactly this inhibition amounts to, and how this desire ‘inhibits’ aside from making disobeying more “difficult” (p. 553).

2.9 Reason-based Explanations

Because existing explanations of the obedient subjects’ behaviour are insufficient, we need an alternative. One concept which surprisingly doesn’t feature much in the explanations cited thus far is that of perceived reasons or judgements. In addition to desires, traits, and practical wisdom, our judgements about reasons play a large role in our deliberations, and so it seems promising to try and explain subject behaviour by appealing to their perceived reasons for action. Not only is this prima facie plausible, I believe that shifting our explanatory focus to subjects’ judgements and perceived reasons for action allows us to identify what was initially appealing about desire and trait-based accounts.
The desire-based explanations were appealing because many philosophers take our desires to be based on what reasons we take to be present (e.g. Gregory, 2017), and in some cases, the fact you have a desire itself provides a further reason for action (Darwall, 2003). Some philosophers also argue that our beliefs about reasons can give rise to new intrinsic desires, where we desire certain things for their own sake (M. Smith, 1994, 2003).

Additionally, many behavioural traits such as obedience and caringness are dispositions to judge, feel and act in response to certain features of a situation. If an agent’s trait plays a role in explaining their behaviour, it will be because this trait was responsive to the reasons (or lack thereof) a person took to be present (or absent). And in citing subjects’ lack of practical wisdom as a partial explanation, we’re implicitly saying that they didn’t exercise or possess the capacity to recognise or act on the reasons that were present in the situation. The presence of the experimenter, or the prestige of the institution could affect subjects’ judgements about how to act. Subjects may construe the situation, the moral demands they are subject to, and the possible choices that are available differently to how we expect.

Similarly, Haslam and Reicher are correct in saying that a number of features of the experiments make different considerations more or less salient to subjects. They are also correct that we can group these considerations into two rough clusters: those relating to goals and interests of the experimenter, and those pertaining to the learner’s interests. And I would agree that the extent to which any subject implicitly thinks of themselves as sharing the experimenter’s scientific ends could help explain why they act on certain considerations. But in order to make a more complete explanation for why so many subjects obey, and why they obey more in particular settings, we need to make specific reference to their subjective construal of the situation, and which considerations subjects were more or less aware of.

---

52 E.g. suppose you, an adult, are deciding what to eat, and your parents recommend something you don’t want. The fact you desire some other food can itself provide you with a reason to dismiss their recommendation.

53 Some philosophers take our desires to be based on what reasons we take there to be, and this could be a way of modifying the desire-based explanations provided earlier. Such a revised account could overlap with my own. However, the philosophers responding to Doris who cite desires in their explanation do not identify whether they take desires to operate in this fashion.

54 Kamtekar (2004) argues subjects did not construe their actions as being wrong, but then uses this to argue for scepticism of our ability to attribute any global traits or lack thereof to subjects, since we cannot access the way those subjects construed their situation. Such scepticism isn’t useful in discussions of moral responsibility, because we lack access to all wrongdoers’ subjective construal of their situation.
My aim over the next two chapters is to provide a detailed explanation of how situationist factors affect subjects’ judgements and thus their decision-making. However, I want to do this by first answering a different question. This regards how we assess an agent’s capacities to recognise and respond to moral reasons in a given situation. In the next chapter, I will consider some evidence that subjects possessed the capacity to avoid wrongdoing in the situationist experiments, but simply failed to exercise this capacity. I will argue that this evidence is misleading, and that we need to reconsider how we assess when an agent has a particular capacity. This will then allow us to answer why subjects act as they do in these experiments, and how situationist factors affect behaviour.

2.10 Conclusion
In this chapter, I examined the situationist experiments and causal explanations of why subjects act as they do. First, I outlined one possible implication of the situationist experiments which has not yet been noticed by philosophers. I argued it would be easy to develop a modified Obedience-like experiment in which over 95% of people would commit egregious wrongdoing. Such an experiment would be very similar to the original Obedience experiment in that it would not involve coercion or duress, but it would achieve an even higher rate of wrongdoing. This kind of result would make it somewhat more plausible that the subjects are somehow impeded from doing the right thing.

I then turned to examining existing causal explanations of subject behaviour. I examined responses to Doris’s and Harman’s argument that global trait theories lack empirical support. These responses try to explain subjects’ behaviour in such a way that does not undermine the existence of global character traits. These explanations included referring to subjects’ desires, traits, lack of practical wisdom, akrasia or identification with the experimenter. I argued that all of these accounts suffer from important explanatory gaps, and we currently lack a sufficiently detailed explanation for why subjects act as they do in these experiments. I suggested that a plausible explanation could be developed by thinking about subjects’ judgements and perceived reasons for action, things which have received relatively little attention in the existing literature on the situationist experiments.
Chapter 3: Reasoning About Capacities

3.1 Introduction

Having reviewed philosophers’ previous analyses of the situationist experiments, I will now examine how we assess people’s capacities. Most philosophers writing on the situationist experiments implicitly take the actions of subjects who act well in these studies to be indicative of the capacities of those subjects who act wrongly. This inference regarding subjects’ capacities is then used to make inferences about subjects’ traits, beliefs, desires and decision-making processes. I believe this inference is unwarranted, and requires critiquing before we can examine whether the subjects who commit wrongdoing in these experiments are blameworthy.

In §3.2, I argue that many philosophers writing on the situationist experiments make an implicit assumption regarding the capacities of the subjects in these settings. This is that, as some people in Obedience were able to disobey, both the obedient and disobedient subjects had the capacity and opportunity to avoid wrongdoing. This means that the obedient subjects are blameworthy for their actions. I argue such reasoning would lead to problematic assessments in other cases where we are assessing agential capacities. For example, it would lead us to conclude that people with dyslexia simply aren’t trying hard enough at reading. We thus need to find another means of assessing when agents lack a particular capacity.

In §3.3, I consider some attempts to assess when agents possess or lack a certain capacity. Ciurria (2013) and Doris (2007) use the reasonable person standard, and Brink (2013) appeals to relevant counterfactuals. I argue that these approaches face problems regarding the reference class from which we apply the reasonable person standard, and which kinds of counterfactuals count as relevant. We should therefore reject these approaches, and thus need another method of determining whether subjects who act wrongly in the situationist experiments had a reduced capacity to do the right thing.

In §3.4, I argue that our best method for assessing whether an agent had the capacity to \( \phi \) is to investigate what is the best explanation for their failure to \( \phi \). This relies on examining agent behaviour in a range of circumstances where we already know something about the agent’s motivations and, from there, inferring what kinds of things impede agential capacities. To give some examples of how we use this reasoning to assess an agent’s capacity to recognise moral reasons, in §3.5, I canvas some explanations for why addicts relapse so
often, and show that addicts have a significantly reduced capacity to avoid wrongdoing. I argue that the capacity to perceive a situation in a particular way, or to see certain considerations, is important for our capacity to recognise moral reasons. In §3.6, I outline the phenomenon of goal neglect. This acts as a case study in how agents can accurately perceive moral reasons, have the general capacity to act on those reasons and yet, in some circumstances, lack the specific capacity to act on those reasons.

3.2 Capacity Comparisons

A common observation is that in no situationist experiment do 100% of subjects act wrongly. This has been explicitly noted by Adams (2006, p. 149); Badhwar (2009, p. 261); Brink (2013, p. 129), Nelkin (2005, p. 193); Solomon (2003, p. 56; 2005, p. 653) and Wielenberg (2006, p. 448). Most of these philosophers use the fact that some subjects did the right thing as evidence that the subjects who acted wrongly had the capacity to do the right thing too. Additionally, nothing seems to physically prevent the obedient subjects from exercising any capacities they possess. As a result, many of these philosophers argue that nothing of moral relevance distinguishes these experiments from other settings where agents are faced with a choice between acting wrongly or rightly. The implication from these observations, and the responses to Doris, is that all the subjects in the situationist experiments are morally responsible for their actions, and thus eligible for ascriptions of praise or blame.

Let’s keep our focus on Milgram’s Obedience as a central case study for a few reasons. First, this experiment has received most of the focus from other philosophers, and this allows me to engage directly with their arguments and interpretations of this experiment. Second, it is the study that most strongly violates our intuitions about how subjects will and ought to behave. Differences between subject behaviour in this setting and subject behaviour in other settings are much more pronounced, making it easier to examine differences in subjects’ decision-making processes. Third, the actions of the obedient subjects in this experiment are much more wrong than the actions of subjects in other situationist experiments. As a result, much more is at stake. It is more important that our analysis of subjects’ responsibility for their actions is accurate, because we risk either blaming agents who are excused or excusing agents who are in fact culpable.
Though most philosophers responding to Doris and Harman don’t explicitly argue that all subjects in *Obedience* had the capacity to avoid wrongdoing, there are two reasons to think they accept this position. The first is that rejecting such a position would provide a far simpler refutation of Doris’s and Harman’s attack on globalist trait theories. If any of these philosophers thought that the subjects had a reduced capacity or opportunity to avoid wrongdoing during the experiment, or that subjects’ decision-making capacities were in some way impaired, this would be sufficient ground for arguing these experiments do not count as trait-eliciting conditions. Settings which undermine an agent’s rational capacities simply are not fair tests of whether an agent has a global character trait.

The second reason is that these philosophers’ interpretations of the experiments make more sense if we assume they hold this position. Arguments that subjects act akratically or display some trait typically presuppose that subjects have a sufficient capacity to act in accordance with their best judgement or to act from a different trait. The person who judges that they should help a stranger but who is physically unable to help, coerced into not helping, or incapable of doing otherwise due to fear does not act akratically.

Against this position, it is worth noting that the mere fact that some subjects did the right thing doesn’t directly show that other subjects could have avoided wrongdoing. Recall we’re examining the behaviour of two groups of people, and trying to ascertain whether one group – the obedient subjects – had the capacities necessary for responsible agency. That we observe some subjects disobey isn’t what leads us to think that the obedient subjects could have too, as there are plenty of cases where we’d agree people could do otherwise without observing anyone in fact doing otherwise. Nelkin (2005) points out that if 100% of customers who enter a bank wait in line, this does not show that they were unable to do otherwise or somehow forced to act as they did by features of the banking environment.

If we’re treating the capacities of the obedient subjects as an open question, the fact that some subjects had the capacity to disobey (shown by the fact that they actually disobeyed) only provides evidence regarding the capacities of the obedient subjects if we already believe that the disobedient and obedient subjects were relevantly similar. But this is exactly the kind of premise someone who thinks that the obedient subjects are excused is going to reject. This does not imply that the question is being begged against them. Thinking the obedient and disobedient subjects are relevantly similar and thus both had the capacity to do the right thing is a reasonable assumption to make. Rather, pointing to the fact some subjects did the right thing is a means of saying the burden of proof is on those wishing to show the obedient
subjects had a diminished capacity to do otherwise. In many other ordinary settings, we’d be inclined to judge that the obedient subjects are relevantly similar to the disobedient subjects in a number of ways, and we wouldn’t attribute a lack of capacity to \( \phi \) to an agent simply because they failed to \( \phi \).

A default presumption in favour of thinking the obedient subjects had similar capacities to the disobedient subjects initially seems reasonable. However, I believe that we should reject this presumption. This is because it seems to generate incorrect verdicts when applied to other important cases. For example, assuming that people who \textit{seem} similar have similar capacities would originally have lead us to conclude that people with dyslexia simply aren’t paying enough attention, and people with depression who fail to get out of their bed are simply lazy. We could also reach the wrong conclusion regarding people with extreme phobias or brain damage.

To avoid such erroneous conclusions, we need to examine what kinds of evidence lead us to ascribe the presence or absence of a capacity to an agent. Applied to our assessment of subjects in the situationist experimental settings, the relevant question now is this: do situationist factors affect people’s behaviour in the way that dyslexia affects a person’s capacity to read, or do they affect people’s behaviour in the way that a bank environment affects people’s capacity to stand in line? Relying only on observable evidence, how can we tell if any individual is not reading well because they’re dyslexic and have a reduced capacity to read, or because they’re simply inattentive?\textsuperscript{55}

3.3 The Reasonable Person Standard and Relevant Counterfactuals

Let me consider two attempts at answering when we should think an agent has a capacity. Though both face objections, the way they fail is instructive.

The first attempt comes from Ciurria (2013) and Doris (2007) who argue that, although not definitive, the number of people who act a certain way in a particular situation can help us determine what can be expected of a reasonable person in that situation. In everyday life and in our legal system, we often make culpability judgements according to what we intuitively expect of a reasonable person, rather than assessing each person’s capacities on an individual basis. We don’t have access to other people’s minds to assess whether their

\textsuperscript{55} Where ‘inattentiveness’ is understood as a trait for which someone can be held accountable, rather than a capacity for inhibition control that an individual has not yet developed.
apparent quality of will was genuine or whether they were reasons-responsive, so we require some other standard by which to judge people’s actions. Ciurria argues that how most people behave in a situation can be an indicator of an individual’s capacities and opportunity to exercise those capacities, as we take most people to be reasonable. If the majority of people fail to act a certain way in a particular setting, this is evidence that we could not reasonably expect most people to have done otherwise in that setting. Doris points out that, although it would be misguided to determine excuses entirely by the rates at which people act, “still, reflection on base rates helps determine what can fairly be expected of a particular individual in particular circumstances; surely it is partly because most people yield under torture that it seems unfair to hold victims responsible for failing to resist it” (2007, p. 527).

The reasonable person standard is deliberately vague, which allows a flexibility to accommodate the many context-specific factors relevant to our judgements regarding capacities and moral responsibility. If we find there is too great a discrepancy between how the average person actually acts and how we expect them to act, this can be evidence that our expectations are misinformed, or that we have not taken some factor into account in forming these expectations. If 100% of people act a way that we think they ordinarily would not (such as electrocuting a stranger at dangerous levels), then Ciurria believes this would serve as strong evidence most people lacked the capacity or the opportunity to do otherwise in that setting. Ciurria herself thinks that the 65% obedience rate in Milgram’s study could “potentially excuse subjects, since all of them surpassed the expected threshold of 150 volts, and they were acting in unusual circumstances under pressure from an apparent authority figure” (2013, p. 190).

The standard of the reasonable person alone is not enough to assess the responsibility of subjects in situationist experiments. This is because the rates at which people behave in certain ways are not the only thing that is considered in determining how a reasonable person could have been expected to act in a situation. The standard also relies on “other evidence” of an individual’s capacities and opportunity to exercise those capacities. This is what prevents Ciurria from having to say an individual who waits in a bank line cannot do otherwise simply because a high proportion of other people did the same. This is also why she does not have to commit to excusing “the majority, or even a significant number of people [who act badly in the normal world]” (p. 190). For example, the actions of Nazi officers “were beyond the pale”, as they were principle instigators of actions performed repeatedly, and they caused more harm than the average citizen in Nazi Germany.
Additionally, they were not culturally insulated or uneducated, and had the requisite moral understanding of their actions. According to Ciurria, these factors allow us to conclude that the officers had the capacity and opportunity to avoid wrongdoing. They could have reasonably been expected to see that what they were doing was wrong and chosen not to take part.

The fact that one can point to ‘other evidence’ of capacity or incapacity when using the reasonable person standard means the standard isn’t especially useful for our current task. Trying to work out what kinds of evidence suggest an incapacity is precisely why we began considering the reasonable person standard in the first place. The reason I have considered the reasonable person standard here is that I want to note a greater problem with it. This is that, without making certain contentious assumptions, we cannot identify the relevant reference class from which to apply the reasonable person standard. For example, if we applied a reasonable person standard to assessing the capacities of ‘all 6th graders’ on a reading task, we would observe that there are no constraints on behaviour and that most 6th graders could pass the task. We would thus be at risk of concluding those with dyslexia are merely being inattentive. It is only once we change our reference class from ‘all 6th graders’ to ‘all 6th graders with dyslexia’ that we can then apply the reasonable person standard and get the correct assessments. It is only after noticing a difference in capacity between these individuals and the other children that we can then see some expectations are unreasonable (e.g. reading a book at the same pace as everyone else) while others are very reasonable (e.g. listening to others to the same extent as everyone else).\textsuperscript{56}

The problem of knowing which reference class to appeal to is very relevant for assessing moral responsibility in the situationist experiments and many other cases. Ciurria does not comment on the responsibility of the lower level Nazi soldiers who were considerably more numerous than the higher officers, nor on how ‘situations’ within an environment can still vary substantially in such a way that affects a particular individual’s capacities. For example, though all the citizens of Nazi Germany were raised in a particular culture and setting, there would be tremendous variation between citizens regarding factors such as peer or family pressures, or propaganda prevalence and susceptibility. It is true that not every citizen joined

\textsuperscript{56} The point that the reasonable person standard fails to take into account the particular history or perspective of people from particular social groups has often been applied regarding cases of provocation or self-defence. See, for example, Donovan and Wildman (1980).
the Nazi party enthusiastically and so, at first glance, it seems a reasonable person within this German setting could have been expected not to join. As a result, those who joined seem accountable for their decision. But were it to turn out that the majority of those who had pressure from their peers or were from certain towns went on to join, and those who were not exposed to such factors did not join, this would indicate there are relevant differences between the ‘settings’ in which individual Germans found themselves. Specifically, it may turn out that while a reasonable person in Germany would not join, a-reasonable-person-in-Germany-who-was-raised-in-a-certain-environment-and-subject-to-strong-social-pressures would. I don’t want to make any particular responsibility assessments regarding Germans in WWII as this is outside the scope of this thesis, but the example is useful for showing how the reasonable person standard’s initial plausibility relies on holding certain factors fixed, and that individuals within a class of people may still have great differences in their capacities.

Consider one of the variations of Milgram’s experiment which had an obedience rate of 25%. When we apply the reasonable person standard to this case, it may be the case that a-reasonable-person-placed-in-this-setting would not administer shocks at 450V. But it could also turn out that a-reasonable-person-placed-in-this-setting-who-was-not-instilled-with-great-confidence-or-suspicion-of-authority-figures would in fact obey. It could turn out that once we remove everyone who was given certain moral training, the remaining reasonable people would overwhelmingly obey.

Some evidence suggests some subjects had a greater capacity to disobey than others, which has not yet been noticed in the philosophical literature. Adrian Dimow was one of Milgram’s subjects who successfully disobeyed very early in the most famous variant of the experiment, which had a 65% obedience rate. What makes Dimow’s behaviour of interest is that he disobeyed specifically because he thought the experimenters were lying to him. He reports he had been raised to distrust authority in a “socialist-oriented family steeped in a class struggle view of society [which] taught me that authorities would often have a different view of right and wrong than mine” (Dimow, 2004; Perry, 2013a, p. 127). He also tells of being harassed by the FBI while a member of the Communist Party. These experiences were clearly not had by the majority of the other subjects, and acted as a relevant difference which led to him finding it much easier to resist the influence of the authority figure.

This case is important when it comes to thinking about our capacity assessments. Knowing about Dimow’s history influences how we understand his behaviour in Obedience. Prior to having this information, we could easily just think that, since he was in the same
setting as the other subjects, he was therefore exposed to an equal amount of social pressure or ‘influence’ from the situationist factors present. We might then conclude he showed greater strength of will in resisting this influence than the obedient subjects, and that the obedient subjects could have used their strength of will to disobey too. But once we look at his particular history, it seems apt to say he found resisting the influence of the situation much easier than the obedient subjects.

Note that Dimow did not resist out of concern for the learner. This consideration barely even features in his account of his deliberative process. His interpretation of the setting, his perception of the experimenter, what considerations he took to be present and what he took to be at stake were all fundamentally different to those of most subjects. In short, Dimow was a subject who disobeyed and who was not relevantly similar to the obedient subjects. He seemed more ‘able’ to disobey in some important sense. It seems likely there would be variation in subjects’ interpretation of their setting, the pressure they felt, their perception of the experimenter and understanding of what was at stake. These are factors which can add up and affect the ease with which a subject can successfully disobey, and thus which group we should apply the reasonable person standard from.

A second means of assessing an agent’s capacities comes from Brink (2013). Brink argues that we should assess an agent’s capacities by examining how the agent would act in relevantly similar counterfactual situations. He argues that the obedient subjects are responsible agents, because he imagines there are relevantly similar counterfactual scenarios in which these subjects would successfully disobey:

Consider the Milgram experiment again. Would the compliant subjects have administered serious shocks if they had been given more time to consider their options, if they had been asked to justify the imposition of apparent harm to innocent parties, or if they had known the learners? If subjects would not have shocked in some of these circumstances, that provides evidence that those subjects had a capacity for recognizing reasons for compassion and for engaging in compassionate behavior.

Brink, 2013 (p.141).

Whereas Ciurria and Doris assess an agent’s capacities by looking at how relevantly similar agents would act in a particular setting, Brink looks at the particular agent across relevantly
similar settings. This enables Brink to avoid the problem regarding reference classes, since we do not have to categorise an agent into a particular class before making our assessment. The account can also explain why dyslexic people are thought to have a reduced capacity to read, because their reading would not improve in a variety of relevantly similar contexts. In contrast, agents in bank lines would leave similar lines with ease in a range of relevantly similar settings, which is why we don’t think they have a diminished capacity to leave lines.

A problem for this account regards which counterfactuals are relevant, and how we interpret the outcomes of those counterfactuals. Grant Brink’s supposition that an obedient subject would have disobeyed had they been given more time to consider their options. This could be taken to show that the subject had the capacity to disobey in the actual sequence all along. But it could also be taken to show that the actual sequence of events was relevantly different to a setting where that subject had more time. The counterfactual could instead show that something we assumed had no effect on an agent’s capacities, namely time pressure, was in fact significant.

One way around this is to imagine multiple counterfactuals or repeated iterations, holding the morally relevant factors fixed. Plausibly, if subjects took part in Obedience on multiple occasions, many more would eventually disobey. Brink would take this to show that subjects always had the capacity to disobey, but simply didn’t exercise it in the first iteration. But this could instead show that subjects’ capacity to disobey is reduced in novel situations, and that doing the right thing in unfamiliar situations without certain information (i.e. what has happened when one has acted in certain ways in this situation in the past) is more difficult than we typically assume. By analogy, the fact that an agent can quickly learn to reliably make successful putts can show they always possessed the capacity to make a putt on their first failed attempt, but it can also show that they gained a capacity they initially lacked.

There is one particular type of counterfactual that seems relevant to assessing the capacities of the obedient subjects. Since we’re assessing the capacity of subjects to resist strong-social-pressures-to-harm-others-in-novel-situations, Brink treats the relevant counterfactuals as ones where the source of social pressure, orders to harm others, and the overall task remain fixed. But if we’re truly interested in this capacity, it is relevant how subjects would behave in other, very different situations that still involve strong social pressures and causing harm to others. For example, we’d be interested how the obedient

---

57 Although they might instead become habituated to the experience and obey even more.
subjects would behave if they were placed in a Stanford Prison Experiment-like setting. We would also be interested to see what would happen in the kind of experiment designed by Hofling et al. (1966). If the obedient subjects in Milgram’s experiment would have done the same in a number of similar situations, these are exactly the kinds of counterfactuals which Brink would have to accept do show subjects that had a reduced capacity to disobey.

3.4 Capacities and Inferences

We are trying to work out a principled method of deciding when we should think an agent has or lacks a capacity to \( \phi \). To articulate the challenge more clearly, imagine we encountered an alien race, who found the following actions incredibly easy to perform:

- Staying awake for 100 hours
- Holding a heavy weight for a long time
- Keeping one’s hand in a bucket of ice
- Running a marathon
- Attending to the horizon for 12 hours for signs of an enemy approaching.

Suppose the aliens could reliably perform each of these actions through simply deciding to do so. Suppose also they saw that nearly all of us regularly fail when we attempt these things, and concluded most humans simply aren’t motivated enough to complete these tasks. How would we show that few of us possess the capacity to do these things in the sense required for moral responsibility?

Here are two prima facie sources of evidence we might point to. The first is that many people report they lack the capacity to do these things. The second is that many people reliably fail at these tasks, despite trying to succeed. The aliens, however, could correctly point out that people’s self-reports are often mistaken, and that people routinely fail due to a simple lack of motivation. Many people fail to stick to their diets, to save money, to go to the gym, and conclude that they lack capacity to do these things. But then one day they succeed,

---

It is sometimes reported in the philosophical literature that the person who blew the whistle on the My Lai massacre, Ronald Ridenhour, had taken part in a replication of Obedience and disobeyed very early on (e.g. Glover, 2003; Kupperman, 2001). This story is incorrect. Someone named Ronald Ridenhour did disobey early on in a replication, but this was a different person. See Bear (2013).
and they realise that their earlier self-reports were mistaken. They had the capacity all along, they just weren’t trying hard enough.

The aliens stand in relation to us in the same way that we stand to the subjects in the situationist experiments. We need some method of working out why some actions which people regularly fail at (e.g. staying awake for 100 hours) are thought to be beyond their capacities, but others (eating healthily) are not.

I believe that Ciurria, Doris and Brink focused too narrowly on finding evidence for the claim that the obedient subjects lacked the capacity to do otherwise, rather than thinking of this claim as a hypothesis which can be more or less plausible than other competing hypotheses. Here’s how I think we should make capacity ascriptions: for any failure to $\phi$ (understood as doing the right thing for the right reasons), we have two competing (but not mutually exclusive) hypotheses:

- The agent failed because they had the capacity to $\phi$, but didn’t exercise it.
- The agent failed because they lacked the capacity to $\phi$.

Our task (and that of the aliens) is to adjudicate between which of these two explanations is more plausible. When we think of our task this way, we can see why Ciurria, Doris and Brink’s accounts were appealing. Although the fact that large numbers of reasonable people act in an unusual way doesn’t prove they couldn’t do otherwise, it does require explanation. That a large number of seemingly reasonable people acted in an unusual way is much more difficult to explain as the product of individual idiosyncrasies than when only one person acts in an unusual manner. Likewise, how people behave under certain counterfactuals acts as evidence for what a person’s motivations are, which, in turn, makes certain explanations of their actual behaviour more likely than others. To investigate whether the obedient subjects had the capacity to do otherwise, I believe we should investigate what is the best explanation of their failure to do otherwise.

Here’s one way in which we can reasonably conclude an agent lacked the capacity to $\phi$ in circumstances C: show they failed to $\phi$ in C even though they were strongly motivated to $\phi$. For example, suppose we observe that I regularly succeed at completing the Stroop test without error in under 15 seconds when offered financial incentives, but fail to complete the test in under 10 seconds when offered equal or greater financial incentives. Given that we know I’m generally pretty motivated to complete quick tasks for money, and given that reducing the time of the test from 15 to 10 seconds doesn’t introduce any factors likely to
make me less motivated to try my hardest, the best explanation for my failure to complete the test seems to be that I lack the specific capacity to complete the Stroop test in under 10 seconds. And if, for some reason, I have a moral duty to perform the Stroop test in under 10 seconds, I will not be blameworthy for failing to do so.\footnote{This is roughly the reasoning used when making claims about subjects’ capacities in ego depletion experiments. In these experiments, subjects who have had to engage in some task requiring willpower (e.g. suppressing their laughter at a funny video, eating radishes instead of chocolate) appeared to perform much worse on later tasks requiring willpower (e.g. squeezing a spring-loaded hand grip, persistence at an unsolvable puzzle) than controls (Hagger, Wood, Stiff, & Chatzisarantis, 2010). These results regularly lead psychologists to talk about mental capacities, and make claims like “The capacity for self-control depends on a single stock of a resource that operates like an energy or strength” (Baumeister, 2002, p. 32). These researchers’ implicit reasoning is that, given the incentives to squeeze the grip or solve the puzzle are held fixed, and given the radishes or suppressed laughter are unlikely to give subjects less reason to change how long they persist at a task for, the best explanation is that subjects literally have a reduced capacity to complete that task, and so willpower is a capacity we can have more or less of. While ego-depletion experiments have been used by philosophers to help explain akrasia and addiction (Levy, 2006, 2011c) I’ve refrained from doing the same due to recent meta-analyses finding the effect sizes are suspiciously close to zero (Carter & McCullough, 2014; Hagger et al., 2016; Lurquin et al., 2016).}

There is a problem with making inferences about an agent’s capacities by comparing their performance across cases where their level of motivation is increased. This is that, in some cases, an increase in motivation can also produce an increase in capacity. To show what I mean, consider the following pair of cases:

*Sleepy*: Neil is told that the longer he stays awake, the less money will be taken from a charity. Neil stays awake for 50 hours.

*Fearful*: Neil is told that if he falls asleep, his children will be killed. Neil stays awake for 100 hours.

Suppose the aliens reasoned about this case in the following way: We know Neil cares about charities, and was motivated to stay awake for 50 hours. And this might make us think that he lacked the capacity to stay awake for longer. But he did stay awake even longer when he was more motivated to do so in *Fearful*. So Neil actually had the capacity to stay awake for 100 hours in *Sleepy* all along. He simply didn’t try hard enough.

This reasoning seems mistaken. Rather than thinking Neil just wasn’t motivated enough in *Sleepy*, most of us would conclude that Neil had a greater capacity to stay awake in *Fearful*. What makes Neil more able to stay awake in *Fearful* is his fear. His appraisal of the situation as dangerous to his children triggers certain responses in his brain and body, namely, greater
arousal and adrenaline which mitigate the drive for sleep. Importantly, these responses are not under Neil’s direct control. So it makes sense to say that Fearful itself is not proof that Neil had the capacity to stay awake (in the sense required for accountability responsibility) for longer than 50 hours in Sleepy, even though Fearful is somewhat similar to Sleepy and Neil in fact stayed awake for longer in Fearful.

This pair of cases is useful for a few reasons. First, it is a reminder of why it is important to hold certain facts fixed when assessing an agent’s capacity to ϕ; in this case, Neil’s level of physiological arousal, which he cannot directly control. Second, this case shows that it is very difficult to assess an agent’s capacity to ϕ by looking only at performance in one or two situations. As we saw when considering Brink’s argument, if an agent fails to ϕ, but would ϕ in other situations, do we say the agent had the capacity to ϕ and simply didn’t exercise it, or do we say that they had a reduced capacity to ϕ in that particular situation?60 Finally, this pair of cases helps illustrate how we tacitly rely on a host of background knowledge regarding people’s motivations and what sorts of things impede capacities when making capacity assessments. This is important to acknowledge, because, if our background beliefs are mistaken, we may overestimate how difficult it is for agents to perform certain actions in certain circumstances.61 Let’s see if we can make some of our implicit reasoning regarding capacities and motivation more explicit.

We’ve already seen that we can’t simply investigate whether subjects would have succeeded at doing the right thing if they’d had more incentive to do the right thing, or if they’d been given more time, or if they were placed in an identical setting again some time later. But tests like these can be useful in another way. By performing them in a wide variety of circumstances, we can make inferences about what sorts of factors do reduce people’s capacities, and then see if those factors are present in the setting we’re interested in.

This seems to be how we gain our background knowledge regarding things like action potentials, muscle fatigue, and adrenaline, which informed our earlier capacity assessments.62 One can readily find evidence that sleep deprivation impedes numerous capacities by

---

60 See Duhem (1954) for the *locus classicus* on the underdetermination of theory by data.

61 Our implicit background knowledge is what typically lets us succeed in selecting the relevant reference class when applying the reasonable person standard, or selecting the relevant counterfactuals for most capacity assessments.

62 Or, at least, scientists’ knowledge, which we accept as authoritative.
observing that people who are sleep-deprived perform much worse at a range of tasks. While it is possible that people who are sleep-deprived are simply much less motivated to perform those tasks, it seems implausible that sleep-deprived people are more likely to crash their car because their sleep-deprivation makes them simply care less about living. One can find similar evidence that adrenaline enhances physical capacities, by noting that people’s physical performance increases in a range of ways, even when we hold people’s incentive to perform fixed. This background knowledge then allows us to make inferences about more complex cases. If we know that someone was sleep-deprived when they failed to $\phi$, this can be a plausible explanation of their failure to $\phi$.

This is the kind of evidence we would rely upon to convince the aliens that people simply lack the capacity to stay awake indefinitely, though it is true most people can ward off sleeping for some time by trying to. The best evidence for the claim that a particular person can’t stay awake for 100 hours isn’t that they failed to stay awake, or that they report they lack this capacity, or even that most other people in similar situations fail to stay awake. It’s that they regularly fail to stay awake in situations where we would expect them to succeed given what we have reason to believe about their motivations. For any given instance in which an agent fails to stay awake for 100 hours, it is simply much more plausible to infer they lacked the capacity to stay awake for longer rather than to infer they possessed this capacity, but simply didn’t exercise it.

It may seem somewhat laborious explaining the structure of our reasoning in these cases, and the background knowledge we draw upon, given how easily we make these kinds of inferences. I’ve done so because, although it can be easy to assess whether an agent has a particular physical capacity, it is particularly difficult to tell when an agent has an impeded capacity to act for a certain reason. If we observe that an agent continually fails to act on a certain reason, do we say that they lack the capacity to act on this reason, or that they simply don’t care enough about that reason? If an agoraphobe is able to leave their house once it is set on fire, why shouldn’t we think he just wasn’t motivated enough to leave earlier? Doesn’t this show he did have the capacity to leave the house all along? (Fischer, 2012; Mele, 2006a).

In the remainder of this chapter, I want to examine two kinds of agents who have impeded capacities to do the right thing for the right reasons. This serves two functions. First, this will help us think about the kinds of evidence that lead us to think that an agent has a limited capacity to perceive and act on moral reasons in particular circumstances.
Second, these cases provide some of the foundations for my explanation of why subjects in the situationist setting have a reduced capacity to avoid wrongdoing.

The first group of agents are addicts. Though addiction and situationism are very different topics, philosophers have been able to show that addicts have an impeded capacity to recognise and act on moral reasons, explaining their tendency to relapse. Addicts particularly have a diminished capacity to attend to certain features of a situation and construe their situation in a certain way when they experience cravings. The second class of agents I want to consider are those with frontal lobe damage or dementia who commonly experience goal neglect. This is a phenomenon where an agent understands what is required in a task, is consciously aware of what to do, and yet seems to lack the ability to perform that task, even though they succeed with very minor prompting. By considering how goal neglect occurs, I aim to highlight how agents can lack the capacity to act on certain reasons in some very particular settings, although they can perceive those reasons.

3.5 Impaired Rational Capacities 1: The Capacity to Perceive Moral Reasons

Let’s distinguish between a few ways in which someone can notice, perceive, or be aware of something in their environment. Sometimes, we ‘see’ something but are not consciously aware that we are perceiving it. For example, you might hear the fridge or air conditioning turn off and realise that you had been hearing it make a humming sound for quite some time, but were not aware that you were hearing it. Alternatively, you might see something, but fail to notice a particular detail. You might think back to an interaction with a friend that occurred earlier in the day, and realise only now that your friend seemed sad.

Here’s a few helpful dimensions in which someone can perceive or be aware of something in their environment. The first is that our perceptual representations can differ in terms of their grain and richness (Siegel, 2005). Walking down the street with my friend, I might see my friend’s face, clothes, and bodily movements in great detail, but only see other people walking by as faint humanoid blobs. Perhaps I can make out some low-level details – e.g. tall, thin, black clothing – but not much else. Or perhaps I can see some of their features very clearly, but fail to notice others in much detail, including particular properties.
Our coarse-gained representations of the world are compatible with multiple states of affairs. These possible states of affairs can themselves appear more or less fine-grained to us when we consider them. Hearing my front door open, I might have a clear sense that one of my two housemates has come home, but cannot yet tell which one it is. In contrast, while talking to a friend, you might glimpse someone running down the other side of the street, but have little or no impression of whether they might be running to someone, away from someone, running happily or scared, or running as part of a race. You might notice only that someone-is-running, while your attention remains on your conversation with your friend.

To resolve ambiguous representations, or to get more information about which kinds of states of the world are more probable, we often direct our attention to certain features to try and generate more fine-grained representations. Perhaps it briefly seems to me like the person running may be trying to get away from someone. I can direct my attention to them to take in more details, quickly notice they are smiling. At this point it will seem to me that they are running for completely benign reasons, even if I’m still not sure whether they are in a race or merely training, and I can then return my attention back to my friend. But just as we can direct our attention towards certain objects or features to make our representations more fine-grained, sometimes features of our environment grab our attention without us having decided to attend to them. This attention-grabbing property is referred to as salience by psychologists (Taylor & Fiske, 1978). Features of our environment can have greater salience in virtue of a number of properties, including contrast (for example, a red circle surrounded by blue circles), novelty (a clown riding a unicycle down the street), importance (a gun) or familiarity (seeing a friend’s face in a crowd; overhearing your name at a cocktail party).

It is worth also briefly noting how much of our perceptual experience is a composite product of individual perceptual outputs. This is demonstrated by the fact that patients with brain damage can lack the capacity to complete very specific tasks. For example, some patients with brain damage can recognise faces but not emotions while others can recognise emotions but not faces, some cannot recognise tools from certain orientations, some can recognise what an object is but not what it is used for, and some can recognise moving objects but not still objects (De Renzi, 2000). Though the majority of us have no trouble exercising these capacities under normal conditions, differences between agents can still be observed in some settings. For example, some of us might have a greater specific capacity than others to recognise a particular tool that is far away, or that is obscured by foggy
conditions, or that is only glimpsed momentarily, or that is surrounded by lots of visually similar tools.

Finally, consider how many of our representations of the world often cause us to experience certain kinds of action-tendencies or dispositions (including directing our attention), which can themselves be stronger or weaker. A strange object coming fast at one’s head usually triggers a strong impulse to duck. A sudden appearance of something growling with fur and bared teeth typically triggers a disposition to get away. Seeing strangers walking on the path ahead triggers a weak inclination to move over so that you don’t bump into them, though this inclination typically increases as you get closer to them. Resisting these action-tendencies is sometimes very difficult, even if we know our perceptions of the world are misleading. For example, a holographic brick coming towards us at speed will cause many of us to duck, even if we know there are no real bricks in our vicinity. A prank mechanical bear might still trigger an impulse to get away if it unexpectedly moves. And actors often report that the fans they meet often erroneously treat them as if they really were the characters they play on screen (Barbas, 2016).

Relatedly, we can also perceive affordances, which are opportunities for action given to us by our environment (Gibson, 1979). For example, we usually perceive shoes as wearable, but not as chewable, even though they are something we can chew. To a dog, our shoes appear chewable, rather than wearable. The affordances we perceive can change depending on our goals. One might perceive a shoe to be throwable, if one is trying to get away from an attacker. But affordances can be perceived independently of our goals, and they can prime us to perform certain actions over others, making those opportunities for action more obvious and easier to consider in our deliberations (Pavese & Buxbaum, 2002).

Examples like these show that we can’t always treat questions like whether an agent noticed a consideration, or is aware of an action available to them, as having simple, discrete answers. Agents can be aware of features of their environment to different degrees and in different ways, and their level of awareness can change dramatically based on what the agent is attending to. Thinking about grain, salience, how our representations are composite images relying on many other capacities for detecting particular properties, and how many representations have associated action-tendencies, all help us understand how certain perceptual experiences can make certain actions more or less difficult to perform. It’s very difficult to do the right thing and save a drowning child if you don’t first notice that there’s
someone in the water, or if you don’t interpret them as being in trouble, or if your fear of water makes it seem like there’s a high chance you will drown if you jump in.

With this background, let’s now look at why (at least some) addicts are thought to be excused for their failures to stay clean. Many people naively think that addicts are excused because they experience a literally irresistible desire to use (thus lacking the capacity to resist), or because the withdrawal from not using would be too painful (thus imposing unreasonable costs on them). Although withdrawal certainly occurs for some users, it doesn’t seem to play a significant role in explaining why so many addicts relapse. Withdrawal for drugs like cocaine is relatively mild, subjective ratings of withdrawal severity have no correlation with objective measures, and many addicts of all drugs relapse several months after any withdrawal symptoms have abided (Loimer, Linzmayer, & Grünberger, 1991; Lyvers, 1998). Addicts also routinely refrain from using in many cases, such as when using has immediate and large penalties, when they have other commitments, or when they need to budget for other things (Neale, 2001; Watson, 1999).

Addicts thus seem to possess quite a lot of control over their actions. They use intentionally, they are often aware of what they are doing as they are doing it and, while their assessments of their reasons may not match our own, this seems able to be explained by simply saying they really enjoy using drugs. Many addicts also mature out of using, and seem to do so in response to incentives (Winick, 1962). This has led some philosophers to argue that addicts simply have different preferences to most people (Foddy & Savulescu, 2010).

Of course, many addicts report being unable to refrain from using. They also express very strong desires to refrain from using, which acts as some evidence that they lack this capacity. But we’ve already seen that there are cases where we don’t believe people’s self-reports, and people routinely fail to stick to their commitments. Very few people who make new year’s resolutions stick to them. So why are addicts often thought to have a diminished capacity to refrain from using, and to therefore be less blameworthy for any consequent wrongdoing?

Philosophers and psychologists have put forward a few recent arguments that addicts lack the capacity to refrain from using due to cognitive and volitional impairments. Though the details differ, they all rely on showing that addicts perceive their situation very differently to how non-addicts do. For example, Levy (2014a, 2016) argues that addicts use because they experience a judgement shift in what they should do, all-things-considered. Importantly, this judgement shift is not simply like an ordinary change of mind. Instead, because addictive drugs release dopamine, and dopamine is interpreted by the brain as a signal that the world is
much better than expected (or about to be), drug-related cues and thoughts cause the brain to predict that the world is about to contain a large amount of unexpected reward. But because this prediction doesn’t match the body’s bottom-up sense data (the addict isn’t actually experiencing any unexpected reward, because they haven’t taken any drug) there is a very large prediction error, which our brains naturally try to resolve. The way addicts’ brains resolve this error is by adopting a different model of the world – one in which the drug is seen as very positive, and worth taking, all-things considered.63

A similar theory comes from Berridge (2007) who argues that addicts’ dopaminergic firing attributes greater ‘incentive salience’ to drug-related stimuli, causing the addict to experience a conditioned pathological motivation to obtain them (see also Robinson, Robinson, & Berridge, 2013; Robinson & Berridge, 2008). Evidence that this motivation is pathological comes from studies showing that the objects which acquire incentive salience can completely dissociate from the original objects of reward. Animals conditioned to expect that a particular stimulus predicts a certain reward will sometimes start treating the stimulus as if it were the reward, even in ways that actively prevent them from getting the reward itself. For example, hungry pigeons that have learned a light at one end of a box predicts food at the other end for a short time can be drawn to the light, and will try to ‘eat’ the light, even though doing so prevents them from getting the clearly visible food at the other end of the box. This doesn’t seem to simply be a classically conditioned response, because animals will engage in novel actions requiring instrumental reasoning to get to the salient stimulus.

Henden (2018) argues that addicts’ primary problem is that they suffer from depleted self-control. Drug addicts notably make many more errors on Stroop tests with drug-related words than non-addicts, indicating that their attention is being pulled towards drug-related stimuli (Field & Cox, 2008). This matters, because our ability to control our attention is widely thought to be an important component of our ability to exhibit self-control (Eisenberg, Smith, Sadovsky, & Spinrad, 2004). By attending to particular features of our situation, we make certain other features or reasons more obvious, and ignore others for deliberation. If addicts experience ongoing cravings, their attention is continually being pulled towards drugs and the benefits of using. This makes the benefits of using seem much more obvious, and the costs much more difficult to imagine accurately. If self-control is a

---

63 A more complete explanation relies on the understanding the predictive processing account of perceptual experience (Clark, 2016; Hohwy, 2013).
limited resource, as many psychologists argue (Baumeister, Vohs, & Tice, 2007), then it is plausible that addicts give in to using because they lack the strength to inhibit these urges indefinitely, and cannot keep the considerations against using at the forefront of their mind.

Schroeder and Arpaly (2013) argue that drug-taking has very strong effects on unconscious learning, making the presence of particular mental states (e.g. seeing a drug on television) very likely to cause others (e.g. remembering all the fun times one has had when using). In particular, drug-related cues are thought to trigger dispositions to act from habit, which must be inhibited for as long as cravings persist. Although this may make it sound like addicts can stay clean by just resisting bad habits, Schroeder and Arpaly note that it can be very difficult for ordinary agents to inhibit habits, even when consciously trying to. Additionally, the habits we develop can include habits of thought (e.g. arguments in favour of using) which are very difficult to notice as being against one’s best interests.

The exact details of why addicts find it so hard to get clean is beyond the scope of this thesis. What I want to highlight is that when we add up all the ways in which brains are affected by using, addicts seem to have a reduced capacity to recognise and act on moral reasons when they experience cravings. Their attention is constantly being pulled towards thoughts in favour of using, positive memories of using, arguments in favour of using, and away from the considerations against using. The considerations against using become much less fine-grained and seem less important. Addicts may be aware that using is in some sense ‘bad’, but downplay the costs of using. They fail to consider consequences like losing their job in enough detail, or to accurately imagine how unpleasant those consequences will. And even if they often manage to take steps that prevent themselves from using, unlike most habits or inclinations that ordinary agents have to inhibit on some occasions, “abstinence requires not using the drug on every occasion” (Schroeder & Arpaly, 2013, p. 229).

Without practice managing cravings, prepared plans for what to do when cravings begin, or support from people who know what is happening, addicts can just find themselves

---

64 Though as noted earlier, the evidence usually appealed to for this argument is facing challenges due to failed replications and evidence of publication bias. See supra 59.

65 As an example, they cite a report from someone being trained to throw grenades in the army, where an instructor would hit cadets with a stick if the cadets watched where the grenade landed instead of ducking. Despite resolving to duck immediately after throwing the grenade, many cadets throw, watch, and get hit with the stick (p. 227).
thinking and attending to considerations favouring using, and not realise that their perceptions are being distorted. All of this means that not using is incredibly difficult.

Understanding how our brains represent the world is also useful for explaining why agents with post-traumatic stress disorder (PTSD) and extreme phobias lack control over some of their actions. Agents with PTSD experience distress, disturbing and intrusive thoughts or feelings related to earlier trauma, heightened arousal, and hyper-vigilance, amongst other things (American Psychiatric Association, 2013). These agents’ earlier trauma is thought to cause their brain to weigh certain predictions or hypotheses – namely, threatening ones – as much more likely than they in fact are (Wilkinson, Dodgson, & Meares, 2017). For example, while most of us will automatically interpret the sound of a squeaking window as the product of wind blowing the window open, someone with PTSD may interpret the sound as coming from a burglar sneaking into the house, something which would be very distressing. Similarly, agents with extreme phobias perceive certain stimuli to be very threatening, and this also seems to be explained by the predictions their brain generates. Although they may consciously believe that what they are afraid of isn’t dangerous, their brain has adopted a model in which that thing is dangerous and should be avoided.66

I have made this digression into the addiction literature to highlight two things. The first is that philosophers and psychologists agree that, to understand the blameworthiness of some marginal agents and why their capacities to perform certain actions are impeded, we need to first understand how they perceive features of their situation. This is relevant because I believe that to explain the behaviour of subjects in the situationist experiments, we also need to examine how they interpret their situation. Once we do this, we will see they also had a reduced capacity to avoid wrongdoing.

The second thing I want to highlight is how difficult it can be to talk accurately about the capacity to perceive moral reasons, especially if we do not reference things like attention, representations, or the salience of certain considerations. Giving a simple yes or no answer to

66 Illusions are a common example where our brains adopt a model we don’t consciously believe. Although we believe that the Muller-Lyer lines are the same length, our brain has encountered similar lines when looking at the edges of rooms so often that it assigns a very high probability to the hypothesis that the arrows are reliable cues for assessing distance, causing us to perceive one line as longer than the other (Vetter & Newen, 2014). Just as we experience the lines as looking to be different lengths, agents with PTSD or extreme phobias experience certain stimuli as dangerous, and it is very difficult to act as if the stimuli are not dangerous even if one consciously believes the stimuli are safe.
whether addicts are aware of the moral reasons against using is likely to be misleading. In some sense, addicts are aware of these reasons when experiencing cravings, in that they can perceive moral reasons in a certain pattern that leads us to ascribe a certain understanding of their moral situation. But in another sense, they are not aware of these reasons, because they cannot accurately perceive the weight of many reasons. What exactly it is to perceive moral reasons, to have moral understanding, or to take moral considerations to be stronger or weaker, is difficult to analyse.⁶⁷

3.6 Impaired Rational Capacities 2: The Capacity to Act on Moral Reasons

Having outlined some of the difficulties addicts face in refraining from using, let me outline another kind of difficulty faced by some agents. Whereas the previous section looked at difficulties in perceiving certain considerations, this section looks at a difficulty in acting. Psychologists have identified a very interesting phenomenon known as goal neglect. Goal neglect occurs when subjects know they are required to perform a certain task, can accurately report the task’s requirements while trying to perform that task, have previously succeeded in performing the task and yet, in some cases, appear unable to complete that task (De Jong, Berendsen, & Cools, 1999; Duncan et al., 2008; Kane & Engle, 2003)

Psychologists demonstrate goal neglect with experiments that usually involve instructing subjects to follow two or more rules (e.g. when you see an arrow, only read out the letters on the side of the screen the arrow points to). Subjects will fail to follow certain cues (e.g. they fail to change responses when the direction of the arrow changes) but are perfectly capable of reporting what the rules of the task are, and that they have just seen a particular cue entailing a particular rule. Some will even say things like ‘I should ϕ’ (e.g. ‘I should raise my arm’)

⁶⁷ As already noted, some philosophers have tried to overcome these problems by avoiding referring to subjects’ subjective construal of their situation entirely. They instead opt to analyse facts about how the agent acts in nearby possible worlds. For example, Coates and Swenson (2013) argue that Fischer and Ravizza’s (1998) account of reasons-responsiveness can account for how difficulty affects praiseworthiness and blameworthiness. One is more blameworthy to the extent that she is able to respond to reasons, where this is measured by the closeness of the nearest possible world in which their mechanism reacts to sufficient reason to do otherwise. Nelkin (2016b) objects that difficulty often doesn’t depend on features of mechanisms, but features of one’s environment i.e. what is at stake, rather than how easily my mechanism is able to resist giving that thing up. Additionally, some possible worlds might be further away but not difficult to access e.g. the agent in question might just be lazy. See also supra fn. 24.
upon seeing the relevant cue, but then make no attempt to $\phi$. Curiously, when subjects receive an external prompt (e.g. the experimenter says ‘you should raise your arm’, or a word appears indicating the appropriate action) subjects are almost immediately able to complete the task again for a number of trials.

Goal neglect is most commonly observed in people with frontal lobe damage or dementia. The factors causing these agents to experience goal neglect seems to explain many other difficulties they experience, such as stimulus categorisation, maze learning, picture description, puzzle solving, story recall, and regard for social conventions. These behaviours had previously puzzled psychologists, because people with frontal lobe damage and dementia can score normal or even high scores on certain IQ tests (Eslinger & Damasio, 1985; Shallice & Burgess, 1991), though not tests of fluid intelligence (Duncan, Burgess, & Emslie, 1995).

Goal neglect is thought to occur due to a particular failure in the cognitive processes underlying goal-directed behaviour.\(^{68}\) Completing any particular goal requires first completing subgoals, which themselves have further subgoals. For example, to make soup, one must chop up vegetables, but to chop up vegetables, one must acquire a knife and vegetables, and to acquire a knife, one must reach for the knife and grab it, and so on. This schema of tasks unfolds until one reaches a basic motor or mental action. Importantly, each goal must be adopted in a particular order to be successfully executed. There is no common association between ‘cooking soup’ and ‘flexing fingers’; one needs to have adopted the intermediary goal of finding a knife and reaching for it before flexing one’s fingers can be efficacious. If one tries to proceed directly from the most abstract goals to the most fine-grained goals, one’s planning would be chaotic. Agents must thus form a ‘model’ of the tasks and the sequence in which they should be completed. For each subtask, behaviour must be appropriately guided by selectively retrieving relevant knowledge from memory, and combining this with current sensory input. If one cannot attend to each particular subtask, and exclude irrelevant aspects, one will not be able to complete the main task.

\(^{68}\) While it is thought to occur due to our limitations in working memory, this phenomenon differs to the more commonly known difficulties that most humans have with tasks requiring working memory e.g. attending to multiple stimuli simultaneously or trying to remember more than nine digits for recall. The difficulty with dual-attention tasks can be resolved if stimuli are separated by more than a few 100ms, and the difficulty with remembering multiple digits can be resolved if one set of digits can be reported before another set is stored. In contrast, in goal neglect tasks, each individual task is typically easily executed, subjects don’t have to attend to two tasks at a time, and they can, in principle, abandon each task before moving on to the next one. The difficulty is specifically with the organisation of, and changes between, easily performed tasks.
Goal neglect is thought to occur when an agent has trouble abandoning a previous goal-directed action that has become inappropriate, leading agents to display perseveration (e.g. repeating a motion or word that was previously appropriate). Changing which subgoal we are engaged in requires being able to update one’s model with information about the world. For example, if you’ve completed the goal of cutting onions, you abandon the goal of cutting onions and ‘select’ a new goal of cutting carrots.

Importantly, people with frontal lobe damage and dementia seem to have little trouble performing tasks that are familiar and well-learned. Such patients also seem to have little trouble performing tasks when given verbal prompts regarding particular subgoals. For example, a patient “may fail to make progress on some task such as preparing a meal, but be perfectly capable of carrying out each component when individually instructed to do so” (Duncan, Emslie, Williams, Johnson, & Freer, 1996, p. 264). Familiarity and prompting seem to resolve goal neglect because they act as sources of bias towards certain subgoals for selection. If a patient is well-practised at making soup, it will be easier for them to automatically select the correct subgoals in the correct order and respond accordingly.

Patients mostly have difficulty with tasks that are novel or complex, i.e. that require spontaneous generation of an unfamiliar task model. This is because “As more information is entered into the task model … multiple task components compete for representation and individual components may be weakly represented and lost, leading to neglect of that requirement during performance” (Bhandari & Duncan, 2014, p. 28). These patients seem to experience task neglect because they have a reduced capacity to ‘select’ the appropriate task at the appropriate time, even if they are consciously aware of what the appropriate response is. For example, “Asked to draw a square, a patient described by Konow and Pribram (1970) began instead to draw an A (repeated from a previous trial). Her comment was, ‘That’s not a square – I guess I’ll draw you an A.’” (Duncan et al., 1996, p. 265). Although the patient knew she should draw a square, she had been unable to abandon the previous task sequence.

Here’s why understanding this phenomenon matters for us: goal neglect occurs in ordinary agents with ordinary cognitive capacities too (Altamirano, Miyake, & Whitmer,
And just like with frontal lobe patients, this neglect is easily resolved with verbal or visual prompts. When agents are carrying out unfamiliar tasks that require attending to multiple unrelated factors, they can get ‘stuck’ when trying to change actions, despite being able to recognise and report what action they should perform at that very time. Crucially, it doesn’t seem to be the case that they form the intention to perform a certain bodily movement, but this intention then somehow fails to effectively result in the relevant bodily movement. Instead, while having the broader overall intention to complete the task, they are blocked from forming the narrower intention to perform the relevant bodily movement.

Without an understanding of how exactly goal neglect occurs, and how it is much more prevalent in agents with cognitive impairments, ordinary agents who experience goal neglect may be interpreted as just not trying hard enough. After all, if an agent has seen the relevant considerations in favour of \( \phi \)-ing, understands that they are supposed to \( \phi \), can report they should \( \phi \), and has \( \phi \)-ed reliably in the past, it will be very tempting to conclude that they must have the specific capacity to \( \phi \). But when we reflect on how inexplicable their failure to \( \phi \) is given what we know about their motivations, and how often they previously succeeded, the best explanation of their temporary failure to \( \phi \) is that they experienced a very temporary incapacity to \( \phi \). Although much of the literature on capacities takes repeated observations of successful \( \phi \)-ing to be evidence of the capacity to \( \phi \) despite a particular failure, goal neglect is a case in which repeated observations of successful \( \phi \)-ing act as evidence of incapacity.

I want to make a similar argument with regard to some of the subjects in the situationist experiments. By thinking about the processes that subjects must engage in in order to do the right thing, and showing that subjects are in the presence of factors which tend to impede our capacity to engage in those processes, I want to argue that we have grounds for thinking many of the subjects who commit wrongdoing are significantly excused for their actions.

In section §3.5 and §3.6, I’ve highlighted the difficulty of talking accurately about the capacity to recognise and act on moral reasons. I did this because when we evaluate whether an agent had these capacities, we often just try to imagine how things would have looked to

---

69 Though goal neglect occurs less frequently in subjects with high levels of Spearman’s \( g \) in these experiments, and more frequently in children (Piek et al., 2004; Towse, Lewis, & Knowles, 2007), this doesn’t mean adult subjects with high \( g \) don’t experience goal neglect, only that more complex tasks are required to induce it. I argue below that subjects in Samaritan plausibly experience goal neglect.
them and ask how easy it would have been to do the right thing. But given the situationist subjects are in highly novel situations that we haven’t experienced, there is good reason to suspect simply imagining what being in that situation would look like will lead us to overestimate the subjects’ capacity to avoid wrongdoing. Psychologists have shown that humans regularly overestimate how easy it is for others to know what we know, calling this the ‘curse of knowledge’. For example, Fischhoff (1975) found that people who have been given the description and outcome of an event overestimate how other people will rate the likelihood of those outcomes occurring based off the description alone. Newton (1990) found that people tapping a song out radically overestimate how easy it is for other people to correctly guess that song. And Keysar and Henly (2002) found that speakers who read out ambiguous sentences (e.g. Angela killed the man with the gun) strongly overestimated how the ambiguity would be resolved by other subjects, when the speakers had earlier privately read a scenario resolving the ambiguity. The common theme in these studies is that people strongly overestimate how agents with limited knowledge will interpret somewhat ambiguous situations. And when we sit at home, reading about these experiments, it seems obvious to us what the correct course of action is, because we know which considerations morally ought to guide action and which are irrelevant.

Having given some examples of agents who ought to be excused for their actions because they fail to perceive the moral features of their situation accurately, or have difficulty forming and acting upon an intention to do the right thing, I will now turn to explaining the behaviour of subjects in the situationist experiments. I will argue that most subjects who fail to do to the right thing perceive their situation differently to how we expect, or have difficulty forming the intention necessary to do the right thing.

3.7 Conclusion

In this chapter, I have examined how we assess whether an agent had the capacity to do the right thing for the right reasons. I began by considering whether the fact that some subjects

---

70 To be clear, I will not be arguing that subjects in the situationist experiments behave like addicts, or that they interpret their situation like addicts do. Unlike addicts, the subjects in the situationist experiments would not experience cravings to commit wrongdoing, and would be unlikely to repeatedly fail if they were repeatedly placed in these settings with forewarning. But I do think that when we ask whether subjects in the situationist experiments can resist whatever inclines them towards wrongdoing, we need to keep in mind the various cognitive processes that add up to give an agent the capacity to recognise and act on moral reasons.
successfully disobeyed counts as evidence that the obedient subjects had the capacity to disobey too, since these groups of agents seem to be otherwise relevantly similar. I argued that such reasoning would lead to erroneous conclusions when considering conditions like dyslexia or depression. I then considered two attempts to distinguish agents who possess a particular capacity from agents who lack that capacity, and outlined objections to these accounts. Appeals to the reasonable person standard are unable to specify the relevant reference class from which to apply the standard. Appeals to counterfactuals fail to specify which counterfactuals are relevant, and the results of these counterfactuals are often consistent with multiple hypotheses. I then argued that the strongest way we can make capacity assessments is through inference to the best explanation. By looking at agent behaviour in a range of circumstances, and combining this with background knowledge about what motivates agents, we can assess whether any particular failure to do the right thing is the product of a lack of capacity, or of insufficient motivation.

I then considered two kinds of agents who find it difficult to act for certain reasons. I examined why addicts relapse so often, and argued that a crucial component in our explanation was that cravings cause them to construe their situation in distorting ways, interfering with their capacity to recognise moral reasons. I then considered the phenomenon of goal neglect, to show how agents can sometimes lack the capacity to act on moral reasons even though they have recently reliably performed such actions in relevantly similar settings.
Chapter 4: Situationism and Selecting Action

4.1 Introduction

In the previous chapter, I considered some common ways we make inferences about an agent’s capacities, and problems with these methods. I argued we should assess whether an agent had a certain capacity to $\phi$ by examining whether their failure to $\phi$ is best explained by their having the capacity to $\phi$, and simply not being sufficiently motivated to exercise it, or by their lacking that capacity. In this chapter, I will finally assess what is the best explanation of subject behaviour in *Dime*, *Bystander*, *Samaritan* and *Obedience*.

This chapter will proceed as follows. In §4.2, I argue that, although situationist factors like those in *Dime* can somewhat reduce subjects’ awareness of moral considerations, they do not do so to a degree large enough to excuse the subjects who commit wrongdoing. I then argue that, given the results from a number of related studies, the best explanation for why some subjects in *Bystander* fail to give aid is that they genuinely do not take the victim to be in need of help. Similarly, many subjects in *Samaritan* failed to interpret the confederate as in need of assistance. Some subjects did notice, however, and they may be partially culpable.

In §4.3, I argue that subjects in *Obedience* have a greatly diminished capacity to avoid wrongdoing. The ambiguity of their situation, the difficulty of interpreting the actions of the experimenter, their reasonable belief that the subject would get the next answer right or that the experimenter would notice the moral considerations present, and the focus of their attention, all made it very difficult to perceive their situation in the way that we expect they would. This prevented subjects from recognising that disobeying was an option available to them, which in turn prevented them from disobeying. In contrast, many of the subjects who disobeyed seemed able to do so precisely because they construed their situation differently. I then consider and respond to a number of objections in §4.4.

4.2 Explaining Behaviour in *Dime*, *Bystander* and *Samaritan*

Let’s start with explaining why small environmental factors can have large effects on moral behaviour, such in *Dime*. The most commonly offered explanation for why items like dimes can affect subject behaviour is that finding a dime is experienced by subjects as an unexpected piece of good luck, putting them into a better mood (Isen & Levin, 1972).
Although moods are thought by some psychologists to make people more motivated to help, the effect of moods on attention is thought to be particularly significant (Carlson, Charlin, & Miller, 1988). It was initially thought that positive moods broaden our attention (Conway, Tugade, Catalino, & Fredrickson, 2013; Fredrickson & Branigan, 2005; Wadlinger & Isaacowitz, 2006), but more recent research suggests this is mistaken. Instead, many emotions generate an urge to act in some way (e.g. fear generates an impulse to avoid or get away from whatever we perceive to be dangerous), and a large body of evidence shows that the intensity of such urges narrow our attention (see Harmon-Jones, Price, & Gable, 2012 for a summary). Intense emotions that generate urges to act in some way narrow subjects’ attention, causing them to notice things in their peripheral vision less, pay less attention to any peripheral items they do notice, focus more on parts and details rather than wholes, and be more concrete/less abstract in their conceptual categorisations. Less intense emotions and moods, such as the happiness one might feel after finding a dime, have the opposite effect. Such results plausibly explain why subjects in Mathews and Canon (1975) were less likely to help strangers when in the presence of a loud lawn mower. Loud noises are unpleasant, and generate an impulse in us to avoid their source, which can, in turn, cause us to focus more on our current task, making other features of our situation much less salient.

Now that we have an explanation for the relative variation in subjects’ behaviour between conditions, what should we think regarding subjects’ level of responsibility? I believe that the subjects in Dime remain blameworthy for their actions. The confederate dropping his papers may be less noticeable to subjects than he otherwise would be, and subjects may be feeling more of an inclination to continue walking than they would experience had they found the dime. These factors might make subjects who fail to help slightly less blameworthy than had they found the dime, experienced the broadened attention and reduced inclination to keep walking, but kept walking anyway. But this difference in relative blameworthiness is minor. Dimes and noises are just part of the ordinary furniture of moral life which we expect responsible agents to be familiar with. If somewhat annoying noises were sufficient to meaningfully reduce people’s capacity to perceive or act on moral reasons, it would be a wonder how pre-school teachers could ever do the right thing. We have no evidence that the subjects are not sufficiently aware of the events around them, or the moral nature of their actions. Doris (2002, pp. 31, fn. 10) reports from personal conversation with Isen and Levin that some subjects in Dime actually trampled on the papers dropped by the confederate. Given the wrongs in these experiments are all relatively minor, subjects likely just don’t care
as much as they ought to. Any remaining intuitions that these subjects should be partially
excused may be responding to the fact that situationist factors can act as a form of moral
luck, as Herdova and Kearns (2015) have argued. While moral luck raises important
questions, they are outside the scope of this thesis. More importantly, these worries about
moral luck do not show that the subjects are undeserving of blame.

Let’s now consider the situationist experiments involving the presence of others, starting
with Bystander. Strong evidence that the subjects in these settings have a different perception
of their environment than we expect comes from passages like this:

Subjects who had not reported the smoke also were unsure about exactly what it was,
but they uniformly said that they had rejected the idea that it was a fire. Instead, they
hit upon an astonishing variety of alternative explanations, all sharing the common
characteristic of interpreting the smoke as a nondangerous event … subjects claimed,
not that they were unworried by the fire or that they were unwilling to endure the
danger; but rather that they had decided that there was no fire at all and the smoke
was caused by something else. They failed to act because they thought there was no
reason to act. Their “apathetic” behavior was reasonable given their interpretation of
the circumstances.


Interpretations of what the smoke was included smog, air conditioning vapours, and
‘truth gas’.

Although it is possible that some subjects lied out of embarrassment, Latané and
Darley (1969) and Herdova (2016) argue it is unlikely all these subjects would think to use
the same kind of lie.

Further evidence that subjects really did not interpret their situation as we do comes from
studies on what factors exacerbate or mitigate the bystander effect. Studies consistently show
that subjects are much more likely to help when the situation is unambiguously dangerous, or
when the victim explicitly labels themselves as in need of help (see P. Fischer et al., 2011 for
a meta-analysis). Subjects are also much more likely to help when other bystanders visibly
react in surprise to the event (Darley & Batson, 1973), are blind (Ross & Braband, 1973), are
children (Ross, 1971), or are visibly unable to help (Bickman, 1971). Shotland and Stebbins

---

71 The smoke used was from burning titanium dioxide. This has a very acrid smell, unlike dry ice
which is more commonly used to mimic smoke, which some readers may assume was used.
(1980) found that bystanders who saw and heard an assault were much more likely to intervene than those who only heard the assault. Cramer, McMaster, Bartell, and Dragna (1988) found that nurses, who notably have experience in both giving assistance and recognising when people need medical attention, did not fall prey to the bystander effect. Beaman, Barnes, Klentz, and McQuirk (1978) found that people who had been educated on the bystander effect are much less susceptible to it.

These studies cast doubt on the hypothesis that people experiencing the bystander effect see the moral considerations but simply fail to care sufficiently about them, possibly due to a fear of embarrassment of desire to conform. Subjects’ behaviour consistently tracks the degree to which the victim’s need for help is recognisable, rather than any prudential factors we could plausibly attribute as relevant to subjects’ motivations. Indeed, we would expect subjects to be less likely to help in unambiguously dangerous situations if they were being motivated by objectionable attitudes, since intervening in dangerous situations carries considerable risk. Learning about the bystander effect is also unlikely to change one’s concern for others in any substantial way, or reduce the risk of embarrassment, and yet it has a significant effect on people’s likelihood of helping. The best explanation of why training reduces the bystander effect is that it enables subjects to recognise that the victim needs help. This, in turn, enables them to ignore other people’s failure to help when interpreting their situation, and trigger the thought that helping is an action they are capable of performing.

Let’s now turn to Samaritan. I believe that the subjects who failed to help in this experiment are a heterogeneous group, in that different subjects failed to stop for different reasons. Some subjects seemed to not notice the confederate while walking by him:

> When we talked to them later, some of them seemed to be thinking about the victim for the first time, as if saying ‘Gee, maybe the guy was sick and needed help’. They really hadn’t noticed him.

Darley, quoted in Shearer (1971, p. 11).

These subjects plausibly suffered from some form of inattentional blindness. This is a phenomenon where attending to particular features of a situation causes agents to miss other features of their situation that normally seem obvious. An example of how strong this blindness can be in situations like Samaritan comes from Chabris, Weinberger, Fontaine, and Simons (2011). These researchers asked subjects to follow a confederate around campus for
three minutes and count how many times he touched his head. While doing this, subjects passed a staged fight only eight metres away, in which two people appeared to be loudly beating up a third person. Though this fight was in subjects’ line of sight for 30 seconds, 58% of subjects did not notice the fight at all.

The subjects Darley is referring to may have ‘seen’ the confederate, in that, upon reflecting on their memory of the walk, they realised that there had been a person present. But it seems plausible that at the time of walking, they did not notice that a person was lying on the ground. This phenomenon of ‘looked but failed to see’ is thought to explain why drivers are at a much greater risk of crashing when using their phone despite keeping their eyes on the road, and why some motorists hit people on bicycles or motorcycles after having looked to see if it was safe to merge (Hyman Jr, 2016). This also plausibly accounts for why so many more subjects helped when told they were not late. Since they didn’t have to concentrate on getting to the room quickly and avoiding mistakes, their attention wandered more freely, enabling them to notice something out of place on their walk. They then attended to this unexpected shape, and saw that it was a person lying on the ground.

Some subjects may have noticed the confederate in a slightly stronger sense, but failed to ‘see’ him in the relevant way. They may have seen that a person was there, but failed to see that he needed help. Consider the following quote:

Our seminarians in a hurry noticed the victim in that in the post-experiment interview almost all mentioned him as, on reflection, possibly in need of help. But it seems that they often had not worked this out when they were near the victim. Either the interpretation of their visual picture as a person in distress or the empathic reactions usually associated with that interpretation had been deferred because they were hurrying. According to the reflections of some of the subjects, it would be inaccurate to say that they realized the victim’s possible distress, then chose to ignore it; instead, because of the time pressures, they did not perceive the scene in the alley as an occasion for an ethical decision.


Based on earlier cited studies regarding the bystander effect, it seems plausible that some subjects genuinely didn’t interpret the confederate as being in need of any help. They likely had a low-to-medium grain representation of a person lying down, but failed to see some
finer-grained properties, like that he was unconscious, appeared to have collapsed rather than lain down, and was likely in need of help. If subjects simply didn’t have the thought that ‘that person probably needs help’, then it is hard to see how they could have had the thought, ‘I should go help him’, which is required for them to in fact give aid. If this characterisation is accurate, then these subjects will be excused for their failure to help, as will the subjects who did not notice the confederate at all.

It doesn’t seem like all the subjects who failed to give help were like this though. Some subjects seemed to have noticed the confederate and made a decision to continue on:

For other subjects it seems more accurate to conclude that they decided not to stop. They appeared aroused and anxious after the encounter in the alley … Because the experimenter, whom the subject was helping, was depending on him to get to a particular place quickly … Conflict, rather than callousness, can explain their failure to stop.


These subjects’ culpability is difficult to assess. On the one hand, it seems clear they saw the confederate as someone needing help, given that they were anxious. On the other hand, it seems plausible that the factors which contributed to some subjects not seeing the confederate as needing help, or not seeing him at all, would still have some effect on these anxious subjects’ construal of their situation too.

Recall from §3.5 that our awareness of objects in our environment comes in degrees. Subjects might have had a coarse-grained thought along the lines of ‘he is not ok’ but failed to perceive the confederate as being in need of help, even though those two thoughts seem very closely linked to us. Recall also that goal neglect is particularly likely to occur for tasks that are not well-trained, in settings where there are no prompts or cues towards performing the relevant action, and when our working memory is engaged in an unrelated task.72

Even if subjects didn’t experience complete goal neglect, the kinds of factors that result in goal neglect plausibly make changing tasks more difficult than we expect. The subjects in *Samaritan* do not normally see unconscious people in the middle of the day, and they are

---

72 “The [process of shaping behaviour between competing goals] is most likely to fail under conditions of novelty and weak environmental support.” Duncan et al. (1996, p. 265).
already engaged in an unfamiliar task which requires sustained attention (preparing to give a talk on the parable of The Good Samaritan, trying to locate a new room quickly, anticipating having to apologise to for being late etc.). These tasks would occupy much of their working memory, directing their attention away from other features of their environment. Their representation of the confederate as someone in need, and the corresponding thought that helping him is an action they can perform, might be weaker than we expect. Subjects may fail to have the thought that they are in a position to help. The unfamiliarity with stopping to check on an unconscious stranger, and the uncertainty with what to do after checking on him (e.g. what if he’s not responsive? How do I get help? What if he’s fine and checking on him embarrasses him?) would increase the difficulty of considering this course of action, while the familiarity with the much simpler task of walking would create a bias towards continuing walking. After passing the confederate, subjects would have more time to consider what they saw, which would in turn make them much more aware of the actions they could have taken.

This explanation would match Darley and Batson’s ambiguous characterisation of subjects as having both noticed and not noticed the confederate. If this explains why so many subjects in the late condition failed to help, this would be partially mitigating. Even if subjects technically retained the specific capacity in the relevant sense to give help, a number of factors plausibly made it more difficult for them to give help than we would expect.

One particular consideration counts against excusing the subjects entirely. The fact that some subjects were anxious upon arrival suggests that they were considering whether to go back and check on the confederate. This would imply that they had been deliberating about whether to help during their walk to the room. The longer they deliberated, the less likely it is that they were genuinely unable to switch tasks. Subjects would be aware that they could go back to check on him, and could plausibly have the sort of thoughts necessary to trigger the relevant task model which would overcome any inclinations against helping. These subjects would merely be suffering from akrasia, which isn’t in itself an excusing condition.

I thus believe that the subjects who fail to give aid in Samaritan are too heterogeneous for any single culpability assessment to be accurate. Some subjects will have failed to notice the confederate entirely, or failed to interpret him as being in need of aid, and they will be fully excused. Some subjects will have noticed him and initially found it very difficult to realise that helping is something that they could do, but later realised this and been capable of switching tasks as they deliberated further. Such subjects are plausibly somewhat blameworthy.
Let me briefly anticipate and ward off one potential objection. Arguing that helping is difficult because it is unfamiliar or requires changing tasks does not risk excuses nearly all wrongful omissions that require an agent to change from some unrelated task. I have argued that these factors are plausible partial explanations of subjects’ failure to help because the confederate is genuinely unexpected, highly out of place, and there are no cues from other people directing how to interpret this situation. Additionally, we have good evidence that other subjects genuinely did not notice the confederate at all, and it seems plausible that these factors still had some effect on the interpretation of those subjects who do notice him.

Most wrongdoing does not occur in situations like these, and in most situations there are cues which can act as prompts towards doing the right thing. Recall agents with frontal lobe damage, who have considerably less ability to change tasks than most subjects, regularly overcome goal neglect when presented with very minor prompts towards certain courses of action. For most wrongful omissions, a failure to care sufficiently will act as a better explanation for wrongdoing. My argument that subjects who failed to help in Samaritan are partially excused thus does not threaten to excuse nearly all wrongful omissions.

4.3 Explaining Behaviour in Obedience

Turn now to Obedience. Though we can’t be sure that all obedient subjects shared a similar interpretation of the situation, I believe there are grounds for excusing many of the obedient subjects. My argument is this: a number of factors stopped many subjects from recognising that they had a choice, and without realising this, the subjects couldn’t form an intention to disobey. Subjects who did disobey seemed to do so precisely because they had this thought, plausibly due to previous experiences, being exposed to certain prompts, or even luck. When asking whether the obedient subjects had the capacity to disobey, we need to hold fixed the fact that they did not realise they could make such a choice. Once we do this, it seems the obedient subjects lacked the capacity to do the right thing for the right reasons.

In Chapter 2, I noted how various existing explanations of subject behaviour might seem more plausible if we began thinking about subjects’ perceived reasons for action or construal of their situation. In Chapter 3, I noted the importance of recognising that agents can be more or less aware of particular features of their environment, and thus particular considerations. And in my analysis of Bystander and Samaritan, I argued that numerous studies show that it is very difficult to recognise particular considerations in ambiguous situations.
Although it seems very unambiguous to us what is happening in the experiment, and very clear what courses of action the subjects can take, a number of factors plausibly had a distorting effect on subjects’ interpretation of their situation.

First, the experimenter seems unperturbed by the behaviour of the learner, which would make most reasonable people question their evaluation of the situation. The experimenter looks like someone who has a greater understanding of what is going on, and who is much more familiar with the experiment. Subjects didn’t have anyone else to look to for cues, which would make it hard for them to reconcile their initial interpretation of the shocks as harmful with what the experimenter’s words and manner seem to be suggesting. Subjects would also have the reasonable belief, at least for some of the experiment, that the experimenter would eventually come to share their concern and stop the experiment.

Perceiving the experiment to be unethical requires perceiving the experimenter as someone who is highly unethical and completely insensitive to the harm they cause others. Seeing the experimenter in this way is very difficult because he seems in every other way to be very pleasant, unworried, authoritative, normal, and to be someone who has apparently gained approval from a reputable university to conduct such experiments. All of these things make it initially hard to perceive the experimenter as a wrongdoer, which in turn makes it hard to see following their directions as doing the wrong thing.

We have already seen from Bystander that which features of our situation we attend to can affect our capacity to do the right thing for the right reasons. It is also worth considering the many things that subjects in Obedience attended to and thought about, which plausibly

---

73 McGeer and Pettit (2015b) argue that when thinking about an agent’s capacity to recognise and respond to moral reasons, we should also think about their disposition to respond to a real or imagined moral audience, since the moral community helps us develop and sustain our moral capacities, teaching us right from wrong. We have a disposition to respond to our audience’s expectation that we respond to reasons, and we want to do so in order to in fact respond to reasons. This account suggests an interesting take on why these subjects in Obedience are excused: as the only person in the room, the experimenter acts as distorting moral audience for subjects. The subjects’ normal audience is absent, removing a crucial feature that ordinarily supports or scaffolds their moral decision-making, and subjects are presumably unaware of this distorting influence (insofar as most people don’t realise the extent to which moral decision-making is supported by our moral audience). The unusualness of the situation makes it difficult for subjects to ignore the present audience in favour of a more reliable imagined audience who would help them attend to the actual moral reasons. However, given their commitment to an agency-cultivation model of responsibility, McGeer and Pettit would argue this distorting audience doesn’t make blame inappropriate, because blame would still support and develop the obedient subjects’ agential capacities. See supra §1.4.
distracted them from cues which might lead them to realise they could disobey. Subjects had clearly arrived at the conclusion that they ought to register their concern to the experimenter, to get him to stop the experiment. Having formed this intention, the experimenter’s lack of uptake would be very puzzling, and subjects’ attention would plausibly remain on trying to figure out why he wasn’t responding in the expected manner, thinking about whether there was something else they could say to impress their concern upon him, and weighing up the likelihood that the learner might get the next answer right. These (very reasonable) targets of subjects’ attention would make it harder to consider other possibilities, such as what would happen if they chose to leave the room, or call for help, or steadfastly refuse to continue.

One objection to the claim that subjects’ construal of their situation was ambiguous regards the subjects’ distress: If subjects couldn’t perceive their situation in the way we do, and so didn’t believe what they were doing was wrong, why were they so anxious about what they were doing? I don’t think the obedient subjects believed that what they were doing was permissible. As we already saw when considering the evidence for why addicts and people with extreme phobias or PTSD are excused for their actions, our perceptual seemings can be very hard to ignore when we deliberate or act. Subjects may have had occurrent beliefs like ‘the learner is in pain’ and ‘I don’t want to do this’ or ‘I should convince the experimenter to stop’, and this would explain their distress. But they may have also simultaneously failed to form any firm thoughts like ‘I could refuse to keep shocking’ or ‘the learner is in serious danger’. Given the obedient subjects’ ambiguous construal of their situation, they oscillated between very conflicting judgements about what was happening, and what was likely to happen if they continued shocking. They did not form definite beliefs or judgements about what they ought to do, all-things-considered.

An important set of studies can help explain why the gradual increase in shocks seems to be so effective in producing obedience. These studies show that humans are reliably affected by hystereresis, a phenomenon in which the point at which we change our judgements, perceptions or attitudes regarding some variable depends on which direction our judgements, perceptions or attitudes regarding that variable came from.74 For example, most philosophers know that when constructing a phenomenal sorites series and leading students through the inductive premise, which vague predicate is applied to items in the middle of the series depends on which end of the series was chosen as the starting point. Although real adults

---

74 Thanks to Mark Colyvan for pointing me to these studies.
presented with heaps of rice or thermostats will change which vague predicate they apply at some point (i.e. they never call a single grain of rice a ‘heap’), the point at which they change from ‘heap’ to ‘not heap’ itself changes depending on which direction they come from.

Raffman (2014) showed subjects colour slides that gradually changed from blue to green, asking subjects to categorise them as either ‘blue’ or ‘green’. Once subjects switched from ‘blue’ to ‘green’, the experimenters then reversed direction, showing subjects the colour they had categorised as ‘blue’ moments earlier. Subjects would now categorise this slide as ‘green’ and continue to do so for a number of slides with increasing blueness. Once they eventually switched to ‘blue’, the experimenters switched direction again, at which point subjects would again judge the next several slides in the sequence to be ‘blue’. Hysteresis has been shown to occur in judgements regarding motion (Nichols, Huisman, Rivera, Hock, & Bukowski, 2005), other people’s emotions (Sacharin, Sander, & Scherer, 2012), speech (Tuller, Case, Ding, & Kelso, 1994), hearing (Chambers & Pressnitzer, 2014), ambiguous sentences (Rączaszek, Tuller, Shapiro, Case, & Kelso, 1999), customer satisfaction regarding delivery services, (Terence, Richard, & Ian, 1992), dating behaviour (Tesser & Achee, 1994), and business investments (Dixit, 1992).

This phenomenon seems to partially explain why most subjects who disobeyed did so when the learner first complained of heart trouble. Complaining of heart trouble acted as a ‘line’ for subjects to pick out, making them interpret any new shocks as being different to previous shocks, and making it easier to form a resolution to stop. It seems likely that had there been larger increases between the shock voltages, and correspondingly larger increases between the learner’s cries, more subjects would have disobeyed. The gradual increase in the shocks, which it is worth noting were initially harmless and thus morally permissible, anchored subjects’ construal of their situation. It would have been difficult to see administering shocks at 120V as impermissible, given everything about this decision looks almost identical to the decision faced at 115V. The gradually increasing scale acted as a psychological slippery slope for subjects, making it difficult to pick out any particular point at which they should stop.

Recall also that certain actions are easier to consider due to perceptual affordances. It seems notable that in Obedience, subjects are placed immediately in front of the electrocution box, and have only the experimenter to talk to, while the learner is in an entirely different room. These features of the environment would make talking to the experimenter and electrocuting the learner very obvious potential actions for consideration. In contrast, disobeying is not an action suggested by the environment at all, and the fact subjects cannot
see the learner makes it very difficult to think of actions that would need to be taken to refuse, or to see the learner as being someone they *can* help. Recall that even those subjects who disobeyed only had to register their concern with the experiment four times. Not a single subject went to the door and left the room, either to find someone for help, or to check on the learner directly.

The subjects had entered the experiment expecting to perform a certain task. They spent time memorising how they were supposed to act, and they administered ten questions and shocks before the learner first demanded to be let out. Though subjects managed to put this task on hold while talking to the experimenter, the experimenter kept prompting them to continue, directing their attention back towards the experiment and the actions that they’d just become familiar with performing. All of these factors plausibly made it difficult for the subjects to realise that there existed alternative courses of action. This seems to be an accurate characterisation of what Sabini and Silver (2005) were trying to talk about when referring to subjects’ confusion regarding the “moral world” and their “inhibition” from acting (p. 551).

Testimony from some of the obedient participants suggests that they really did not identify any alternative action to continuing the experiment:

The thought of quitting never occurred to me … just to say: ‘You know what? I’m walking out of here’ — which I could have done. It was like being in a situation that you never thought you would be in, not really being able to think clearly.

Participant Bill Menold, quoted in Perry (2013b).

This was a sheer reaction to a totally impossible situation … being totally helpless and caught up in a set of circumstances where I just couldn’t deviate and I couldn’t try to help. This is what got me.

Participant Morris Braverman, quoted in Milgram (1974, p. 54).

I believe that many of the disobedient subjects had the thought that disobeying was an option. This triggered something like a gestalt switch in how they saw their situation, resolving the ambiguity and enabling them to clearly see the moral considerations. This then allowed them to attend to certain features of their situation, ignore others, deliberate and form a firm intention to disobey. One source of evidence for this comes from Adrian
Dimow’s testimony, who reports having an easy time disobeying because his earlier experiences caused him to construe his situation differently to the other subjects. But another source of evidence that other disobedient subjects had different gestalts of their situation to the obedient subjects comes from looking at the effectiveness of the experimenter’s prods. It is notable that of the four prods used to get the subjects to continue, the last – ‘You have no choice, you must go on’ – was completely ineffective. Not a single person who heard this prod continued to administer shocks. This prod is much more direct than the previous three, and it resulted in a number of subjects refusing in a way that directly referenced their capacity to choose to not continue. Consider these transcripts of Milgram’s experiment taken from Gibson (2013):

Experiment: You have no other choice you must continue.
Participant 2032: I have another choice. I won’t continue.

Experiment: You have no other choice, sir, you must go on.
Participant [unknown]: If this were Russia maybe, but not in America.

Experiment: You have no other choice you must continue.
Participant 2036: [Oh I] certainly do have, you can have your cheque back sir.

Experiment: You have no choice, really.
Participant 2005: Why?

It seems plausible that the reason the fourth prod was ineffective was that it made it obvious to subjects that they had a choice. By telling subjects they ‘must’ continue, subjects’ attention would be drawn towards considering the supposed force of this ‘must’, which in turn made it much easier to recognise that no such obligation existed. It intuitively seems

---

Prod 1: ‘Please continue’, or, ‘please go on’. Prod 2: ‘The experiment requires that you continue’. Prod 3: ‘It is absolutely essential that you continue’. (Milgram, 1974, p. 21). It is also possible that the reason this prod was ineffective was due to selection bias, in that the kind of people who had resisted three prods may be the kind of people who would successfully resist four or more, regardless of the fourth prod’s content. But it does seem plausible the prod’s content had an effect, given more subjects probably would have disobeyed if the fact they had a choice was made salient earlier, and given these quotes show the refusals to continue are much more definite than subjects’ earlier responses.
likely that if the experimenter had explicitly asked the obedient subjects what their final
decision was, or if there was a large sign saying ‘you have a choice’, more subjects would
have disobeyed. These prompts would have triggered a certain way of perceiving their
situation, which would make disobeying much easier. The ineffectiveness of the fourth prod
relative to the previous three also seems difficult to account for in terms of any of the
previously offered explanations that I considered in Chapter 2.

Individually, the components I have pointed to don’t necessarily show that subjects had a
diminished capacity to avoid wrongdoing. But put together, they act as strong evidence that
subjects construed their situation differently to how we expect, and did not realise that
disobeying was something they were capable of choosing.

I think it is worth considering again our project of trying to explain why subjects in this
experiment obey, and my earlier discussion regarding how we assess when an agent lacks a
particular capacity. It is not enough to simply identify a plausible explanation for why so
many subjects obey in this experiment. We need to explain the relative variation in obedience
rates that occur between conditions. We need to explain why more subjects disobeyed when
the learner was in the same room or when they received the fourth prompt.

These subjects are ordinary people, who presumably care quite a lot about the well-being
of others. Almost all of them make several attempts to make the experimenter notice that the
learner is being harmed, and many more would have succeeded in not shocking the learner
had the fact they had a choice been more obvious. The subjects are not just distressed about
what is happening to the learner, they are distressed about what they are doing, and despite
not seeing any value in continuing, they do not stop. Prior to the experiment, the disobedient
subjects didn’t seem any more virtuous, caring, or strong-willed than the obedient subjects.
The main thing that seems to distinguish them from the obedient subjects is the way they
happened to interpret their situation. To interpret any given situation, our perceptual
capacities depend greatly on cues from our environment, other people, and previous
experiences, often without us realising this. When these ordinary sources of support are
removed and replaced with factors specifically intended to mislead our interpretations, it
becomes very difficult to know what to attend to in order to resolve ambiguities and
deliberate clearly. Given all of these considerations, I believe that many of the obedient
subjects are excused for their actions. They lacked the capacity to disobey in the sense
required for moral responsibility. They are not deserving of blame for their actions.
4.4 Objections

Let me now consider a number of objections to my argument thus far. One objection is that my account threatens to excuse too many wrongdoers. After all, our awareness of moral considerations is also affected by physiological drives for food, water, and even novelty.

I believe that physiological drives like hunger can reduce blameworthiness for some wrongs. While feeling a bit peckish after breakfast won’t excuse wrongdoers, an agent who has not eaten for days could be excused for not accurately seeing any moral considerations that are present against eating. In the Minnesota Starvation Experiment, Keys, Brožek, Henschel, Mickelsen, and Taylor (1950) famously found that volunteers who were forced to exercise, but fed only 1570 calories per day, became obsessed with food. They talked about recipes at length, fantasised about food by themselves, ate up to 40 sticks of chewing gum a day and become very agitated any time a meal schedule was changed. The hunger one feels when one is starving bears some similarities to addicts’ urge to use: it causes you to see eating as good, it makes food-related things pull your attention, causes you to imagine in great detail the reasons favouring eating, and one must continually act to inhibit themselves from eating if food is available. If, for some reason, an agent had to starve themselves to meet their moral obligations, at a certain point we could not appropriately blame them for finally eating. It seems true that most people lack the capacity to starve themselves indefinitely in the sense required for moral responsibility, even if some people do in fact manage to starve themselves to death in hunger strikes. Similar responses apply to people who are thirsty, tired, and even bored. Sensory deprivation has been used as a form of torture (McCoy, 2007; Wolfendale, 2009), psychologists have found that people will willingly induce pain in themselves to avoid the sense of boredom (Havermans, Vancleef, Kalamatianos, & Nederkoorn, 2015), and researchers find that boredom plays a significant role in accounting for why many homeless people spend money on alcohol rather than goods we might think are more essential (Burraway, 2018; Reeve, Green, Batty, & Casey, 2009). It could be argued that we take these things to be excusing conditions because they act as a form of costs agents have to bear in order to do the right thing. But even if we stipulate that an agent has a duty to bear these costs, the agent could still be excused because extreme thirst, tiredness and boredom would make the moral reasons that generate this duty far more difficult to attend to.

Another worry is that many of the morally relevant features we encounter are not very salient, or are usually only perceived in a coarse-grained manner. One might object that
failures to act on these considerations will thus entail a very low level of culpability. I don’t think this objection is convincing for a few reasons. First, while some situationist factors make us less aware of certain moral considerations, other factors make us more aware of these considerations. Cues from other agents in particular usually cause us to attend to moral reasons when deliberating. Second, the thought that a less-than-perfect level of awareness entails low culpability only holds if we assume that full responsibility can be achieved when we are ‘fully’ aware of moral considerations, whatever that amounts to. This is not the case. To be blameworthy for a wrong, agents need only be sufficiently aware of the moral considerations against wrongdoing. It is plausible that we are sufficiently aware of most considerations in most of the moral decisions we face, and so I am not committed to saying most ordinary wrongdoers are excused.

Most situations we find ourselves in are simply not like Obedience: There are not usually multiple tasks competing for our attention, we typically have cues regarding how to interpret our situation, and we can often consult other agents for advice. Another reason we don’t excuse wrongdoers with a lower-than-perfect level of awareness is that people generally have had a lot of experience with their surroundings and, over time, this adds up to an awareness of the morally relevant factors. People learn what features of their situation to attend to. Obedience, Bystander and Samaritan are unfamiliar situations with which subjects have no prior experience. People simply haven’t had much practice standing up to authority figures or recognising the social factors which influence moral behaviour.

Another objection is that my account threatens to excuse other moral wrongs that are not very salient, or which we find it difficult to be greatly aware of. Consider individual efforts to reduce carbon dioxide emissions, in order to mitigate global warming. Psychologists have argued that part of the explanation for why our collective responses have been lacklustre is that carbon emissions and their effects are not particularly noticeable to us (Gifford, 2011). Carbon dioxide is colourless, any individual’s emissions have a tiny effect on the earth and take years to have tangible effects, and many of our carbon-emitting activities seem very normal to us. These things all make it less likely that we’ll think about our emissions as we go about our days. These facts arguably make me committed to saying most people are somewhat excused for failing to reduce their emissions.

I think this is the right result, all else being equal. I think people are less blameworthy for their emissions, relative to cases where driving one’s car (say) would visibly result in rising sea levels destroying someone’s home. But I don’t think this means most people are, in fact, not
blameworthy for their emissions. Although the consequences of emitting carbon dioxide aren’t immediately obvious, scientists have been engaged in prolonged campaigns to make us very aware of them. We have had more than enough time and education to learn about the effects, and we have been given enough opportunity to develop habits and routines that reduce our emissions. The fact that the harms of individual carbon dioxide emissions are not very salient is no longer an excuse for most people.

If the International Panel on Climate Change had released their first report on the harms of carbon dioxide emissions, and no further media attention was given to the issue, then I would agree that most individuals are considerably excused for their emissions. But that is not today’s state of affairs. Likewise, if the subjects in Milgram’s experiments had been forewarned that there exist unethical experimenters who try to dupe volunteers, that many people obey these experimenters when they shouldn’t, and of the various psychological tendencies humans have which can inhibit them from helping, and then subjects obeyed in the experiment, they would be much more blameworthy. It would be much easier for them to attend to the features of their situation that made them fully cognizant of the moral considerations, and realise they could easily prevent the learner from being harmed.

A real-world example of how our assessments of blameworthiness seem to track the salience and obviousness of moral considerations is that of our responses towards consumer spending. Many people accept Singer’s (1972) argument that failing to donate to charity is just as wrong as failing to save a drowning child from a pond. However, few of us seem to act as if the two acts are equally blameworthy. If I told you I had just let a child drown on my way to meet you, you would be aghast. Compare this to if I told you that I spent $20 on coffee instead of Oxfam this week, which would hardly change your feelings towards me. If we accept Singer’s argument, the discrepancy in our blaming attitudes between these two cases seems hard to account for in terms of any objective moral considerations. But pointing to a difference in the obviousness of the wrongness of each act seems like a very plausible explanation for why there seems to be a difference between each act’s blameworthiness.

One last objection to my argument is along the following lines: Sure, the obedient subjects failed to have the thought that disobeying was an option available to them. And if we hold this fixed, it does seem like they lacked the capacity to disobey. But this is the sort of thing that we should not hold fixed. Indeed, this failure is something that they are blameworthy for! They could have realised that disobeying was an option. They had the capacity to realise they could disobey, and they should have known better.
This thought is very tempting. But as I noted in §1.7, philosophers have struggled to explain how an agent can be culpable for the fact that a consideration failed to occur to them. It seems difficult to explain what the agent should have done to remove their ignorance, given that, in most cases, they would also be ignorant of the fact that they ought to take steps to remove their ignorance. I think that we can explain why this intuition seems tempting in a number of ways that give us reason to not rely upon it when assessing the blameworthiness of the obedient subjects.

First, the fact that an agent fails to notice certain considerations is often reliable evidence that the agent does not care sufficiently about those considerations, which can make them attributability responsible. They can thus be appropriate targets of certain negative responses, even if they are not blameworthy. Second, many cases where an agent ‘should have known better’ are actually cases of recklessness, for which it is easy to ground culpability. The agent did, in fact, know better. They just didn’t care enough, and chose to ignore certain moral considerations. Even in cases where a consideration genuinely did not occur to the agent, it is often implausible that the agent did not earlier realise that they could take steps to prevent this lapse from occurring. For example, we have calendars and smartphones to prevent us from forgetting our friend’s birthdays.

Finally, the fact that the obedient subjects could have realised they could disobey is true, if we understand the sense of ‘could’ at play to be referring to general capacities. We feel similarly frustrated when a sports star misses the game-winning shot, when a chess master falls into the trap that their opponent set for them, and when an artist’s daring aesthetic choice ends up being ridiculed by the art community. All of these agents had the capacity to succeed at doing what we expect them to, and indeed, should have known better or performed better. They might also be appropriate targets of what we could call sporting, epistemic, and aesthetic ‘blame’, whatever those amount to (Wolf, 2015).

These factors all plausibly cause us to have the intuition that the obedient subjects could have and should have known better. But we need to remember that we are evaluating whether the subjects had the specific capacity to do the right thing for the right reasons, for the purposes of assessing their accountability responsibility, and that we are setting aside whether they could have taken steps prior to the experiment to avoid wrongdoing. When we keep this in mind, we should be suspicious of moving from the thought that the subjects ‘should have known better’ to the conclusion that they are deserving of blame.
4.5 Conclusion

In this chapter, I considered why so many subjects in the situationist experiments fail to do the right thing. I argued that subjects’ helping behaviour in *Dime* is best explained by saying that finding a dime placed them in a better mood which broadened their attention. Though this made them more aware of certain considerations, this does not show that subjects who failed to help without finding the dime are excused. Nothing shows that they were not sufficiently aware of the confederate dropping their papers. Given the wrong is rather minor, the best explanation of their failure to help is that they just did not care enough.

I then argued that subjects in *Bystander* who fail to help are excused. This is because a variety of studies consistently show that people’s likelihood of helping is affected by how unambiguous it is that someone needs help. Further evidence that subjects interpret their situation differently to how we expect comes from their testimony to the experimenters. This evidence also shows that a number of subjects in *Samaritan* genuinely didn’t notice the confederate, or didn’t interpret him as in need of help. These subjects are also excused. Some subjects seemed to notice him needing help but continued to keep walking. These subjects are blameworthy, but their culpability is mitigated because the factors causing other subjects to not notice the confederate at all would still have had some effect.

I then argued that most obedient subjects in *Obedience* are excused. I argued many factors plausibly made it very difficult for subjects to realise that disobeying was an option available to them. The unfamiliarity of the situation, the lack of other agents to look to for cues on what to do, the experimenter’s directions and hysteresis all made the situation very ambiguous, which in turn prevented subjects from considering that they could disobey. The disobedient subjects were able to disobey precisely because it occurred to them that disobeying was an option. This is evidenced by the ineffectiveness of the fourth prompt, the transcripts describing the subjects’ refusals, and testimony from the subjects.

Finally, I considered objections to my argument. These were that my argument entails that physiological drives and the fact some moral considerations are sometimes not very obvious can also act as mitigating factors. In response I argued that these factors can be mitigating, all else being equal, but in most cases other factors make us sufficiently aware of the relevant moral considerations, which make failures to do the right thing blameworthy. I also considered whether the subjects are blameworthy because they ‘should have known better’ and considered a number of factors which plausibly explain why we have this response.
Chapter 5: Excuse and Repair

5.1 Introduction

In the previous three chapters, I have argued for the conclusion that the obedient subjects in Milgram’s experiments have a reduced culpability for their actions. In this chapter, I consider an objection to my argument. This is that the victims of this kind of wrongdoing seem intuitively entitled to blame the obedient subjects. Had the learners actually been harmed, even partially excusing the obedient subjects would do a disservice to the learners, and this suggests that the obedient subjects are blameworthy. Call this ‘the victim objection’.

This chapter responds to the victim objection. In doing so, I aim to examine an aspect of our moral practices that have not received adequate attention in the moral responsibility literature: what kinds of responses are appropriate towards excused wrongdoers. Much of the moral responsibility debate seems to suggest that excused wrongdoers occupy the same moral status in our practices as someone who never committed any wrongdoing. I think this is mistaken, and that examining the status of excused wrongdoers is a worthwhile project.

In §5.1, I outline the victim objection. In §5.2, I examine another case where we have mixed reactions to an agent who appears to be excused for their actions: Williams’s (1981) lorry driver. I argue that our seemingly mixed responses are not inconsistent, as the driver has acquired new duties of moral reconciliation in virtue of accidentally killing the child, even though he is excused for the killing. In §5.3, I consider and respond to an objection. This is that excused wrongdoers who have incurred these duties could also be unable to recognise they have these duties, and thus be excused from fulfilling those too. In §5.4–5.5, I detail the nature of these duties, drawing from Radzik’s (2009) account of atonement. I argue that excused wrongdoers have duties to become worthy of moral reconciliation with their victims by showing a commitment to morality, providing reparations for the harm caused, and withdraw the insult that their wrong expressed.

Let me first add a brief note on terminology. Thus far I’ve adopted Nelkin’s (2011) account of moral responsibility, according to which agents require the capacity to do the right thing for the right reasons in order to be blameworthy. My argument in this chapter engages with arguments from a few different Control theorists, who identify slightly different capacities as necessary for responsible agency. Doris and Murphy (2007) talk about ‘normative competence’, and Brandenburg (2019) adopts Fischer and Ravizza’s (1998)
‘reasons-responsiveness’. For the purposes of this chapter, it doesn’t matter which particular account of control or capacities one takes to be necessary for responsible agency. For simplicity, I will follow the current philosophical literature and adopt ‘reasons-responsiveness’ as a placeholder term to refer to the capacity agents must possess in order to count as responsible agents under Control accounts of moral responsibility.

5.2 The Victim Objection

Recall from Chapter 1 that advocates of Control accounts of moral responsibility maintain that agents must have a certain kind of control over their actions in order to be directly morally responsible for those actions. Attributionists deny this, and it is their objections against Control accounts that I want to consider.

The Attributionists’ main objection to Control accounts is that Control accounts appear to be under-inclusive regarding who qualifies as a morally responsible agent, excusing wrongdoers who we commonly take be blameworthy. Additionally, while we often excuse agents who lack control over their actions, close attention to key thought experiments purportedly reveals that lack of control is not what excuses these wrongdoers. Rather, factors which undermine agential control often, but not always, show that a wrong is not connected to the wrongdoer in the way required for culpability.

Talbert (2008, 2012), for instance, argues that when a wrong action expresses an objectionable evaluative judgement that the wrongdoer did not have sufficient reason to do otherwise, this judgement makes the wrongdoer blameworthy. On this account, even if an agent cannot perceive or understand the moral features of a situation, this will not excuse the agent if they were aware of all the relevant non-moral facts. If I back up my car over you, but didn’t know you were there, then my actions don’t express anything objectionable about you and I’m not blameworthy (assuming I checked my mirrors and took adequate care). If, on the other hand, I back up my car over you knowing you’re there, this is objectionable even if I am incapable of understanding how your presence gives me any reason not to back up my

---

76 Talbert (2009b) does look at Obedience but doesn’t come to any explicit conclusion about the obedient subjects’ culpability. He appears to think there isn’t enough information to determine whether the obedient subjects were sufficiently aware of their actions, or whether the pressures of the situation were somehow blocking these subjects from acting on good reasons (p. 428). However, he believes that if the obedient subjects are blameworthy, it will be because the wrongdoing come about in such a way that their actions can be considered to express objectionable attitudes.
car. I am thus an appropriate target of blame. Using arguments along these lines, Talbert argues that psychopaths (2008), the morally ignorant (2013), the morally insane (2012), and agents manipulated into having artificial judgements (2009a) are all appropriate targets of blame for wrongdoing. And if the actions of the obedient subjects can be considered to express objectionable attitudes about the learner, then the subjects will be blameworthy.

While these debates are important, I don’t want to enter into them directly. Because Control theorists are willing to be more revisionist than Attributionists, they can reply by saying that our moral practices are mistaken when presented with counterexamples, or our blaming practices simply don’t track who is blameworthy in the basic desert sense. Instead, I want to consider a more substantial objection that Attributionists have against Control theorists, which has only received indirect attention. Recall that many Control theorists see blame as only being appropriate when it is deserved, as blaming agents for things they lacked a fair opportunity to avoid seems unfair. Though Talbert (2012) has not directly argued against the claim that agents are blameworthy only if they are deserving of blame in Pereboom’s (2014) basic desert sense, he has made some points about fairness which could be fashioned into such an argument. Talbert agrees that intrapersonal fairness matters to our assessments of blameworthiness, but argues that Control theorists are focusing on the wrong party. He argues that excusing perpetrators who lacked control is radically unfair to victims, and this should take priority in our theory of moral responsibility. Victims deserve to be able to blame perpetrators who have wronged them, even if those perpetrators lacked a fair opportunity to avoid wrongdoing. Conversely, denying victims an entitlement to blame their perpetrators does a further disservice to them.

This point is best summed up in Talbert’s reply to Doris and Murphy (2007), who provocatively argue that the perpetrators of the Abu Ghraib prisoner abuses are excused for their wrongdoing. Few people are aware of the circumstances that the Abu Ghraib soldiers were in prior to their wrongdoing. The area was subject to grenade and mortar attacks up to twenty times per week, fostering high levels of fear and anxiety. One soldier was so affected by seeing another soldier get blown up that he never took off his flak jacket. The prison had no sewage system, and the porta-potties regularly overflowed. The prison would turn into a

---

77 Talbert also does not find this charge convincing, because the point can generalise in a problematic way. He argues ill-willed agents who are reasons-responsive and thus blameworthy on Control accounts can make similar complaints of unfairness: “If anything makes it difficult to respond to moral reasons, being naturally bad or sadistic will do so” (2009b, p. 430).
sauna, heating to over 45°C in the summer, and the walls were unable to keep out wind storms. There was no top-down supervision, as the General in charge was also the head of sixteen other prisons and abandoned her duties at Abu Ghraib because of the danger and awful living conditions. The soldier in charge of the night shift would work up to forty days straight at a time and slept in a prison cell. Some prisoners were violent, with corrupt Iraqi guards smuggling in guns and weapons to the prisoners, making the soldiers even more fearful, distrusting and hypervigilant (Zimbardo, 2007). Doris and Murphy argue that this environment, lacking in supervision, support, or basic amenities for months, led to incredible levels of stress, anxiety and cognitive degradation, entailing that the soldiers lacked the capacities required for accountability responsibility.

Talbert objects that, even if we stipulate that the soldiers were cognitively degraded, and even if we do nearly always excuse people who are cognitively degraded, these soldiers are appropriate targets of blame for their actions. This is because on almost any account of moral responsibility which gives weight to our actual moral practices, it seems appropriate for these victims to blame their perpetrators:

Let us suppose that the soldiers at Abu Ghraib abused detainees because, under the pressures of military life, they unavoidably came to view the detainees as contemptible creatures who did not have the standing to object to being used as objects of sadistic amusement. In this case, if we were to say that, from the standpoint of strict moral fairness, the detainees should not regard their abusers as morally accountable for their actions, we would be telling the detainees that they are mistaken to regard their own victimization as good grounds for moral blame. But this seems to require the abused detainees to view their own victimization as something other than what it was and to view the abuse they suffered as morally analogous to a regrettable accident, a justified punishment, an attack by a wild animal, or some other harm not intentionally inflicted on the basis of morally offensive attitudes. But the treatment the detainees received was not accidental or justified. The abuses were intentionally inflicted, and if they were motivated by certain kinds of objectionable attitudes, then the people who suffered on account of the attitudes have as much right and reason to blame the people who abused them as anyone ever has.

Call this ‘the victim objection’. In this chapter, I want to take the victim objection seriously. The victim objection is that excusing perpetrators who are not reasons-responsive seems to treat victims unfairly, and this appears to be a regular feature of our everyday moral practices. We should therefore reject the claim that wrongdoers are blameworthy only if they are deserving of blame in the basic desert sense, which would undermine my argument that the subjects in Obedience are not blameworthy.

In principle, Control theorists could argue that concerns about victims are simply unrelated to questions about basic desert, and that there might be all-things-considered good reasons to allow victims to blame excused perpetrators, even though, strictly speaking, their blame would be inappropriate. I think such a response would be unfortunate. Radzik (2009) points out that “the problem with a wrong committed against another person is not merely that a norm or rule has been violated but that a valuable being who deserves respect has been harmed and demeaned” (p. 90). Control accounts would appear much stronger if they could show that when we let victims of excused wrongdoing believe their blame is appropriate, we are not merely indulging them. In this chapter, I want to respond to the victim objection without simply asserting that Attributionists are using the wrong sense of blameworthiness, that basic desert should be our primary concern for theorising about moral responsibility, or that revisionism does not count as a cost against a theory.

The parallels between Doris and Murphy’s argument and my own are obvious. Both look at settings in which ordinary people end up doing terrible things we would not expect them to do. Both argue that the wrongdoers in these settings ought to be excused for their actions due to a lack of reasons-responsiveness. And both of our conclusions are incongruent with many people’s intuitions and our moral responsibility practices. The victim objection is thus clearly applicable to my argument thus far. Even if the obedient subjects in Milgram’s experiments entirely lacked reasons-responsiveness, they harmed the learner, and this might entitle the learner to blame. The obedient subjects felt some concern for the learner, but they did not feel enough concern to motivate them to disobey. Their actions could be considered to express a judgement that the learner’s interests were not important enough to warrant standing up to the experimenter, and this judgement is morally objectionable.

Talbert’s point is powerful, even if one is not an Attributionist. A large proportion of us are deeply uncomfortable saying that the soldiers at Abu Ghraib are not blameworthy for their actions. We would not think it appropriate for them to plead innocent in court. We would not be comfortable with them being found ‘not guilty’ by a judge. We would believe
something had gone wrong for any account of moral responsibility which entails that the soldiers were allowed to go home and act as if they never did anything wrong. Even Doris and Murphy agree with Talbert’s overall sentiment. In presenting their argument, Doris and Murphy recognise their conclusion is intuitively “monstrous” (2007, p. 26). Likewise, it does seem that many people would blame the obedient subjects, and that the learner in Obedience would be particularly licensed to blame the subject who had just electrocuted them. This resentment would likely persist even if, after reading up on the theories of moral responsibility, the victim believed that subjects had a reduced capacity to avoid wrongdoing.

In order to make their conclusion more palatable, Doris and Murphy argue that we ought to separate moral responsibility from criminal liability, and adopt a doctrine of what is known as ‘strict liability’ for wrongs such as those committed by the Abu Ghraib soldiers. Strict liability is a legal term which amounts to finding that an agent or institution is to be held responsible and penalised for their actions, regardless of whether they are culpable or not. For example, in some US states, having sex with someone under the age of sixteen is a strictly liable offence known as statutory rape, and perpetrators can be criminally liable even if the minor had greying hair and a birth certificate saying they were born in 1950 (Carpenter, 2003). Unfortunately, endorsing strict liability doesn’t help us with the victim objection, as strict liability is widely regarded by philosophers as an incoherent account of moral responsibility. This is because to be excused for an action just is one of the ways in which you can be causally responsible for an action but not blameworthy, the others being if you were justified or exempt.

I believe that it is possible to respond to the victim objection in such a way that does not require doing a disservice to victims of wrongdoing, abandoning a Control account of moral responsibility, or endorsing something as contentious as strict liability. This is the task that I shall focus on in the remainder of this chapter. My argument relies on a new understanding of what being excused for wrongdoing entails. In short, I argue that excused wrongdoers can still gain new duties of reparation. However, arguing that new duties can account for our practices risks being unconvincing to Attributionists, since they can maintain that such duties...

---

78 Rosen (2004) writes that “in morality there is no such thing as strict liability” (p. 301). Nagel (1979) agrees that “strict liability … may have its legal uses but seems irrational as a moral position” (p. 31). And Brink and Nelkin (2013) argue “strict liability is extremely problematic within the criminal law, where guilt results in blame and imprisonment … conviction without culpability denies the fair opportunity to avoid wrongdoing” (pp. 306–307).
themselves are evidence that such wrongdoers really are blameworthy. A stronger argument could thus be made if it were shown that wrongdoers who both Attributionists and Control theorists agree are excused still incur such duties. Let me now consider such a case.

5.3 Agent-Regret and Appropriate Blame

In *Moral Luck*, Williams (1981) describes a thought experiment in which a lorry driver hits and kills a child while driving. The accident is not the fault of the driver. The driver was doing the speed limit, paying attention to the road, and had perfect reaction times, but he nevertheless kills the child. Since the driver lacks control over the final result, and no ill will or negligence was displayed in his actions, Williams’s question is whether the driver ought to feel guilt over what has happened. This case is relevant to theorising about blameworthiness, because guilt is often thought to be the first-person analogue of blame (Carlsson, 2017).

We have conflicting attitudes towards the driver. On one hand, given the death was an unavoidable accident, it seems that the driver ought not to feel guilt. Were he to feel guilt and blame himself, we would console him and draw his attention to the fact that there was nothing else he could have done. On the other hand, however, Williams points out that we also think that the driver ought to feel bad about what has happened. We may initially expect him to feel remorse, given that remorse “in general is something like ‘how much better if it had been otherwise’” (p. 27). But further reflection reveals this is not enough. An uninvolved passer-by should also feel some remorse at the death of a child, but we would not expect her to feel remorse in the same manner as the driver. This is evidenced by how we would react if the driver did not feel any remorse at all. Imagine if the driver were to say, ‘After thinking about it, I’m excused for this death, so I actually don’t feel bad at all. There’s nothing I could have done, so I’ll be on my way’. If we saw this, we would be horrified and immediately reproach this callous driver to reflect on what has just happened. We would demand he realise not only that a child has died, but that he has killed a child. While a passer-by who shows no concern and keeps walking is callous and worthy of some reprimand, the driver’s callousness is much more objectionable, and in a different kind of way. For the driver to maintain that he had nothing to feel bad about “would be a kind of insanity … and it would be an insane concept of rationality which insisted that a rational person never would” (p. 125).
While Williams argues that our conflicting attitudes can be explained by invoking the existence of agent-regret, I would like to provide an alternative diagnosis.\(^79\) I believe that we in fact do not have contradictory responses to the question of whether the driver is morally responsible, because our responses arise in two different cases. In one case, the driver has agent-regret, and in the other he does not. Philosophers who analyse this case make the mistake of interpreting our varying reactive attitudes towards the driver to be a source of evidence about the appropriateness of blaming him for hitting the child, instead of taking the driver’s response to be a separate behaviour which may be a legitimate target for our reactive attitudes (e.g. Domsky, 2004; Moore, 2009; Rosebury, 1995).

Claiming that our reactive attitudes are responding to the driver’s own response towards the child’s death (rather than responding to the fact that he killed the child) is promising because this is something he does have control over, potentially making him an appropriate target of our reactive attitudes according to Control accounts. This would give us a way to grant that the child’s family are entitled to blame the driver if he doesn’t express remorse, while also giving him a fair opportunity to avoid blame. This would make it possible to agree with the thrust of the victim objection, without having to abandon the claim that blame is appropriate only if it is deserved.

But as noted earlier, an uninvolved bystander can also respond in various ways towards the child’s death. And since their response doesn’t seem to trigger the same varying attitudes in us, it appears that the driver’s response alone can’t be what explains our varying attitudes towards him. This seems like a problem, but I nevertheless believe there is something connecting the driver’s causal involvement in the death with how he responds which our reactive attitudes then target. The challenge is to explain why our attitudes towards the callous driver differ to our attitudes towards the callous bystander.

The relationships that people are in often determine which moral reasons or obligations are present, and this can in turn affect one’s degree of blameworthiness for wrongdoing. For example, the degree to which an agent is blameworthy for walking by someone who is unconscious will differ depending on their relationship to that person. If they are strangers, the passer-by may be described as callous. If a doctor walks by an unconscious patient, they

\(^{79}\) Agent-regret is regret at “actions in which he regards himself as a participant … In this case, the supposed possible difference is that one might have acted otherwise, and the focus of the regret is on that possibility, the thought being formed in part by first-personal conceptions of how one might have acted otherwise” (p. 27).
can be guilty of malpractice and failing in their professional duties. And if a parent walks past their unconscious child, they will be guilty of negligence and possibly child abuse. In each case, the wrongdoer’s level of blameworthiness depends on what duties they have. That is, their blameworthiness depends on what they are morally required to do in virtue of their relationship.

Though philosophers writing on moral responsibility are aware of this, I am drawing attention to it because I believe that the way many thought experiments proceed leads philosophers to misdescribe some cases. Philosophers typically begin their arguments by focusing on thought experiments where agents clearly have a duty but violate it. These philosophers then think about our reactive attitudes and moral practices regarding this wrongdoing. This is useful, as it could be confusing to focus on cases where an agent’s blameworthiness is unclear in virtue of our uncertainty regarding whether they violated a duty. But in doing so, these philosophers risk missing cases where our reactive attitudes are themselves evidence that an agent had a certain duty that we had not previously recognised.

Williams’s lorry driver thought experiment is one such case. My diagnosis of our intuitions is this: harm-doers and wrongdoers often acquire new duties to try and repair the relationship with their victims, even if they are not deserving of blame for their wrongdoing. That the driver was the cause of the child’s death is enough to make it the case that he now has duties to apologise to the family and to show concern, even if he is completely excused. Being excused for killing someone does not place one in the same category of obligation-holders as someone who never killed at all. Being excused does not make one uninvolved.

Furthermore, the manner in which the driver takes up or rejects these duties is something which expresses regard for the harmed parties, and responds to moral reasons. If the driver apologises profusely and tries to make up for his action, he is displaying a considerable level of positive regard towards the dead child and her parents. His attitudes and behaviour communicate the tragedy of the situation, and act as evidence that no ill will was present in causing the harm, even if we already knew the death was unintentional. Conversely, if the driver denies any wrongdoing, treats the child’s death as simply unlucky and rejects having

---

80 Wolf (2001) tries to resolve our mixed reactions to the lorry driver thought experiment by arguing we are responding to the existence of a ‘nameless virtue’ which we expect people to exhibit in such situations. Wolf thinks the driver should engage in self-blame, but he is not a fitting target of third-party blame from others, since he simply couldn’t have done otherwise. Whereas her virtue-based account places much emphasis on feeling certain things, my approach is duty-based, and emphasises engaging in certain ways with the family and other affected parties.
any new duties, he thereby displays a level of disregard for which we can blame him. I will spell out the nature of these duties in §5.4 and §5.5. Let me first make some comments regarding the broad idea that excused wrongdoers can incur certain duties.

While claiming that excused wrongdoers can incur new duties may be somewhat novel in the moral responsibility literature, upon reflection it is not that surprising or controversial. This is something we experience regularly and take part in reflexively. If I knock your vase over, the required response is to apologise, clean it up and offer to buy a new one, even if I was in no way negligent and took reasonable precautions to avoid breaking anything. Richards (1986) provides a thought experiment in which your neighbour goes to pick up your baby. Unbeknownst to her, your baby has been crawling in soap. As a result, your neighbour drops your baby, killing her. Richards argues this “would change for ever your feelings about [your neighbour] – no matter how clear it was that no one could have held that tiny, slippery, wriggling body” (p. 208). I believe this fact greatly affects what responses from your neighbour are morally permissible. Our apologies in these settings are not just politeness or a social norm, they are the morally required response to the situation, and those apologies or lack thereof are an appropriate target of our reactive attitudes. Again, if the wrongdoer in these cases focused on how they were excused and therefore need not feel any agent-regret or guilt, we would blame them and think doing so is appropriate, even if we agree they are in fact excused for the original wrong.

I take this intuition to be shared by a number of philosophers. Similar sentiments have been expressed by Thomson (1980), Lomasky (1991) and Feinberg (1978), regarding a case where a backpacker caught in a blizzard breaks into an occupied hut and uses the food there to keep himself alive. Although the remedial duties in this case are primarily cashed out in terms of compensation, these philosophers agree that the backpacker has incurred some duties, even if his actions were perfectly justified or excused. Other philosophers agree that similar principles apply in other cases. Cane (2002) argues that “Our responsibility to repair bad outcomes can extend further than our responsibility to avoid faulty conduct … once we take account of the interests of the victim, it seems less clear that victim-focused obligations of repair should always depend on fault” (p. 107). Miller (2001, p. 458) argues that if I accidentally knock someone over, I have an obligation to help that person back to their feet and make sure that they are not hurt. This obligation is stronger than that of other passers-by, even though I did nothing wrong, and could not have taken any other steps to avoid knocking him over. And Kramer (2005, p. 328) points out that if an innocent man is sent to
prison, the fact that there was overwhelming evidence suggesting he was guilty doesn’t absolve the state of any remedial duties. At the very least, they are required to make a formal apology, and the man would be justified in feeling aggrieved at his treatment.

I believe that the intuition of these philosophers generalises to the wrongs performed by the subjects in *Obedience*. Had the harms been real, these would strongly affect the learners’ view of the world, trust in others, trust in shared moral norms, and sense that their dignity is respected by others (Walker, 2006). These factors put the obedient subjects in a particularly unique position to remedy the wrongdoing, and distinguish our expectations of them wrongdoer from other expectations we might have of uninvolved parties.81

If one finds it unintuitive that agents could involuntarily gain duties in virtue of being an excused wrongdoer, it is worth noting that agents can involuntarily incur duties towards certain victims without being wrongdoers at all. Arguments for this claim come from the literature on collective responsibility. Much of the literature on collective obligations focuses on basic duties of assistance, duties of compensation or duties to cease collective wrongdoing (e.g. Collins, 2019; Lawford-Smith, 2012). But some philosophers also believe that individual agents can incur duties in virtue of being a member of a group, even if that agent has not previously behaved in any wrongful manner (e.g. Miller, 2004; Räikkä, 1997; Wringe, 2010, 2014). For example, Radzik (2014) argues that contemporary German citizens have particular political duties to preserve the memory of Nazi atrocities. This is in order to restore trust between Germany and previously victimised groups, and to demonstrate a commitment to never allowing such atrocities to occur again.

Another source of support for my position is that we expect members of groups to denounce wrongs committed by fellow members. At the time of writing, there are many calls for the Catholic Church to condemn the actions of Cardinal George Pell, who was convicted of multiple counts of child sexual assault (C. Smith, 2019). We expect representatives of the

---

81 Miller (2001) points out that there seem to be some cases where causing someone to be worse off does not generate new remedial duties, such as outcompeting a business rival. I have the intuition that there do seem to be some instances where businesses owe something to people they cause to be worse off, such as in cases where a big corporation’s lower prices result in workers at small local shops becoming unemployed. One possible explanation for this difference is that entering into a business for profit and failing merely deprives business rivals of a benefit they had no prior entitlement to (customers’ patronage). In contrast, unemployment substantially impedes workers’ ability to provide for themselves and their families. Another possible explanation is that unlike most forms of employment, starting a business is an inherently risky task, where one implicitly consents to bear the costs of failure.
Catholic Church to do more than, say, the Anglican Church, even if those Catholic representatives were equally as unable to prevent the assaults as the Anglican clergy. If the Catholic bishops insisted they did nothing wrong, and so this matter had no significance for how they ought to behave, we would think those bishops are appropriate targets of blame.

One final set of cases where people involuntarily incur duties despite not being deserving of blame comes from cases where we cause ‘hurt feelings’. Shoemaker (2019) argues that there are many instances where we hurt someone’s feelings, by making them realise that how we feel about them is worse than how they had hoped, expected or believed. For example, we can create hurt feelings by failing to like a present that was bought for us, failing to reciprocate someone’s interest in us, or forgetting the face of someone we’ve met several times before. These are cases where we do not violate any obligation, do not express a lack of regard or ill will, and our actions do not reflect objectionable attitudes. However, Shoemaker argues that it is appropriate to feel guilt when one causes hurt feelings in such cases, and I believe such agents are obliged to apologise and try to repair the relationship.

Returning to the tension between the positions held by Doris and Murphy on one hand, and Talbert on the other, my approach is a promising means of resolving our conflicting attitudes without having to bite any bullets. Assume that Doris and Murphy’s argument is correct. The Abu Ghraib soldiers were not reasons-responsive, and so are excused. We can still maintain that they have a duty to make up for the harms they caused, and that they are blameworthy if they do not do so. On this approach, we do not have to treat the harm done to the victims as analogous to an unfortunate natural disaster. We do not have to adopt a doctrine of strict liability. And we do not have to blame people who are undeserving of blame due to a lack of control over their actions. That my account is able to resolve our initial conflict without giving up one of these commitments counts considerably in favour of it. If the soldiers are excused for their actions, this alone does not yet settle whether they are appropriate targets of blame. They are still subject to further obligations which are relevant to our assessments. If the soldiers were cognitively degraded, they are not to be blamed as

---

82 One might still want further details regarding the scope of such duties, given worries that a mere causal connection to a harm seems insufficient for duties of reparation. Some previously offered suggestions regarding what kind of connection between the harm-doer and victim in the strict liability literature, which can help us narrow down the relevant conditions, are when the agent was the primary cause of the harm, was the instigator of the harm, or violated the victim’s rights. See Capes (2019); Couto (2018); Gardner (2005, 2011); MacCormick (1977).
instigators of violence because they lacked the control or reasons-responsiveness required for moral responsibility. But they nevertheless have a duty to ‘make things right’.

If the soldiers were to feel overwhelming guilt and disavow their actions, our negative reactive attitudes towards them would lessen considerably. When setting up the victim objection earlier, I described the discomfort we’d feel were the soldiers to plead innocent and be found not guilty for their actions. I believe that a large part of what motivates the victim objection is that we presume that excusing the soldiers entails this outcome. We presume that to excuse the soldiers is to grant them the liberty to act as if they never did anything wrong, equal to that of an uninvolved bystander, and this is an outcome which does seem unfair to the victims. But being excused and being granted the liberty to go home and not make up for what has occurred are two different things. The former does not necessarily entail the latter. When we recognise this distinction and have an alternative account of what it means to be excused, the victim objection loses much of its force.

The relevance of this argument to my previous analysis of moral responsibility in situationist settings should be clear. Even if we accept that the obedient subjects in Milgram’s experiments had a reduced capacity to disobey, we may still feel that the learner would be licensed to blame the teachers for the harm caused if the teachers did not at least apologise. We would be uncomfortable with the obedient subjects feeling no regret, or showing no concern for the learner and going home, even if we believe that the subjects are completely excused. Once we have an account of responsibility and excuse which avoids this, however, the claim that the obedient subjects are excused is much less unintuitive.

In the remainder of this chapter, I will examine two important questions to clarify my account. The first regards whether victims are entitled to blame wrongdoers who cannot yet recognise these duties of reparation. The second regards what kinds of responses would amount to meeting these duties. I will then consider some objections.

5.4 Reproach

One preliminary objection to this account is that the things which excuse perpetrators for their wrongdoing may also prevent them from recognising their present duties of reconciliation. They could thus be excused for failing to take up their duties, and therefore not be appropriate targets of blame for this failure. The victim objection could then apply again to these kinds of cases. For example, suppose the obedient subjects are excused while
electrocuting the learners because their rational capacities are impeded, but regain these capacities afterwards. The learners are entitled to blame the subjects after the experiment if the subjects fail to take up their duties of reparation. But surely it’s appropriate for the learners to blame the subjects while the subjects are unable to recognise the duties they have incurred. Surely it’s appropriate for the learners to blame while they are being shocked. I agree that the victim objection could once again apply to cases where wrongdoers cannot recognise their duties, and that it intuitively seems appropriate for victims to blame perpetrators while they are being wronged. My answer is that, although blame is impermissible, certain blame-like responses are permissible, as they do not amount to sanctions.

Brandenburg (2019) claims that communicative accounts of blame should make a distinction between blame and reproach. While both communicate on her account, and often overlap, they communicate different things. Both are “an affective response that, when expressed, communicates the relevance of some moral norm and invites the addressee to reflect on her conduct” (p. 173). But reproach does not communicate anything about what is deserved, is not hostile, and does not necessarily carry any suggestion “that we consider the addressee’s conduct to be an expression of ill will or moral disregard” like blame does (ibid.).

Brandenburg argues that the distinction between blame and reproach is important. This is because it helps us make sense of an important class of responses we have to marginal agents, who fail to possess reasons-responsiveness, but who still participate in our moral practices in some ways. Brandenburg cites the case of Earl, a person with a severe brain injury in a documentary by Louis Theroux (Pickup & Theroux, 2016). In one scene, Earl seems unaware that he has hurt his mother with his words, despite her crying right in front of him. Louis sternly addresses Earl and tells him he should apologise, at which point Earl appears to realise his wrongdoing and comforts his mother. Earl is not a responsible agent because of his brain injury, and thus is not deserving of blame. But being reproached causes him to recognise and react to moral reasons, and Louis’s reproach seems appropriate.

An appealing feature of Brandenburg’s distinction between blame and reproach is that it allows us to identify which blame-like responses are appropriate towards children and explain
why, something philosophers have found difficult to provide a convincing account of. Children are clearly not responsible agents, but we do respond to them in ways which look and feel a lot like blaming. Brandenburg points out that we regularly reproach small children, particularly when they are in the middle of wrongdoing, in order to make them more attentive to the moral reasons that are present. Though the child is able to successfully act on a certain reason after being reproached, it remains the case that the child is not a reasons-responsive agent. They could not perceive or react to the relevant set of reasons prior to being reproached. This account is very simple, and I believe more true to the phenomenology of engaging in moral address with children than other proposed accounts.

Brandenburg’s aim was to make sense of an important way in which marginal agents take part in our moral practices, assuming that blame primarily has a communicative function. But her distinction is also helpful for making sense of our responses to wrongdoing committed by responsible agents who temporarily fail to be reasons-responsive. I do not subscribe to a communicative account of blame, and thus cannot distinguish blame and reproach in virtue of their communicative content like Brandenburg does. Instead, as I see it, reproach differs from blame because it is justified by different considerations. Because reproach’s point is to make wrongdoers recognise their wrongdoing, or which moral reasons are present, as soon as this occurs and receives the right kind of uptake, reproach ceases to be justified. In contrast, blame, being a deserved sanction, remains appropriate from the moral community and victims even after the wrongdoer apologises and regrets their actions. There are limits to this – one can blame disproportionately, or for too long – but generally, blame does not suddenly become inappropriate when the wrongdoer recognises their moral character of their actions. In short, the difference is that culpable wrongdoing makes one deserving of blame, while one’s failure to perceive moral reasons makes one liable to reproach.

That reproach is only instrumentally justified up until it succeeds in making the wrongdoer recognise moral reasons can be seen by reconsidering the case of Earl. Imagine if Louis had continued to express negative reactive attitudes to Earl after Earl had apologised.

---

83 E.g. Vargas (2012) argues that our blaming responses to children are ‘feigned’ but very few philosophers have explicitly looked at what kinds of responses to children are appropriate. Control theorists would say children aren’t blameworthy for a wrong action if they lacked the capacity to avoid performing it. Others argue in various ways that children can’t really express an objectionable quality of will in the sense required for moral responsibility (e.g. McKenna, 2012). Neither of these accounts seem to accurately account for how we do seem to respond to children with responses that look and feel a lot like blame, and which seem appropriate.
to his mother. Louis clearly wouldn’t be trying to get Earl to recognise the moral considerations present. By hypothesis, Earl already sees them. Louis’s continued indignation would instead be a species of blame, one which strikes us as unfair and undeserved. As evidence of this, it would sound odd if Louis used negation to argue he wasn’t blaming: ‘I’m not blaming you, I’m just telling you I think you should feel remorseful about what you did in a very raised voice’. Once Earl has recognised what he has done and begins responding as he ought to, both blame and reproach are inappropriate.

Upon reflection, we often blame agents after they have apologised or recognised what they’ve done, and we think this is appropriate. This seems like an objection to communicative accounts of blame, since blame seems to lack any point once its message has received uptake. Fricker (2016) avoids such an objection by treating blame as a cluster concept, and arguing she is focused only on paradigmatic instances of expressed blame because our blaming practices are so varied and ‘disunified’. But for those of us interested in a more unified account, the blame qua sanction account can easily explain why blame remains appropriate. Blame remains appropriate because it continues to act as a sanction, which the wrongdoer is deserving of even if they feel remorse.

Here’s what recognising a distinction between blame and reproach means in practice. Once reproach succeeds and our temporarily excused wrongdoer perceives the moral features of their situation, the appropriateness of continued negative reactive attitudes depends on how the wrongdoer acts. They now have a choice. They can begin acting in accordance with their duties, in which case they are no longer liable to reproach nor deserving of blame. Alternatively, they can ignore these duties in which case they are now knowing wrongdoers and deserving of blame, though not deserving of blame for their initial harmful action. To apply this distinction to the wrongdoing in Abu Ghraib or Obedience, I believe that victims are entitled to reproach their respective wrongdoers until those wrongdoers recognise the moral duties they are presently subject to. If they accept those duties, neither reproach nor blame is appropriate. If they reject those duties, they knowingly act wrongly and are deserving targets of blame.

One might worry that what I am calling reproach is just blame by another name, gerrymandered in order to avoid the objection of unfairness usually levelled against Attributionists’ account of blameworthiness. To put the worry another way, why not think that being reproached when one is not a responsible agent is also unfair or undeserved? Can’t reproach sanction just like blame does?
Worries about desert or unfairness do not apply to reproach like they do to blame. Consider an instance in which an agent temporarily loses reasons-responsiveness, is reproached, and regains their normal capacities. They could not reasonably object to having been reproached earlier, or demand some apology from those who reproached them. Indeed, if their wrongdoing was entirely caused by the fact they were not reasons-responsive, they shouldn’t feel wronged or mistreated. Reproach doesn’t say anything about them in the way that blame *qua* sanction does. Being reproached while one is not reasons-responsive is analogous to being locked up while one is in the middle of a hallucination and at risk of harm to others. So long as the negative treatment ceases when one regains their normal rational capacities, one has no grounds for objection. One could not complain that their being locked up was a form of punishment, even though it is true that being locked up usually is a form of punishment. This is not the case with blame. If one is continually blamed for things one did when one lacked a fair opportunity to do otherwise, one does have grounds for complaint.

My account is thus able to explain why excused wrongdoers seem like such appropriate targets of blame-like responses while engaged in wrongdoing, or while they are yet to take up their duties incurred in virtue of their wrongdoing. We are entitled to reproach them, insofar as our aim is to have them recognise relevant moral reasons and duties. We may want them to feel remorseful because this is the most effective way for them to recognise the moral character of their actions. But this should not be thought of as a sanction or as deserved, even if it is unpleasant for the person on the receiving end.

5.5 Atonement and Reconciliation

The next question concerns what these duties consist in. In *Making Amends: Atonement in Morality, Law, and Politics*, Radzik (2009) develops a detailed account of what duties wrongdoers incur towards victims, the community and themselves. Although she focuses on culpable wrongdoers, her account can be extended to cases of excused wrongdoing too.

Radzik argues that wrongdoers incur a duty to atone in virtue of their wrongdoing. While a ‘duty to atone’ might sound punitive, Radzik’s understanding of atonement differs from our folk concept, which has associations with suffering, paying off a debt, and religious overtones. Instead, the original sense of the term was ‘at-one-ment’, implying a reconciliation in the relationship between perpetrator and victim. Radzik argues that the suffering that
comes with guilt is typically good evidence that one takes their wrongdoing seriously, but wrongdoers don’t have any obligation to suffer.

To atone is to take steps that make one worthy of redemption and thus reconciliation with those they have wronged. A wrongdoer becomes worthy of being ‘redeemed’ by making themselves worthy of re-evaluation: literally, worthy of being deemed (evaluated) again. They become worthy of re-evaluation in virtue of their response to their earlier wrongdoing. In particular, wrongdoers must merit being given esteem and trust to conduct themselves appropriately, in order to re-establish a proper relationship with the moral community. When the wrongdoer atones, she gives the victim good reason to stop structuring their relationship to one another in terms of the roles of wrongdoer and victim.

Radzik claims that wrongdoers do not have an obligation to successfully reconcile with victims. Since some victims will refuse to reconcile, this would make any duty to reconcile beyond the wrongdoers’ capacities. However, wrongdoers do have a duty to make themselves worthy of being reconciled with, and to create the conditions in which such reconciliation could reasonably take place. This may be impossible if the victim is dead. But partial atonement is possible in these cases, and worth aiming for.

In order to merit reconciliation, wrongdoers typically need to do three things. First, they must morally improve themselves. It is not enough to just convince others that they are trustworthy, they must actually become trustworthy. Second, they must change the meaning of their wrong act, by withdrawing the insult and the threat expressed in the wrong act (this will be explained in §5.6). This requires communicating to the wrongdoer, in such a way that is likely to be understood as a withdrawal of the insult, typically by apologising. Third, the wrongdoer must make reparations for any material, physical, psychological, and relational damage that has occurred. Reparations should be made as symbolic gestures communicating that one is remorseful, and that one cares sufficiently about the victim. This is why symbolic gifts are usually more conducive to reconciliation than mere money, and why receiving $1000 from someone poor is more meaningful than receiving the same amount from someone rich. Symbolic apologies show that the perpetrator has thought about the victim and is sensitive to their cares. It is notable that psychologists have found that apologies are much more effective when there is a cost to the wrongdoer, with this being the case across cultures and religions (Ohtsubo & Watanabe, 2009; Ohtsubo et al., 2012).

Most cases of atonement, but not all, will have all three components, and many actions can fulfil more than one at the same time. For example, giving an expensive gift can help
compensate for damages, express that one is sorry, and count as the kind of action that reflects a commitment to morality. The degree to which wrongdoers must atone is highly context-sensitive, but should be proportional to the severity of the wrong done. Given that victims are often well-placed to understand the impact of the wrong, atonement will be proportional if the victim would judge it to be proportional were they well-informed, rational, and not influenced by any vice such as greed or arrogance.

Radzik indicates that she doesn’t intend for her specific account to apply to excused wrongdoing. For example, she notes that wrongdoers ought to feel guilt, which involves recognition that one has violated a norm “under conditions that neither justify nor fully excuse its violation” (p. 35). But a number of considerations show that the main parts of her account can be extended to excused wrongdoers.

First, the motivations for the account apply to cases of excused wrongdoing too. Radzik is motivated by the thought that in the aftermath of wrongdoing, we ought to give adequate consideration to victims’ interests and their perceptions of the wrong. In many cases, we think that victims ought to have some say in how wrongdoers are to make amends. These concerns are also present when we look at victims of excused wrongdoing. The victim objection is motivated by the thought that we ought to give adequate consideration to victims’ interests qua the person being wronged, and that they ought to have some say in how the perpetrators ought to behave, even if those perpetrators are excused.

Second, the link that Radzik draws between culpable wrongdoing and our affective responses seems applicable to cases of excused wrongdoing too. Radzik thinks that it is impossible for humans to take our wrongdoings seriously without feeling remorse and guilt. If a wrongdoer is genuinely committed to morality, their adherence to moral norms matters to them, and in normal humans this will be reflected in their affective responses. Such responses will often be painful (which is why atonement often has connections with suffering), and move us to seek reconciliation with those we have wronged. The point of examining Williams’s lorry driver case is that our excused wrongs also matter a lot to beings like us, and for this not to be reflected in our affective responses would be “a kind of insanity” (1981, p. 125). If someone is committed to morality, their conduct usually matters to them, including parts of their conduct that they cannot completely control.

One particularly strong example of this comes from Sophie’s Choice (Styron, 1979). In this novel, Sophie must choose one of her children to be killed. If she refuses, both will be killed. Sophie chooses her daughter and is understandably wracked with guilt about her decision,
even decades later. She is not blameworthy in any way, does not deserve to feel guilty, and yet we would feel something had gone wrong if she didn’t feel guilt. It also seems relevant that certain kinds of agents who are taken to be excused by many philosophers nevertheless report feeling guilt for their wrongs. Guilt is felt by people with agoraphobia who miss out on important events in their child’s lives (Illawarra Mental Health Information Service, 2005; Pendergraft, 2017), and people with post-traumatic stress disorder report feeling guilt at their irritability or outbursts (Starcevic, 2009). Philosophers often argue gambling and drug addicts are excused for their actions (Franklin, 2013; Kennett, Vincent, & Snoek, 2014), but these agents report feeling guilt for the wrongs their addiction leads them to commit (Locke, Shilkret, Everett, & Petry, 2013; Merritt, 1997). This evidence is not conclusive – perhaps addicts remain somewhat blameworthy for their actions (Schroeder & Arpaly, 2013) – but it is yet to be shown that addicts feel significantly less guilt than fully responsible agents who perform similar wrongs, which is what we’d expect if addicts’ level of blameworthiness was driving their level of guilt.

Finally, and most importantly, there seems to be no reason why excused wrongdoers could not be expected to meet Radzik’s three subgoals of atonement in order to reconcile with victims. These were to make reparations, make a commitment to morality, and to retract the insult expressed in the wrong. Let me examine each in turn.

As noted earlier, many philosophers have previously argued that excused wrongdoers ought to compensate their victims. This is similar to the subgoal of making reparations. The difference between providing compensation and making reparations is that when making reparations, (i) the wrongdoer must take into account the psychological, emotional, and physical harm they have caused, (ii) any gifts provided must not be implied to be morally comparable to the harm caused, and (iii) the gesture should be made as a symbolic expression that one takes the wrong seriously and cares sufficiently about the victim. I see no reason why philosophers who endorse the claim that excused wrongdoers ought to compensate victims could not endorse the stronger claim that excused wrongdoers ought to make reparations in such a way that meets these conditions too.\(^84\)

The second subgoal of atonement is to make a commitment to morality. Although an excused wrongdoer may not have conducted herself in any blameworthy manner, it seems

\(^{84}\) This would be difficult in the *Sophie’s Choice* case, given the victim is dead, but my point would be just as applicable if Sophie had to pick a child to lose one of their legs.
reasonable that, in many cases, she still ought to commit herself to ensuring that something similar does not happen again in the future. Her experiences will make her better able to recognise factors likely to lead to similar wrongdoing, which puts her in a position to recognise situations where she is at an elevated risk of wrongdoing. For example, an alcoholic can display a commitment to redemption by staying away from places likely to be serving alcohol, and by actively helping other alcoholics get sober. Applied to Obedience, the obedient subjects might have duties to be more on guard against ending up in situations with social pressure to act immorally. Additionally, if an agent is blamelessly disposed towards wrongdoing in certain situations, they can take steps to reform their character prior to being present in those situations. Perhaps an adult wrongdoer’s bad childhood makes them excused for their angry outbursts, but they can still attend counselling or anger management, and it seems reasonable for us to blame them if they do not.

One might ask whether it is reasonable to demand that all excused wrongdoers demonstrate a commitment to morality. If the wrong was an unusual, one-off event, such that we do not think that the wrongdoer has failed in their commitment to morality, is there really any benefit to making a new commitment to morality? I am happy to accept that in some cases, this demonstration of commitment is unnecessary, just as Radzik accepts that not all culpable wrongdoers need to meet all three subgoals. This seems particularly plausible if one is not disposed towards committing similar wrongs in the future, and the commitment will not act as a form of apology or help make reparations. It is worth noting, however, that we often think that people ought to renew commitments that have not been broken. Couples often renew their marital vows to one another, and birthdays can act as an occasion for friends to re-affirm their commitment to friendships, even if that commitment was never in doubt. One will also have considerable reason to renew a commitment if agents in similar circumstances regularly do so, as a failure to follow the norm may be interpreted as a lack of commitment.

This leaves us with only one subgoal of atonement from Radzik’s account: to apologise. Although it may seem intuitive enough that excused wrongdoers ought to apologise to

---

85 Sophie might be an example of an agent for whom this step seems unnecessary, particularly given another agent is directly causing the harm to the child. This case is complicated by the fact that (i) it is a moral dilemma, and so is isn’t clear whether Sophie’s actions are wrong, but excused, or permissible, and therefore justified, and (ii) there seems to be something unfair about having someone re-commit to morality if morality doesn’t allow Sophie to avoid wrongdoing in.
victims, Radzik’s specific account of apology is quite nuanced. It also allows me to respond to one last objection readers may have.

5.6 The Social Meaning of ‘Sorry’

While thus far I’ve argued that there isn’t anything implausible about expecting wrongdoers to perform the same actions that culpable wrongdoers ought to, my opponents might want something stronger than this. They might want a positive account for why excused wrongdoers ought to apologise. Why is it not enough for the wrongdoer to simply explain their excusing condition to the victim, and convince them that any hard feelings they have are inappropriate? Shouldn’t victims and the community believe that the wrongdoer merits esteem and trust upon seeing the evidence that they are not a deserving target of blame?

My answer relies on some insights from communicative and conversational accounts of blame, which Radzik endorses. Communicative accounts of blame treat blame as communicating something to wrongdoers, in response to something that has been communicated or expressed in the wrongdoer’s actions. A number of philosophers argue that wrongs have expressive power (Hampton, 1988; Hieronymi, 2001; McKenna, 2012). As Murphy (1982) puts it, “[moral] injuries are also messages – symbolic communications. They are ways a wrongdoer has of saying to us, ‘I count but you do not,’ ‘I can use you for my purposes,’ or ‘I am up here on high and you are down there below’” (p. 25). McGeer and Pettit (2015a) argue “the offender displays a disregard for the protections that we in the society had provided for you. Those protections signal a clear message of respect: that you are one of us, a person adequately and equally protected against offences of the kind you suffered. But the offender, speaking from within our ranks, communicates a message that denies you this status, giving the lie to the public signal” (p. 330). Inwronging, wrongdoers act as if the victim’s status is not a reason to not perform the wrong action. This insults the victim. Similarly, our responses to wrongdoing also have an expressive function, and many philosophers argue that expressing disapproval is one of the functions of state punishment.

I don’t want to commit to the claims that all of our moral responsibility practices should be understood as an analogue of a conversation, or that all actions are expressive, or that the meaning of actions is what we target when we blame, or that an agent is blameworthy if their actions have an insulting message. I am simply endorsing the weaker claim that many of our actions have a certain meaning or express certain things, and that this is often morally
important to understanding the nature of particular wrongs. Adultery typically expresses a lack of commitment to one’s spouse, even if the adulterer hopes that it does not. Gift-giving generally expresses some appreciation or care for the recipient, and it does so in a way that simply writing on a piece of paper that one cares, or pointing to a list of reasons why the other person should think one cares, does not.

The insult that Murphy describes can also act as a form of testimony that this treatment is acceptable. The severity of the wrong can correlate with the severity of further potential harms that the victim has reason to fear, and others may be encouraged by the perpetrator’s apparent claim to superiority. Because of this, Hieronymi (2001) argues that wrongful actions are not merely insults, but also threats. They make a threatening claim about how we are allowed to be treated by others. Without any retraction, the insult or threat persists against the victim. Along similar lines, McGeer and Pettit (2015a) argue that in being wronged, victims suffer arbitrary interference, threatening their status as someone whose liberties are protected by the community, and thus their status as citizens free from domination.

Radzik and Hieronymi argue that the meaning that wrongs express cannot only be determined by wrongdoers. If this were the case, it would make no sense for wrongdoers to ever seek forgiveness, or feel the pain of remorse. Wrongdoers could simply now disapprove of their past actions, and nullify the claim previously made. Instead, the meaning of wrongs is essentially social, and this is what wrongdoers who feel guilt want to change. The social meanings of wrongs persist after the wrong has occurred, and this can affect relations of mutual regard between the wrongdoer and victim, and the wrongdoer and the community.

This account of the social meaning of wrongs provides us with two explanations for why excused wrongdoers ought to atone, and why it is sometimes not enough to simply explain to the victim and moral community that they were excused. The first is that some excused wrongs can also have a social meaning that insults and threatens. One example of how such claims can persist comes from the actions of schoolyard bullies. These are people who are not blameworthy for their actions, in virtue of their having been a child at the time of wrongdoing. Nevertheless, they often report experiencing guilt as adults, and feel compelled to make amends with their victims (Anonymous, 2016; Wargadiredja, 2017). Perpetrators of such harm in their youth have an obligation to apologise and make amends, even if they were technically excused for their wrongdoing. Making amends with their victims seems required
in virtue of the fact that they are well-placed to withdraw their earlier insult, and ensure that any threat to their victims’ social standing is removed.\textsuperscript{86}

For some cases of excused wrongdoing, it seems implausible that the wrong could insult or threaten the victim. Anyone who understood all the facts of Williams’s lorry driver case could not reasonably interpret the driver as insulting or threatening the child. But there is a second way in which the social meaning account of wrongdoing can explain why the driver ought to make amends. This comes from the fact that one’s actions after excused wrongdoing also carry a social meaning.

Many philosophers have argued that, if the state fails to punish a criminal act, this amounts to condoning that crime (Duff, 2000; Feinberg, 1965). It seems that in the moral domain too, failure to condemn a wrong is often equivalent to condoning the wrong, particularly when one stands in close relation to the wrongful act. Wrongful actions must be condemned because “repudiating immoral action is constitutive of a commitment to morality” (p. 45). If a victim’s friend does not repudiate the wrong done to the victim, or if a parent does not apologise to their child’s victims, this can communicate that they find the treatment acceptable, even if they had nothing to do with the original wronging. As I argued earlier, the relationships one is in affect what duties one has, excused wrongdoers are in particular relationships with their victims, and the appropriateness of our reactive attitudes seems very sensitive to how excused wrongdoers treat their earlier wrongdoing. This is not to say that all failures to condemn other people’s wrongful actions thereby condone those actions. Often, we need to respect people’s privacy. But when we stand in a close relation to

\textsuperscript{86} At first glance, Hieronymi seems to disagree that excused wrongdoers can make threatening claim, saying only ‘authored’ events make claims: “An action carries meaning by revealing the evaluations of its author. The event could not make a claim or carry meaning (positive or negative) if its perpetrator were not capable of making moral statements with his actions. The past event would not be a threat to your worth if it were not authored. We don’t resent accidents, because they make no such claim” (p. 546–547). This seems to imply that excused wrongdoers cannot make threatening claims. But shortly after this passage, Hieronymi claims that we do not blame children, not because they do not author claims, but because “their claims don’t carry enough weight” (p. 547). Since children are widely considered to be excused for many wrongs they commit, it seems that if children can make threatening claims, then in principle there could also exist excused wrongdoers who succeed in making threatening claims that do carry weight. I believe the obedient subjects in Milgram’s experiments, and the Abu Ghraib soldiers, are wrongdoers of this kind. In later work, Hieronymi (2004) does argue that blame is only fair if it targets a wrong which issued from an agent’s real self and reveals the agent’s evaluative attitudes. However, this argument makes no reference to the idea that wrongs have a social meaning, and doesn’t explicitly comment on whether children can be blameworthy.
the wrongful act, failing to condemn it can send the message that one condones it. Similarly, failing to apologise or show concern for a victim can send the message that the victim was deserving of the treatment they received.

Additionally, even if members of a particular class of wrongdoers are excused for their actions, we should be sensitive to the fact that reasonable third parties will not always reach this conclusion. Members of the moral community can reasonably expect to make assessments about the culpability of its members for wrongdoing, as can the victim. These assessments should not be dismissed out of hand on the grounds that they are mistaken. Dismissing the community’s concerns, particularly when the community has a justified (albeit false) belief that serious culpable wrongdoing has occurred, can show disrespect and disregard to that group. Excused wrongdoers have a duty to at least engage with the concerns of people affected by the wrong.

Furthermore, it is not always possible to demonstrate to the community and victim that one was excused for their wrongdoing. There may be contradictory evidence, and it is very unlikely that discussions of culpability will proceed using the concepts and sophisticated arguments found in philosophical articles. In these cases, both atoning and refusing to atone carry costs, which excused wrongdoers ought to consider. If they go through the motions of apologising, making reparations and showing a commitment to morality, they may feel like they are incurring an unfair burden. But refusing to atone risks making them look like someone who does not care at all about the victim, and who believes that the wrong is none of their concern. This will often be interpreted as a lack of regard for the victim, and add to the insult the victim has suffered. This, in turn, means that if a wrongdoer is aware that not apologising may look like a lack of regard and cause victims to feel even more insulted, the choice not to apologise can then in fact express a lack of regard. For example, if you non-culpably forget your friend’s birthday, failing to apologise for this and take steps to show that there was no lack of care on your part will itself then express a lack of concern, precisely because you know that failing to apologise will likely be interpreted this way.

When we understand apologies as the retraction of insults, and atonement as the steps that one takes in order to become worthy of reconciliation, it seems plausible that excused wrongdoers can incur duties to atone. They are not deserving of blame for the original wrong; they do not deserve to be sanctioned. But they ought to communicate to the victim that they are sorry and that they take the victim’s interests seriously, and engage in reparations for any harm caused. They should ensure that any insult or threat that has come
about through their actions is retracted and acknowledge that the victim deserves to be treated better. If they fail to do these things, they risk condoning the original wrong. They fail to show sufficient regard for the victim, and this can make them a deserving target of blame.

We are now in a position to respond to the victim objection as it applies to *Obedience*. The obedient subjects are excused. But their actions are likely to carry a social meaning, which will be interpreted as a threat to the victim. Their actions are going to be understood in a certain way by reasonable people, namely the learner. Even if the moral community and learner’s assessments of the obedient subjects’ culpability are mistaken, the subjects ought not to simply dismiss these assessments. It might simply be too much to ask of the learner to have them think about rational capacities, desert or the fairness of blame, and to revise their understanding of the harm that was done to them. Talbert is right that we shouldn’t merely ‘indulge’ the learners and let them blame anyway. But this doesn’t mean we therefore have to say the wrongdoers are in fact deserving of blame. We can recognise that the learners’ interests and understanding of their treatment are morally relevant considerations which affect how the obedient subjects are permitted to act. The obedient subjects ought to apologise, demonstrate that they care about the victim, take steps to make reparations, and commit themselves to ensuring they do not end up in similar situations in the future.

5.7 Objections

My argument that excused wrongdoers can acquire duties of moral reconciliation is able to explain many of the conflicting attitudes we have towards wrongdoers in key thought experiments, coheres well with our moral practices, and doesn’t give up our commitment to blame being appropriate only if it is deserved. Let me now consider possible objections.\(^{87}\)

The Attributionist could insist that my account still fails to give the victim what they are entitled to, which is to blame the wrongdoer. They could argue that directing wrongdoers to redeem themselves goes some of the way to refuting the victim objection, but is not enough.

---

\(^{87}\) As already noted, one potential worry regards wrongdoers whose victims are now dead. Duties towards the dead are a tricky area of ethics, and raise questions outside the scope of this thesis, such as whether the dead can be harmed or wronged (e.g. Johansson & Ekendahl, 2014; Wisnewski, 2009). It seems plausible that, even if one cannot directly reconcile with the dead, one can engage in symbolic actions which show one takes their wronging of the dead agent seriously, and perhaps this is enough to ground the claim that excused wrongdoers can acquire duties of reconciliation when their victims are dead.
In response, I would like to point to empirical evidence showing that duties directing wrongdoers to redeem themselves are well-suited to accommodate our concern for victims (Braithwaite, 2002; Daly, 2017; Strang, 2001). Walker (2006) notes that the empirical literature on the aftermath of severe and widespread wrongdoing shows that better outcomes are achieved when repair and restoration are emphasised, rather than punishment and retribution: “A large number of studies … show fairly consistently that restorative justice satisfies victims, offenders, and communities better than does existing (retributively based) criminal justice practices. In particular, many studies confirm that restorative justice practices such as mediations and conferences get high marks for procedural fairness and respect for participants, tend to lead to higher rates of completion of restitution agreements, tend to elicit apologies from offenders much more frequently, and tend to reduce fear, anger, and vengeful feelings among victims” (p. 213). Relatedly, Funk, McGeer, and Gollwitzer (2014) found that punishment alone is not satisfying for victims; rather, victims are more concerned with whether the punishment results in a change to the perpetrator’s moral attitudes and behaviour. This evidence adds support for my position, because what motivates the victim objection is a concern for the victim’s interests, and a wariness that deeming their blaming attitudes as inappropriate could be unjust. The real-world cases Walker refers to suggest that victims’ interests could, in many instances, be better served if our practices emphasised wrongdoers’ duties of moral reconciliation instead of victims’ entitlement to blame.

Another objection might come from counterexamples showing that duties of reconciliation are too demanding, particularly in cases where the wrongdoer really couldn’t have done otherwise, or has already incurred significant costs. It is important to note that, when I say that agents who fail to meet their duties of moral reconciliation are appropriate targets of blame, this does not entail they ought to be blamed all-things-considered. Their duties are pro tanto duties, and so other considerations can excuse wrongdoers from fulfilling these duties, and I suspect that this will be doing most of the work in counterexamples that could be proposed. Suppose that Williams’s lorry driver had a stroke, and this caused him to hit the child. It would be very inconsiderate to insist he should immediately uphold his duties of moral reconciliation towards the family, and some may take this to show that he does not have any such duties. But my account can accept this result by providing an alternative explanation. The fact the driver has suffered something life-threatening, is in hospital, and will likely be immobile for quite a while, are considerations which excuse him from immediately carrying out these duties. Such an example does not show that the acquisition of
duties of moral reconciliation is implausible, only that in some cases wrongdoers will not be blameworthy for failing to fulfil these duties.

A relevant question here regards cases where the excused wrongdoer never regains the capacities necessary for responsible agency. For example, if psychopaths are incapable of recognising moral reasons, this would mean their victims are never entitled to blame them for failing to engage in moral reconciliation. Reproach would also be inappropriate, strictly speaking, because the psychopath cannot come to recognise present moral reasons by being reproached. The victim objection would thus apply to these cases.

My response to this objection is that insofar as we think that victims of psychopaths are entitled to blame, this is indeed merely an indulgence because psychopaths simply cannot be held to account as members of the moral community who have flouted our demands. It is worth noting that our intuitions about these cases may be tracking confounding factors, because whether psychopaths are genuinely unable to recognise moral reasons is disputed (Jurjako & Malatesti, 2018; Maibom, 2008; Vargas & Nichols, 2007). Additionally, Robert Harris appeared to repent for his actions, and had something of a return to the moral point of view while in prison (Watson, 1987). If this is right, he would indeed have duties of moral reconciliation. A thought experiment which is less intuitively messy is that of the psychlops from Watson (2011). The psychlops is “a creature with two basic ends: to eat food that it finds tasty, and to preserve its own life. This is all that basically matters to it” (p. 311). This creature genuinely does not understand morality, does not care for it, and simply likes eating people in the same way many people like eating ice cream. Though blaming such a creature for harming you or a loved one would be understandable, I don’t believe such a being can be considered anything like a moral agent who is deserving of blame. Such a being does seem more like the ‘wild animal’ referred to by Talbert in setting up the victim objection, and not a moral agent whom we can blame for failing to comply with moral demands.

In cases where wrongdoers can recognise their duties of moral reconciliation, are capable of fulfilling these duties, and there are no competing considerations which outweigh these duties, failure to meet these duties makes blame appropriate. The Abu Ghraib soldiers had many opportunities to show remorse during their court proceedings, and could have tried to engage in atonement and reconciliation once they were released from prison. They have failed to take even minimal steps towards doing this. Charles Graner, considered the main ringleader, pleaded ‘not guilty’ at his trial. When asked if he had any regrets or apologies, he replied “No ma’am” (Pandiotti & Polk, 2005). Lynndie England, who appeared smiling in
many photos with the abused prisoners, has said “I’m sorry” in one interview (Moore, 2008). But she largely seems unapologetic and cavalier about her history in later interviews. In 2012, England said, “[The Iraqi victims’] lives are better. They got the better end of the deal … They weren’t innocent. They’re trying to kill us, and you want me to apologize to them? It’s like saying sorry to the enemy” (Hall, 2012). Interviews in Kennedy’s (2007) documentary The Ghosts of Abu Ghraib show that the other soldiers generally fail to see themselves as owing anything to the prisoners, focusing instead on assigning blame to the overall circumstances and military for the abuses. The fact that the soldiers have made no effort to apologise or reconcile with the prisoners, despite having the capacity and ample opportunity to do so, makes them appropriate targets of our reactive attitudes.

In contrast, Milgram’s subjects had little opportunity to apologise to the learner. They were quickly informed that the learner had not actually had a heart attack, and then rushed out the door by the experimenters (Perry, 2013a). Whether they actually incurred some duties of reconciliation is complicated by the fact that no harm actually occurred, and the learner was an actor complicit in inducing the subjects to obey. Had the shocks been real, and the learner an innocent victim, our assessments of whether the obedient subjects are blameworthy would depend on how the subjects responded after the experiment. If they apologised and made sufficient effort to ensure that the learner was okay, both physically and psychologically, then they are not blameworthy. But if they denied any wrongdoing or duty to make amends, then they would be deserving targets of blame.

5.8 Conclusion
In this chapter, I considered a widely held intuition that victims are entitled to blame their perpetrators, even if those perpetrators lacked reasons-responsiveness. This suggests that subjects who commit wrongdoing in the situationist experiments are appropriate targets of blame for their actions, which acts as an objection to my previous chapter’s argument. This is the victim objection. I argued that Control accounts can accommodate the intuitions supporting the victim objection, without thereby giving up their commitment to blame being appropriate only if it is deserved. I did this by arguing excused wrongdoers can still incur duties of moral reconciliation, finding support for this thesis from a number of areas.

First, many philosophers share the intuition that excused wrongdoers can incur new duties, and some philosophers have argued that agents can involuntarily incur new duties
without being wrongdoers at all. Second, our seemingly contradictory responses to another case of excused wrongdoing, Williams’s lorry driver, can be rendered consistent by accepting that he has incurred duties of moral reconciliation. Third, and most importantly, there seems to be no principled reason why the duties that culpable wrongdoers incur could not apply to excused wrongdoers too. Using Radzik’s account of reconciliation, I argued that excused wrongdoers ought to show a commitment to morality, make reparations for any harms caused, and apologise to the victims. Failure to do this displays a disregard for victims’ interests, and acts as a second wrong making wrongdoers deserving targets of blame.

I considered a worry that the victim objection could apply to cases where agents are temporarily incapable of recognising their duties of reconciliation. In response, I argued that reproach responses remain appropriate, as these do not constitute a sanction. I also considered a worry that the victim objection applies to agents who are permanently incapable of recognising these duties, and argued that such agents are simply not appropriate targets of blame. Though blame from victims may be understandable, these agents should be thought of as more like wild animals than moral agents.
Chapter 6: Situationism, Subjunctive Hypocrisy and Standing to Blame

6.1 Introduction

In the previous five chapters, I have examined whether subjects who act wrongly in the situationist experiments are appropriate targets of our blame. I have argued that most subjects who committed wrongdoing in *Obedience*, *Samaritan*, and *Bystander* are excused for their actions, because they lacked the capacity to do the right thing for the right reasons, making it inappropriate for us to blame them. The final two chapters of this thesis will also argue that it is inappropriate for us to blame these subjects, but from a very different angle. In these chapters, I want to make the novel argument that it is inappropriate for us to blame these subjects even if they are morally responsible and thus blameworthy for their actions.

Though I’ve argued that subjects who act wrongly in *Obedience*-like settings are excused, I recognise that many philosophers believe that these subjects are blameworthy for their actions and therefore appropriate targets of blame. To briefly summarise: Brink (2013) argues that since it is likely that the subjects in Milgram’s experiment would have disobeyed given more time to deliberate or greater incentives, they had the capacity to disobey but simply failed to exercise this capacity. McKenna and Warmke (2017) argue that though the combined situationist literature suggests that we have less capacity to avoid wrongdoing than we like to think, subjects in these settings are still reasons-responsive to a degree sufficient for responsible agency. Levy (2014c, 2015a) argues that even if situationist factors can change which action an agent performs, they can’t make subjects act in such a way that conflicts with their values entirely. Because subjects are aware of what they are doing, and have the capacity to do otherwise, they are morally responsible for their actions.

In this chapter, I will grant these philosophers their conclusion, and examine whether it follows that blaming the obedient subjects is inappropriate. Though it may seem like granting that the obedient subjects are blameworthy entails that it is appropriate for us to blame them, this is not necessarily the case. This is because the appropriateness of blame not only depends on facts about the wrongdoer, but facts about the blamer too. Specifically, many philosophers believe that it is possible for someone to be an appropriate target of blame, for there to be no weighty consequentialist considerations against blaming them, and yet, for it
be inappropriate for certain agents to blame them. One notable case in which blaming a wrongdoer is thought to be inappropriate is when the blamer has committed a similar culpable wrong, making their blame hypocritical. Philosophers describe the fact that hypocritical blame is somehow inappropriate in virtue of the hypocrite’s own wrongdoing by saying that hypocrites lack the standing to blame.

If hypocrisy undermines standing to blame, we can explain why many people are reluctant to blame the obedient subjects in Milgram’s experiment, despite believing that those subjects are culpable wrongdoers. If we can lack the standing to blame people for wrongs we have also culpably committed, it seems plausible that we might also lack the standing to blame people for wrongs we would have culpably committed if we had been in their situation. This would entail that most of us lack the standing to blame subjects who obey in Milgram’s experiment, because we likely would have obeyed too.

This argument has a number of steps which need to be unpacked and defended. First, let me say some things about the nature of this chapter and sketch out what is to come. My argument is presented as independent of those in previous chapters, in that I am assuming hypothetically that all subjects in the situationist experiments had sufficient capacity and opportunity to avoid wrongdoing, and are therefore blameworthy. I am assuming that all subjects had the capacity to recognise and act on moral reasons, but simply did not exercise these capacities, and are thus deserving of blame. I make this assumption to examine which considerations affect whether it is appropriate for us to blame wrongdoers, independent of the wrongdoer’s blameworthiness. For the purposes of this chapter, I will also be neutral regarding which account of moral responsibility is correct. My argument is compatible with numerous accounts of moral responsibility, including those accounts which reject the claim

88 Another commonly identified condition on standing to blame is that the blamer be suitably related to the victim or affected by the wrongdoing. Smith (2007) has argued that members of the moral community can lack the standing to blame wrongdoers when that wrong is relatively minor, when the wrongdoer and the would-be blamer are strangers, or when the wrongdoer has only wronged themselves. The basic idea is that we lack the standing to blame people for things which are ‘none of our business’. It is disputed whether this ‘business condition’ should really count as a necessary condition on possessing standing to blame (Bell, 2013; Todd, 2019) so I will set this debate aside. I mention these cases as they are another useful intuition pump for thinking it can be inappropriate for certain individuals to blame, even if the wrongdoer is blameworthy.

89 The next chapter will look at what it means to say that an agent lacks the standing to blame in more detail, and examine what sense of ‘inappropriate’ is at issue.
that agents must possess a certain level of control over their actions in order to be appropriate targets of blame, or that ‘appropriateness’ should be identified with basic desert.

My argument proceeds as follows. In §6.2, I identify some uncontroversial aspects of blame and hypocrisy. In §6.3, I summarise a number of recent explanations for why hypocrites lack the standing to blame, and present objections to these accounts. I then present Todd’s (2019) argument that hypocrites lack the standing to blame a wrongdoer for $\phi$-ing when they are not sufficiently committed to the kinds of values which would condemn $\phi$-ing. In §6.4, I argue that subjunctive wrongdoers who are not sufficiently committed to the kinds of values which would condemn $\phi$-ing also lack the standing to blame people who $\phi$. In §6.5–6.6, I outline why the obedient subjects in Milgram’s experiments are not sufficiently committed to the kinds of values which would condemn their wrongdoing. Combining this with the results from the many replications of Milgram’s experiments leads to the conclusion that most of us lack the standing to blame subjects who act wrongly in these settings.

### 6.2 Blame, Standing and Hypocrisy

Since we are interested in standing to blame, let me identify some features of blame which seem uncontroversial, without thereby committing to any particular account of blame. First, blame is distinct from merely judging someone to be culpable for having done wrong. It is perfectly intelligible to believe that someone did wrong, was culpable for doing so, without thereby blaming them for that culpable wrongdoing (Coates & Tognazzini, 2013b). Second, blame need not be actually expressed. One can privately blame their boss for making them work late, but choose not to express this for fear of being fired (Driver, 2016). Third, blame involves some negative component. This component could be certain reactive attitudes (Wallace, 1996), communicative actions (Smith, 2008, 2013) or dispositions towards certain responses (Sher, 2005).

Multiple philosophers have recently argued that hypocrites lack the moral standing to blame wrongdoers for certain wrongs (Coates & Tognazzini, 2013b; Friedman, 2013; Roadevin, 2018; Todd, 2019; Wallace, 2010). Hypocritical blaming is often thought to evince some kind of moral fault, even if the person being blamed is indeed culpable. If Assassin 1 and Assassin 2 both shoot the same political dignitary, Assassin 2 is entitled to judge that Assassin 1 has done something wrong, to think that Assassin 1 has a callous character, and to
inform Assassin 1 that there are strong moral reasons to feel remorse. Blame, however, is off the table. If Assassin 2 were to openly rebuke Assassin 1, Assassin 1 is within his rights to say, ‘I know what I did was wrong, but who are you to blame me?’ or ‘Look who’s talking!’ This response acknowledges that the speaker has committed a wrong and that certain responses from the moral community are appropriate, but denies that the hypocrite is in any position to blame.

As a first pass at identifying what makes an agent a hypocrite, it seems that an agent must meet at least two conditions. They must presently blame a wrongdoer for performing a wrong action (\(\phi\)), and they must have performed a similar kind of wrong action. There is an important question here regarding how similar these two actions must be in order to undermine standing to blame. Suppose that instead of shooting the dignitary, Assassin 2 kills the dignitary by blowing up their car. Since shooting is a different kind of action to detonating a bomb, if we adopt a very narrow sense of \(\phi\), Assassin 2 will have the standing to blame Assassin 1 for shooting the dignitary. Intuitively though, Assassin 2’s action is not relevantly different to Assassin 1’s action, and Assassin 2 has still lost their standing to blame Assassin 1. Similar considerations would apply were Assassin 2 to have killed someone else. If Assassin 2 kills a different but relevantly similar political dignitary, it will not be appropriate for him to blame Assassin 1. Because of this, we should adopt a slightly broader characterisation of \(\phi\).

However, adopting too broad a sense of \(\phi\) also has problems. Without some limit on which \(\phi\) qualify for hypocrisy, agents who commit nearly any wrong whatsoever will lack the standing to blame all wrongdoers. Both an unfaithful husband and a murderer are guilty of failing-in-one’s-moral-duties, but this does not undermine the husband’s standing to blame murderers, nor does this constitute what most people would consider to be hypocrisy. Because of this, our sense of hypocrisy should not be too broad either. For now, I take it we have an intuitive sense of the relevant reference class for describing most immoral actions, and which relevant moral principle an action violates. Though failing-in-one’s-moral-duties is a necessary component of Assassin 1’s wrongdoing, the relevant moral principle he violated is a prohibition on murdering innocent people. He will lack the standing to blame other people who have also violated this prohibition.
6.3 Previous Work

A number of arguments have recently been offered for the claim that hypocrites lack the standing to blame people for wrongs of which the hypocrite is also guilty. To show that we currently lack any plausible explanation for why hypocrites lack the standing to blame, let me briefly summarise the existing literature, and identify the weaknesses of each account.

Fritz and Miller (2015) argue that hypocrites have a ‘differential blaming disposition’, which implicitly rejects the impartiality of morality and thereby denies the equality of persons. Fritz and Miller take it to be a plausible assumption that our right to blame others is grounded in the equality of persons, so anyone who rejects the equality of persons thereby loses their standing to blame. While there is much to like about this argument, I reject their account because it entails that agents who inconsistently blame wrongdoers for particular wrongs lose their standing to blame everyone for similar wrongs. For example, when citizens of Western nations blame terrorists for an attack in France, those terrorists would be able to claim that this blame is inappropriate because those citizens failed to blame the terrorists for an attack in Turkey some months earlier.

Roadevin (2018) argues that hypocrites are blameworthy for failing to take seriously the demands of morality. In blaming the wrongdoer, they are demanding that the wrongdoer apologise, while culpably failing to do this themselves. Unless the hypocrite themselves apologises and self-blames, this treats the wrongdoer unfairly, and Roadevin takes for granted that any unfair practice is unjustified. I reject Roadevin’s account because her claim that blame seeks an apology from wrongdoers seems dubious, as we often continue to blame people even after they’ve apologised, or without wanting an apology at all. Additionally, her account can’t accommodate our intuition that some agents seem to lack standing to blame

---

90 Fritz and Miller try to downplay this kind of objection by noting two kinds of inconsistent blamers will retain their standing to blame: the person whose blaming disposition varies with their mood, and the person who finds themselves unable to blame one particularly charming friend. They argue that in the first case, the inconsistency does not arise from a rejection of the impartiality of morality, and in the second, this person’s disposition is masked by the wrongdoer’s charm. However, they accept that inconsistent blaming of the kind in the terrorist case rejects the equality of persons and removes standing.

91 Her account also treats the questions of whether hypocritical blame is fair, justified, appropriate, or something hypocrites are entitled to as all being synonymous, which I argue is problematic in the next chapter.
despite not being hypocrites, or despite not treating anyone unfairly (respectively shown by *Diet* and *Mistaken Blame* below).

Wallace (2010) claims that when we blame, we undertake a commitment to a principle of equality. In blaming, the hypocrite elevates their interests above the wrongdoer’s, denying the equal standing of persons that is ‘practically constitutive of morality’ in the first place, and which the hypocrite necessarily relies upon in blaming. This explanation may seem plausible at first glance, but it generates unintuitive results. Todd (2019) points out that on this account, anyone who blames without self-scrutinising is open to the same objection we make against hypocrites, even if the blamer is a moral saint.

Isserow and Klein (2017) have an interesting paradigm-based account of hypocrisy. They argue that some people are treated as moral authorities in communities, with this being a social status granted to them by others. This authority is practical rather than epistemic, in that it is *pro tanto* action guiding for others. Isserow and Klein argue that hypocrites undermine their claim to moral authority in virtue of the inconsistency between their pronouncements and their actions. One problem for the account is that Isserow and Klein don’t specify *when* agents are entitled to claim moral authority. Additionally, thinking about the way in which authority is lost reveals further conditions need to be specified. Suppose that I act wrongly, and then make a moral pronouncement. I am open to a charge of hypocrisy. According to Isserow and Klein, this inconsistency between my act and pronouncement undermines my claim to authority. But here, it can’t be the case that the *inconsistency* between my actions and pronouncement undermines my authority. Before the announcement, I hadn’t claimed any authority, and making such a claim *would* be objectionable. I had thus *already* lost any entitlement to make any claim to moral authority, prior to actually making any such claim. I lost this entitlement even if I didn’t end up blaming anyone. But this means we need some additional story about which factors make claims to authority impermissible or inappropriate prior to any claiming.92

The most intuitively plausible account of standing to blame, which I will build upon, comes from Todd (2019). Through careful consideration of a number of thought

---

92 Isserow and Klein briefly examine this question of entitlement, suggesting that blaming is an illocutionary act, and that past wrongdoing can be a “disabling fact” about an agent which makes them no longer “in a position” to criticise (pp. 198–199). Before we can determine whether the obedient subjects in Milgram’s experiment possess the standing to blame, we need to explain why certain things disable one from condemning, or remove one’s entitlement to condemn. Their account in its present form does not have the resources to answer our main question.
experiments, Todd convincingly argues that agents lack the standing to blame a wrongdoer when they are not sufficiently committed to the values that would condemn the wrongdoer’s actions. Todd has four sources of evidence for this claim. First, we often think that agents who are complicit in wrongdoing also lack the standing to blame people for that wrongdoing, even if they haven’t committed the same wrong. Consider the following thought experiment:

*Diet:* Charlie knows that Linus, who has a weakness for sweets, is trying to lose weight. Nevertheless, he takes Linus to a place for dinner that he knows is located next to an incredible ice cream shop. Quite predictably, after dinner Linus visits the shop next door and has some ice cream. Charlie blames Linus.  

Linus is to blame for violating his commitment to his diet, in that it would be appropriate for his family and friends to blame and hold him responsible for this. But even though Linus is fully responsible for his wrongdoing, Charlie is also partly responsible, and not in a position to appropriately blame Linus.

Second, agents who do commit the same action as the blamed can escape the charge of hypocrisy and retain standing to blame if they perform that action for different reasons:

*Nazis:* Jonas and Thomas are Nazi commanders in a WWII death camp. Jonas orders Thomas to investigate the fence and sound the alarm if anyone is trying to escape. Thomas investigates, sees someone escaping and sounds the alarm. However, Jonas is actually a double agent, working to sabotage the camp. He only gave that order to keep up appearances, and he picked Thomas to investigate because he thought Thomas would be the most likely to have mercy and not sound the alarm, allowing the prisoners to escape. Jonas blames Thomas.

Jonas’s blame is appropriate, and Thomas has no grounds on which to object. Jonas’s motives mean he is not the kind of agent to whom we would target with a pejorative charge of hypocrisy.

---

93 *Diet* is originally from King (2015, p. 5–6). *Nazis, Abandoned and Mistaken Blame* are from Todd (2019, pp. 354, 357, 362, respectively).
Third, agents who have lost their standing to blame can regain this standing by now being committed to the relevant values that condemn their prior wrongdoing:

*Abandoned*: Paul, when he is 25, abandons his wife and newborn infant to go touring around the world and “find himself”. His wife and child suffer terribly as a result. Later in life, at the age of 45, he comes to realize the virtues and obligations of fidelity and becomes a committed husband and father to his wife and child, with whom he is somehow reconciled. He now holds values, very deeply, which condemn his past actions. Paul hears of another 25-year-old, Peter, who has abandoned his wife and child to explore the world just as he once did, and blames him.

Even though Paul’s past actions are just as blameworthy as Peter’s present actions, and even though Paul actually lost his standing earlier (unlike Thomas), our judgements regarding standing are sensitive to how agents are presently constituted.

Fourth, people can lack the standing to blame agents they mistakenly believe exist, even though this doesn’t treat anyone unfairly:

*Mistaken Blame*: Samuel is a shop owner who has himself never tried to shoplift, but not for want of wishing to do so. Samuel simply lacks the means to try; a fact he regrets. He simply has no vehicle to travel to other stores, and there are no suitable targets within walking distance. If there were, he would try to shoplift from those stores. One day, Samuel gets a call: someone has been shoplifting from his own store. On hearing this, Samuel burns with rage and incredulity and a desire to confront the criminal. However, in fact the call has been made to the wrong number, and no-one has stolen anything from Samuel’s store.

What is important to note is that Samuel’s hypocritical blame doesn’t, in fact, wrong anyone or treat anyone unfairly. Nevertheless, he isn’t entitled to express blame in virtue of his prior wrongdoing, and third parties would point this out to him. It would be appropriate to respond with ‘Who are you to blame them?’ or ‘Look who’s talking!’

Todd’s conclusion is that agents lack the standing to blame wrongdoers when they are not sufficiently committed to the values that would condemn the wrongdoer’s actions. This account of standing to blame is an improvement on other accounts in that it is simple, able
to explain many of our intuitions across a range of cases, and doesn’t seem to face any obvious counterexamples. Note it has the interesting consequence that hypocrisy doesn’t undermine standing to blame directly. Rather, a hypocrite’s prior wrongdoing is strong evidence that they are not sufficiently committed to the values which would condemn the other person’s wrongdoing. These thought experiments allow us to see that hypocrites aren’t merely people who blame others for \( \phi \)-ing while having culpably \( \phi \)-d themselves, as I noted in my first pass at characterising hypocrisy. To be a hypocrite, one also needs to lack sufficient commitment to the kinds of values which would condemn \( \phi \)-ing.

Todd’s account is attractive in many respects, but it has some weaknesses. Todd never explicitly investigates what counts as ‘sufficient commitment’, how we can assess whether an agent is sufficiently committed, nor how wide or narrow in scope these values are. I will answer these questions in \( \S6.5–6.6 \), thereby strengthening Todd’s account. For now, I follow Todd in noting that commitment to a value consists in some endorsement of that value, and some degree of motivation to comply with that value.

A major weakness of Todd’s account is that, as he admits, he has no explanation for \textit{why} sufficient commitment to relevant values is necessary for standing to blame (p. 25). I also intend to provide this explanation, but for several reasons I am going to set this aside until the next chapter. First, my answer is quite complicated, and examining this question now would introduce too many moving conceptual parts.\(^94\) Second, my answer is not neutral regarding the nature of blame or which account of moral responsibility is correct, while my argument in this chapter is. If you agree with me that Todd has identified the relevant set of circumstances in which agents lack the standing to blame wrongdoers, you can accept my argument in this chapter, even if you disagree with my specific explanation of why these agents lack the standing to blame. Third, even if we cannot explain why lack of sufficient commitment undermines standing to blame, Todd’s account has interesting implications regarding our standing to blame subjects who act wrongly in the situationist experiments. For the remainder of this chapter, I will assume that Todd’s account is correct.

To avoid unwieldy sentences or possible ambiguity, let me stipulate that for the remainder of this chapter, when I say that an agent ‘lacks the standing to blame’, I mean that they ‘lack

\(^94\) Briefly, my answer is that since blame expresses demands, one must first possess the authority to make such demands. But in not being sufficiently committed to the relevant values, agents do not recognise other people’s authority to make demands on them, and this undermines the agent’s authority to make demands of others by blaming.
the standing to blame people for wrongs which would be condemned by values that the agent is not sufficiently committed to’. It is not the case that hypocrites lack the standing to blame any wrongdoer whatsoever. I earlier noted that it seems intuitive that someone who is unfaithful to their partner retains the standing to blame murderers and thieves, for example. This can be accounted for on Todd’s account, because the kinds of values which condemn murder are different to the kinds of values which condemn infidelity. Let me also stipulate that ‘the moral value which would condemn the wrong action or omission under consideration’ is what I will call ‘V’. Agents who have some minor endorsement of the relevant moral value, but not enough to gain standing to blame, will lack sufficient commitment to V. Since I accept Todd’s account for the purposes of this chapter, to assess whether an agent has the standing to blame, we need to assess whether they are sufficiently committed to V.

6.4 Subjunctive Wrongdoing and Moral Luck

If one believes that hypocrites often lack the standing to blame, it is then only a small step to think that agents who would count as hypocrites, but for chance, will also lack the standing to blame. Suppose that we have a wrongdoer who is culpable for φ-ing. Call agents who would have culpably φ-ed in similar circumstances, for similar reasons, subjunctive wrongdoers. Additionally, call subjunctive wrongdoers who blame an agent that in fact φ-ed subjunctive hypocrites. Almost nothing has been written on the relationship between standing to blame and what I am calling subjunctive hypocrisy, but Todd’s account can easily identify which kinds of subjunctive hypocrites lack standing to blame and explain why.

Agents who φ but blame others for φ-ing only lack the standing to blame when that agent lacks sufficient commitment to V. This entails that agents who would have φ-ed due to a lack of V, but who haven’t φ-ed simply due to not being in the relevant setting, will also lack the standing to blame. We don’t necessarily need to investigate whether an agent would have φ-ed to determine whether they lack standing to blame, as we can just independently determine whether they possess sufficient commitment to V. But the fact that an agent would φ can be strong evidence that they lack sufficient commitment to V.

At first glance, thinking about agents who would have done the wrong thing, but haven’t simply due to chance, seems to introduce questions of moral luck. Philosophers who claim that moral luck exists believe that an agent’s level of responsibility and blameworthiness can
be affected by factors beyond her control (Nelkin, 2014a). But Todd’s account actually doesn’t depend on whether moral luck of any kind affects responsibility or blameworthiness. Agents lack standing to blame when they lack sufficient commitment to V, and this is not sensitive to whether factors beyond one’s control can affect one’s level of culpability. In some cases, luck in the way an action turns out might cause an agent to become committed to the relevant values and thereby cause them to regain standing. But this is distinct from moral luck, the thesis that luck can affect one’s level of responsibility or blameworthiness.

For example, if two assassins try to murder the same person simultaneously, but a chance gust of wind causes Assassin 2’s bullet to miss, luck has affected the outcomes of their actions. This luck hasn’t had any effect on their commitment to the relevant values, however. Assassin 2 is thus not now suddenly licensed to blame Assassin 1 for committing murder.

Similarly, if two agents would act wrongly in a certain situation due to both lacking sufficient commitment to V, but only one agent gets placed in that situation and in fact acts wrongly, luck has affected whether each agent committed any wrongdoing. It is an open question as to whether luck has thereby affected these agents’ degree of blameworthiness. However, luck hasn’t affected whether each agent presently is sufficiently committed to V, and so luck hasn’t affected either agent’s standing to blame.

Luck in how agents are constituted can change one’s standing to blame. The environment that an agent is raised in, their genetics and thus their constitution later in life are all significantly affected by chance. But there is no reason to think that this unfairly affects standing to blame, and so I am happy to accept this consequence of Todd’s account. The fact that you would have done the same thing, and thus would have lacked standing, if you had another person’s genetics and upbringing says nothing about how you are actually constituted. Additionally, this line of thinking faces a well-known objection. The thought that you would have done otherwise if you had a different history and constitution simply amounts to the thought that ‘I would have done the same thing if I was a different person’, and it is hard to make sense of how exactly ‘I’ could be a different person (Nagel, 1979).

Instead, the thought that we would have done the same thing if we had been a different person is more relevant to the question of whether the other wrongdoer had a fair opportunity to avoid wrongdoing, or whether it was reasonable to expect them to do otherwise. This is a question about whether the other person is an appropriate target of blame at all, not whether it is appropriate for a blamer to blame given certain facts about the
blamer. And this is a distinct question for philosophers working on problems of constitutive moral luck or free will and determinism, which we can set aside.

6.5 Situationism, Subjunctive Hypocrisy and Commitments

If insufficient commitment to V undermines standing to blame, and if subjunctive wrongdoing can be strong evidence of insufficient commitment to V, the situationist experiments have important implications for our blaming practices and the moral community. Given Milgram’s experiment has been well-replicated in a number of cultures, across a number of decades with very similar results (Blass, 1999a; Burger, 2009; Doliński et al., 2017; Edwards et al., 1969; Shanab & Yahya, 1978), we should believe that roughly 65% of us would have continued administering shocks all the way to the end of the scale had we been subjects. We would have continued administering shocks even after the learner asked to be let out, was audibly in pain, stopped responding, and appeared to have had a heart attack.

It is important to recall that, although it is commonly reported 35% of subjects ‘disobeyed’, most of these subjects still went far beyond what is morally acceptable. Over 80% of subjects continued administering shocks after the learner demanded to be let out and complained about his heart. Had the experiment been real, 80% of subjects would have given the learner a heart attack, and failed to leave the room to check on him.

However, the fact that most of us would obey in this setting doesn’t yet show that we lack the standing to blame subjects who in fact obeyed in this setting. To show that 80% of us lack the standing to blame the obedient subjects in Milgram’s experiments, I need to show that those of us who would obey lack sufficient commitment to V. As noted earlier, there are three components of Todd’s account that we need to provide answers to: how widely or narrowly we should think of V, what counts as sufficient commitment to V, and how to determine whether a particular agent is sufficiently committed to V.

How narrow we should understand V is an important question, but I don’t think worries about scope apply to the case at hand. This is because the subjunctive wrongdoing under consideration is the exact same kind of wrongdoing as the wrongdoing that we are considering blaming the obedient subjects for performing. Whichever values condemn the wrongdoing of the obedient subjects, these will be the same values which would condemn the wrongdoing of most blamers were they to end up in a Milgram-style experiment. There may be some exceptions in which agents obey for completely different reasons, in such a way that
shows they are sufficiently committed to V. But such cases would be rare, and it seems hard for them to arise without it being the case that the subjunctive wrongdoer would be blameworthy in a different way. Suppose, for instance, that one subject would obey only because she had a deep phobia of scientists in white coats. This person’s obedience might come about in such a way that renders their wrongdoing different to the wrongdoing of subjects without this phobia. But in such a case it seems this person had a reduced capacity to disobey, and this could affect the degree of our blame, if we blame at all.

Since these studies are well-replicated in a variety of populations, and those who obey do so in very similar ways, most subjunctive wrongdoers would be blameworthy in the same way as the obedient subjects. They thus be condemned by the same kinds of values those wrongdoers would invoke when condemning the obedient subjects. This means that, if it can be shown that the obedient subjects lacked sufficient commitment to V, it will follow that the majority of us who would obey in the experiment will also lack sufficient commitment to V and thus lack the standing to blame the obedient subjects.

Our remaining questions are what counts as sufficient commitment to V, and how to determine when agents possess sufficient commitment to V. Intuitively, it could be argued that there are two good sources of evidence the obedient subjects did possess sufficient commitment to V. We generally expect agents who value something to act in ways consistent with valuing that thing across a range of domains. Someone who values friendship will generally wish their friends happy birthday and visit them when they are sick. We also expect agents who value something to experience certain affective responses indicative of a commitment to that value, such as caring that the value is adhered to and being distressed when it is not. If someone values their friendship with a person, they will care about how that person is doing, and experience appropriate affective responses in response to the other person’s well-being. The obedient subjects seem to meet both these criteria. They presumably had strong track records of not harming people, despite having the opportunity to do so. They were also clearly distressed about their actions, and would probably report having a strong commitment to not harming strangers.

This evidence is not enough to establish that the obedient subjects possessed a sufficient commitment to V. Appealing to the agents’ distress as evidence they possessed sufficient commitment to V relies on a common, but mistaken, standard by which we evaluate people’s degree of commitment. Suppose that Isabelle volunteers at the library every week instead of using her spare time to watch Netflix, and seems to experience the appropriate responses to
her patrons or hearing that the library has received extra funding. We would usually conclude that Isabelle is committed to volunteering at the library, and values this volunteer work. But this method is problematic because it only assesses how committed an agent is relative to their other preferences and options. To show what I mean, imagine we now find out Isabelle would stop helping at the library permanently if she was offered $10 to quit, or if her boss asked her to come one minute earlier from now on, or if a new show was added to Netflix. We would be inclined to think that she was in fact never very committed to helping at the library after all. Helping at the library had simply been what she preferred to do out of a narrow set of options.

Because we are assessing the obedient subjects’ commitment to moral values, we can’t ascribe sufficient commitment to those values simply because the subjects had a desire to act in a manner consistent with those values while acting immorally. The subjects’ commitment to those values might still be very weak on some absolute or objective standard. Their commitment to those values certainly seems very weak relative to the level of commitment they ought to have had.

One might think that the subjects’ histories are evidence they were committed to V on some absolute measure of commitment. But it is important to note that while subjects’ histories are good evidence that they don’t intrinsically value harming people, we are blaming the subjects for not refusing to administer shocks when directed to by an authority figure. The fact that the obedient subjects had a strong history of not harming people doesn’t show that they were strongly committed to the values which would condemn their wrong actions, as they simply hadn’t previously been in situations where there was strong social pressure to obey someone in a lab coat. If the learner had, in fact, been harmed, they would blame the subject for not being more motivated to act in accordance with their better judgement, and for not having the fortitude to stand up to the experimenter. It is very possible for someone to feel distressed about acting wrongly, while still choosing to act wrongly due to insufficient commitment to the kinds of values which would condemn that wrong action.

6.6 Diagnostic Acts

Whether the obedient subjects possessed sufficient commitment to V remains an open question. We thus need some means of evaluating when an agent has sufficient commitment to V. The strongest evidence of whether an agent is sufficiently committed to V isn’t just
that we observe a lot of behaviours consistent with V. It’s that they regularly display V-consistent behaviour in a range of V-eliciting circumstances. Likewise, a failure to display these behaviours in these circumstances is our strongest evidence that the agent lacks sufficient commitment to V. This is evidence we currently don’t have, as the obedient subjects haven’t been placed in a range of relevantly similar settings. But I don’t think we need to start speculating about how the obedient subjects would act in such settings because of the nature of the wrong that we are considering.

For many values, a single instance of severe and culpable wrongdoing can be diagnostic of the fact that the wrongdoer is not sufficiently committed to that value (Doris, 2002). While someone who eats one unhealthy meal can still be considered sufficiently committed to their diet, someone who murders once cannot (at that time) be considered sufficiently committed to valuing innocent human life. A single act of wrongful intentional murder, committed when one has full control over their behaviour, is not merely in tension with valuing human life. It is entirely incompatible with having this value, even if the agent was distressed while carrying it out and felt regret afterwards. Recall that, had the experiment been real, the obedient subjects would have been guilty of manslaughter, and 80% of subjects would have been guilty of shocking someone until he had a heart attack. They didn’t call the ambulance, and they didn’t leave the room to check on the learner. Although the subjects didn’t show any malice in their actions, they nevertheless committed severe wrongdoing.

Let me remind readers that we are setting aside my arguments from previous chapters, and assuming that subjects had the capacity and opportunity to avoid wrongdoing necessary

---

95 Badhwar (2009) puts the point nicely when assessing whether the obedient subjects possess certain virtues: “It is true that the failure to act virtuously in one situation does not, typically, reveal a lack of the relevant virtues, whether global or domain-specific. But all situations are not equal. Suppose a man murders his wife simply to inherit her property and marry his wealthy mistress. Must we wait for him to murder this new wife in order to inherit her property and marry his new mistress before we can justifiably conclude that his respect for human life is not very deep? Evidently not; the situation is “diagnostic” (to adapt Doris’s term for my own purposes) of a lack of respect for human life. The situation in the Milgram experiment is similarly diagnostic. The experiments offer a clear choice between good and evil: both common sense and the direct evidence of their senses tell the subjects that the shocks are painful as well as dangerous, and basic moral decency tells them that it is wrong to harm an innocent person against his will for a trivial purpose. Moreover, the choice facing them is an easy one, since disobedience poses no threat to their lives or happiness” (p. 263).
in order to be deserving of blame. They were aware that their actions were wrong, but they didn’t follow their better judgement, perhaps due to a fear of embarrassment or desire to comply with perceived social norms (Sabini & Silver, 2005). To be sure, the subjects were in an unusual setting, and we might think that it was difficult in some psychological sense to disobey. But when philosophers and lawyers discuss the difficulty of avoiding wrongdoing as a mitigating factor, they have in mind difficulties which directly impede the exercise of an agents’ capacities. They have in mind cases where agents are either coerced into wrongdoing, or where avoiding wrongdoing carries with it significant cost (Brink & Nelkin, 2013). The difficulty of acting against a strong desire usually isn’t taken to be mitigating, unless it can be shown that this desire somehow reduced the agent’s capacity or opportunity to avoid wrongdoing, as in cases of compulsion or addiction. These kinds of mitigating factors were not present for the obedient subjects. It is also worth emphasising again just how low of a bar subjects had to clear to qualify as having ‘disobeyed’. Subjects didn’t have to outright refuse or challenge the experimenter. They merely had to voice some kind of concern or question four times in a row. At this point, the experiment was terminated by the experimenters.

Given the severity of this wrongdoing and the low cost of disobeying, the fact that subjects had not harmed people previously and felt distressed during the experiment is not enough to ascribe to them a strong commitment to the kinds of values which would condemn their wrongdoing. I thus believe that the obedient subjects were not sufficiently committed to the values which would condemn their wrongdoing prior to and during the experiment (though they could plausibly gain V after the experiment). If they were sufficiently committed, they would have been sufficiently motivated to disobey. If an opponent believes that the subjects retained the standing to blame each other for their wrongs, they need to explain how the subjects could have kept administering the shocks, while simultaneously being sufficiently committed to the kinds of values which would

96 Proponents of Real Self and Attributionist accounts of moral responsibility argue that agents are blameworthy because their actions are attributable to them, and so may not grant that the subjects had the capacity to do otherwise. I don’t think that this is a problem, because locating responsibility in how the agent is constituted seems rather close to locating responsibility in the way the agents’ values and motivations are arranged. If the agent possessed V but still committed wrongdoing, since V entails acceptance of a value and motivation to comply with that value, Attributionists would be inclined to say that the wrongdoing therefore wasn’t attributable to the agent. This would make the obedient subjects excused, thus rejecting our starting premise.
condemn administering those shocks, and possessing the capacity and opportunity to stop administering those shocks. Such an account seems implausible.

Since the obedient subjects in these experiments are ordinary agents just like us, and since numerous replications show that most ordinary people obey the experimenter in this setting, this means that approximately 80% of us lack the standing to blame the obedient subjects. Of course, it is difficult to identify whether any particular one of us falls into this group, and whether being well-informed about situationist experiments might make us more likely to disobey. But all else being equal, it is inappropriate for approximately 80% of us to blame the obedient subjects for their wrongdoing.

6.7 Objections
One possible objection to the claim that standing to blame requires sufficient commitment to V is that it doesn’t seem like we often make the ‘Who are you to blame me?’ response to subjunctive wrongdoers. Getting exercised about subjunctive wrongdoers who blame us may happen sometimes, but it doesn’t appear to be a regular feature of our moral practices, and therefore my account seems revisionary. While I agree with this observation, my account is not revisionary as there is a good explanation for why we don’t commonly have these responses. It’s not that my account conflicts with our moral practices; rather, it’s that we are rarely in settings where we can recognise that someone lacks the standing to blame due to a lack of sufficient commitment to V. This is due to the simple fact that we are epistemically limited in a number of ways, and our primary source of evidence for people’s commitments is their behaviour.

We cannot see the future behaviour of others or ourselves, whereas we can have directly observed people committing wrongs in the past. We thus have much more evidence and greater ability to identify actual wrongdoers who lack the standing to blame, even though there exist more subjunctive wrongdoers who also lack standing to blame. It is also very difficult to assess whether we would violate a moral principle in certain circumstances, because asking and answering that question can change how we would behave in those circumstances. Reflecting on your own dispositions to wrongdoing can motivate you to avoid acting in that way, thereby changing your disposition and thus whether you are a subjunctive wrongdoer. Likewise, presently judging someone to be a subjunctive wrongdoer will often motivate them to avoid acting in the way you predict they would act. Because of
these epistemic limitations, it is very difficult to make assessments regarding who possesses sufficient commitment to V and adjust our blaming practices accordingly, even if it is true that subjunctive wrongdoers who lack sufficient commitment to V lack the standing to blame.

Trying to overcome these epistemic difficulties would come at a considerable cost. Because it is difficult to assess whether an agent is a subjunctive wrongdoer who lacks sufficient commitment to V, reaching any shared conclusion on which members of the moral community have standing to blame would often take a lot of time, or be outright impossible. Furthermore, making this a regular feature of our practices could give culpable wrongdoers a strong incentive to begin questioning the standing of blamers in an attempt to avoid being held responsible. As a result, our need to avoid perverse incentives, our epistemic limitations, and our interest in making sure that wrongdoers are held to account leads to a reasonable default assumption that members of the moral community have standing to blame. These factors have meant philosophers haven’t had much motivation to analyse the relationship between subjunctive wrongdoing and blame, even though the claim that subjunctive wrongdoers who lack sufficient commitment to V lack the standing to blame is consistent with our moral practices.

Another objection to my conclusion is that, if subjunctive wrongdoers who lack sufficient commitment to V lack standing to blame, this leads to a number of undesirable consequences. We are forced to condone the actions of wrongdoers, or at least, we aren’t entitled to complain about their wrongdoing. This arguably removes incentives against wrongdoing, and could make us unable to engage in moral address in settings where this would be effective. Additionally, this conclusion may support a world where wrongdoers, when being blamed, have license to start bringing up faults that the blamer has committed, instead of focusing on their own faults and making amends.

I don’t think any of these consequences follow from my position. First, nothing I have said prohibits other reactions to wrongdoing, or other forms of engagement with wrongdoers. One is still free to engage in moral education, or to convince wrongdoers that they’ve done something wrong and need to make reparations, for example. What is prohibited depends on what blame is, but most philosophers take blame to be a reactive attitude such as resentment (Wallace, 1996), or an expression of demands (Darwall, 2006). These are the kinds of responses my account does rule out.
It is also worth emphasising that agents can regain standing to blame by becoming the type of person who is now sufficiently committed to the relevant values. I agree that anyone who focuses on others’ faults instead of their own when being blamed displays a serious moral fault. I do not intend to promote such a mentality. Instead, my aim is to encourage would-be blamers to take a step back and examine themselves in order to have a greater appreciation of what kind of situation the wrongdoer was in. Such reflection ensures our blame stems from a commitment to moral principles, rather than being a knee-jerk reaction stemming from problematic motivations such as moral grandstanding (Tosi & Warmke, 2016).

Recognising our own subjunctive shortcomings also has benefits. When we are in a position to blame others, it is easy to think that recognising an act is wrong should be enough to motivate people to avoid committing it. But when we realise that we would have done the same, we can appreciate that wrongness alone is not sufficient to motivate people to act well and so other solutions are needed. This makes us much more likely to begin thinking of more effective ways to prevent those settings from developing altogether. We are able to use our understanding of what would cause us to become wrongdoers to design systems with checks and balances, and look more broadly for preventative measures. Instead of just telling people in positions of power (e.g. experimenters) to be ethical, we can design environments in ways that help foster moral behaviour (Russell, 2009).

6.8 Generalisability

A relevant question here concerns whether we lack the standing to blame subjects who act wrongly in other situationist experiments. Since there are hundreds of situationist experiments, and each identifies a different factor which can increase our chances of wrongdoing, this might make it seem like we lack the standing to blame far more wrongdoers than is ordinarily assumed.

Most of the situationist experiments do not directly show we lack the standing to blame subjects who act wrongly in these settings. This is because the wrongs under consideration in many experiments are very minor, such as failing to help pick up some papers (Isen & Levin, 1972) or lying about one’s score on a test (Gino, Norton, & Ariely, 2010). While subjects’ behaviour falls below the standard we think it ought to have been, it is not grossly lower. These single instances of behaviour which are inconsistent with a moral value are not enough
to demonstrate those agents are not sufficiently committed to that value. Subjects (and we) may in fact lack the relevant values which would condemn these wrongs, but these situationist experiments are not proof of this.

It is harder to assess the commitments of subjects who acted wrongly in other situationist experiments featuring strong social pressures. These experiments show subjects failing to help an unconscious person (Darley & Batson, 1973) or failing to call for help after hearing someone have a seizure (Latané & Darley, 1970). Though these actions are less wrong than those of Milgram’s obedient subjects, they still seem like significant wrongs, and the kind of thing for which strong blame is usually appropriate. I am inclined to say that we do not have enough evidence to settle the matter. We can only say that, if we would fail to help in these settings, and if that failure to help is attributable to insufficient commitment to V, then we would lack the standing to blame those subjects who, in fact, did not help in these settings.

My indeterminate assessment of subject behaviour in these experiments may seem unsatisfying, but our ability to assess people’s degree of commitment to moral values inherently limited by the kinds of evidence we have available. Assessing people’s degree of commitment to certain values is difficult because most people simply don’t find themselves in the kinds of settings which act as the clearest evidence. Even when people are in those settings and we can observe how they in fact act, we don’t know how those people would act in relevant counterfactual settings. And trying to assess how people would act in these counterfactual situations via introspection is likely to be unreliable or self-serving. One of the key lessons from the situationist experiments is that we are generally very bad at predicting how we will act in certain situations, and at identifying what sorts of factors influence our behaviour.

One final point on the generalisability of these studies is worth noting. The situationist experiments suggest another way in which fewer of us may the standing to blame than is commonly assumed. This comes from the literature on the consistency of people’s behaviour across normatively similar but otherwise distinct situations. For example, while we usually think that honest people won’t cheat on exams, steal candy bars or lie to their friends, Hartshorne et al. (1929) notably found that any observed instance of honesty in one setting is a very poor predictor of honesty in other settings ($r = .23$). Mischel and Peake (1982) and Chaplin and Goldberg (1984) also found very low correlations when assessing people’s behaviour across a range of situations. Though they weren’t specifically interested in the kinds of behaviour which enable us to discern an agent’s commitment to certain values, this
is nonetheless worrying because we expect agents who are committed to certain values to consistently act in accordance with those values across a range of relevant circumstances.97 This lack of consistency matters, because agents who successfully avoid severe wrongdoing in one setting, and to whom we are inclined to attribute a sufficient commitment to V, may commit a similar and severe wrongdoing in other normatively similar circumstances.98

We are now in a position to see the importance of my first argument from Chapter 2. Recall that in this chapter I showed that numerous situationist factors contribute to people’s willingness to obey in Milgram-like settings, and many of these factors were not present in Milgram’s original experiment. I argued that, by making slight modifications to Milgram’s experimental set-up (e.g. adding a second experimenter and other confederate teachers), it would be relatively easy to design an experiment in which over 95% of people would obey. A consequence of this is that nearly all of the 35% of us who would disobey in Milgram’s original experiment would obey in this relevantly similar experiment. Nearly all of us are thus subjunctive wrongdoers with regard to some Obedience-like settings. To be sure, there are some virtuous and resilient individuals who would consistently disobey in these kinds of settings for the right reasons, and these individuals would retain their standing to blame. But the number of people in this category is likely to be very small. The question now is whether this means those of us who would obey lack the standing to blame. I believe that we do.

Agents who would disobey in Milgram’s original experiment, but who would obey in a modified experiment, possess standing to blame only if they possess sufficient commitment to the kinds of values which would condemn obeying experimenters and electrocuting a stranger. I don’t think that it can be shown that these agents possess standing to blame, while also maintaining the obedient subjects in Milgram’s actual experiment lack standing to blame. The hypothetical changes to the experiment’s design would not impose additional costs on disobeying, coerce the subjects, or reduce their capacity or opportunity to disobey. The fact that an agent would obey in this modified setting undermines their standing to blame other

97 This is similar to Doris’s (2002) point that we expect agents with global traits to manifest trait-relevant behaviour in a broad range of trait-relevant eliciting settings.

98 Doris (2002) acknowledge this point: “Nevertheless, each subject was observed only in a single trial. Damn the obedients and hail the defiants if you will; the experiment does not motivate confidence about how particular subjects would behave in [other trait-relevant] situations. There’s little reason for confidence that the disobedient subjects, however inspiring their behavior in the experiment could be counted on to exhibit Socratic self-mastery in other situations” (p. 49).
obedient subjects, even if that agent would have successfully disobeyed in the original experiment. And since the overwhelming majority of us would obey in this modified setting, this is very strong evidence that nearly all of us are not sufficiently committed to V.

To see how this conclusion plays out, recall from Chapter 3 my discussion of Adrian Dimow, who was a subject that disobeyed early in the original experiment. He disobeyed not out of concern for the learner, but because he thought he was being duped by the experimenters. It seems likely that Dimow would have obeyed a more convincing authority figure in some other relevantly similar setting. And since this obedience would result in Dimow committing severe wrongdoing, we would be inclined to say he, in fact, was never sufficiently committed to V, even though he successfully disobeyed in the original experiment. He thus lacks the standing to blame subjects who obeyed in the original experiment. Blaming those subjects would be analogous to Assassin 2 rejecting a contract, blaming Assassin 1 for taking that contract, but being constituted such that he would have accepted the contract if he’d been offered $1 more. I am arguing that many of us who might successfully obey in Milgram’s original experiment are like Dimow. Disobeying in Milgram’s original experiment is necessary, but by no means sufficient, for possessing the standing to blame subjects who obeyed, and the fact nearly all of us would obey in relevantly similar scenarios is evidence we in fact are not sufficiently committed to V.

I think it is worth reiterating my earlier comments regarding our epistemic limits and competing considerations. While I maintain that we lack the standing to blame people for wrongs when we are not sufficiently committed to the kinds of values which would condemn those wrongs, I recognise that it is usually very difficult to discern other people’s level of commitment to any particular value. There are numerous competing considerations which favour a default presumption that most people possess standing to blame. But in cases where we do have strong evidence that an agent’s wrongdoing was attributable to their lacking sufficient commitment to V, we should think that agent lacks the standing to blame. And when we also have strong evidence that we would have behaved in the same way as those wrongdoers had we been in their situation, we should think that we too lack the standing to blame those wrongdoers. Milgram’s obedience to authority experiment is a rare case which meets this criteria, but an important one.
6.9 Conclusion

In this chapter, I argued that nearly all of us in the moral community lack the standing to blame subjects who obey in Milgram’s experiment, even if the obedient subjects are blameworthy for their actions. First, I showed that existing explanations for why hypocrites lack the standing to blame all face strong objections. I endorsed Todd’s recent account of hypocrisy, which coheres with our intuitions across a range of cases. A consequence of Todd’s account was that hypocrisy doesn’t undermine standing to blame directly. Rather, hypocrisy is good evidence that an agent is not sufficiently committed to the kinds of values which condemn the wrong act they are blaming someone for performing. Subjunctive wrongdoers who lack sufficient commitment to the relevant values will also lack standing to blame. I then argued that the fact that 80% of subjects in Milgram’s experiments committed significant wrongdoing, and the fact these studies are very well-replicated, is good evidence 80% of us would also have been obedient in that setting. This is strong evidence that most of us lack sufficient commitment to the relevant values which would condemn the subjects’ wrongdoing. This is because our subjunctive wrongdoing would have been very severe, would have been carried out when there was a low cost to disobeying, and because we would have possessed the capacity and opportunity to disobey.

I then considered the objection that my conclusion seems inconsistent with our moral practices, and leads to a number of undesirable consequences. I argued that my conclusion is, in fact, consistent with our moral practices, but we rarely notice when someone is a subjunctive hypocrite due to our epistemic limitations. Nothing in my argument prevents us from engaging with wrongdoers to try to get them to apologise or reform, and that recognising our lack of standing to blame could have a number of positive consequences. I then reintroduced my argument from Chapter 2, where I claimed that it would be easy to design Obedience-like settings in which over 95% of us would obey. Combining this with my argument in this chapter, I argued that, in fact, nearly all of us lack the standing to blame the obedient subjects, because nearly all of us are not sufficiently committed to the relevant moral values.
Chapter 7: Standing to Blame and Normative Powers

7.1 Introduction

In the previous chapter, I argued that most of us lack the standing to blame the obedient subjects, even if those subjects are, in fact, blameworthy. By using Todd’s account of standing to blame, I argued that both the obedient subjects and most of us lack sufficient commitment to the kinds of values which would condemn their wrongdoing. However, in §6.3 I set aside an important question which was unanswered by Todd. This was why sufficient commitment to relevant values is necessary for standing to blame. I noted that one could accept my argument that we lack the standing to blame the obedient subjects without having any answer to this question, if one agrees Todd has identified the relevant set of cases in which agents lack the standing to blame. However, my argument would be much stronger if we had such an explanation for why agents lack standing to blame in these cases, and it is this task I take up in this chapter.

This chapter proceeds as follows. In §7.2, I examine previous analyses of standing that have been made by philosophers, and note that they seem to centre around the idea of ‘entitlement’. In §7.3, I argue that thinking of standing to blame as a purely moral entitlement faces problems. By examining how the concept of standing is used in other contexts, I argue that we should think of standing to blame in partly metaphysical terms. That is, we should think of it as a status which grants agents the *ability* to do certain things. Using Darwall’s (2006) account of second-personal obligations, in §7.4, I argue that we should think of blame as expressing demands. For these demands to successfully impose obligations on others, however, we must first have the authority to make these demands. This requires that our demands be made from the perspective of a free and rational agent among equals. In §7.5, I show that agents who lack standing to blame lack the authority to blame, and thus the ability to impose second-personal obligations on others through blaming. Because they fail to accept other people’s second-personal authority to make certain demands on them, expecting others to treat their own demands as legitimate entails that they expect others to treat them as if they have an authority that others lack. Such a commitment violates the conditions required for second-personal demands to be successful.
7.2 What is Standing, Anyway?

To get clearer about different senses of ‘standing’, let’s briefly look at how the term has been used in a range of contexts.

‘Standing’ was originally understood as a legal notion, referring to a status granted to certain citizens. Citizens who wish to bring a lawsuit to court must have the ‘standing’ to do so, which means that they (or someone they are representing) have been (or are at risk of being) negatively affected by the events in the case (*Lujan vs. Defenders of Wildlife*, 1992). In short, one has standing if one is entitled to bring a case to court, with this entitlement being granted by the legal system. One retains this standing even if they choose not to bring the case to court.

This language was then picked up in the forgiveness literature, in discussions over who is entitled to forgive. A key question for philosophers in this domain concerns under what conditions forgiveness has actually taken place. Many philosophers argue that only victims have the standing to forgive (Murphy & Hampton, 1990; Zaragoza, 2012; cf. Pettigrove, 2009; Radzik, 2010). These philosophers argue that if I am not the victim (or at least closely related), then even if I forswear my resentment and declare that I have forgiven the perpetrator, forgiveness will not have taken place. Forgiveness is simply not something I am able to do. Notice that in each of these contexts, the conception of standing at issue is *metaphysical* standing. Agents are entitled to forgive or sue, in the sense that they are able to forgive or bring a case to court if they so choose. They have been granted an ability to do something in virtue of meeting certain conditions. It seems that this sense of standing can’t be what we are investigating when examining hypocrisy, since hypocrites clearly can blame. Indeed, their blaming is precisely what we object to.99

The language of entitlements then gets used in the moral domain, and this is where some lack of clarity risks occurring. When φ-ing is morally impermissible, we say that agents are not entitled to φ, even though they are physically or metaphysically capable of φ-ing. To say φ-ing is morally impermissible is to say that an agent is not entitled to φ by morality, or that they are unable to φ in such a way that conforms with the requirements of morality.

---

99 Cohen (2006) claims that hypocrites literally cannot perform the speech act of condemning, but doesn’t offer any sustained argument for this, and his claim has not received support from other philosophers. I argue below in §7.3 that we shouldn’t dismiss thinking of standing to blame in this way.
The language of moral entitlement is used by numerous philosophers when they briefly reference standing to blame. For example, Coates and Tognazzini (2012, 2013b) describe standing as concerning whether a wrong is within the blamer’s “jurisdiction” (2012, p. 203) or whether certain “propriety conditions” obtain. Friedman (2013) identifies the issue as whether the blamer is “entitled” to have their blame taken seriously by the moral community (p. 17). King (2015) writes that standing “enables one to blame with license or justification” and without it, blame is “morally problematic” (p. 1–2). Scanlon (2008) describes hypocrites as being “not in a position” to blame (p. 137). However, none of these philosophers specify how to understand these new terms in sufficient detail, and Smith (2007) doesn’t use any other descriptors to explain what standing is at all.

Thinking of standing to blame as an entitlement to blame seems very intuitive. But this sort of terminology comes with a risk. Since we’re already not entitled to act in morally impermissible ways, thinking of standing to blame as an entitlement risks not doing any explanatory work when we try to explain why hypocrites lack the standing to blame. If we ask, ‘Is it permissible for Shaun to blame Katia?’ the answer, ‘Shaun’s blame would be morally impermissible because he is a hypocrite, and therefore lacks standing to blame’ may be equivalent to ‘Shaun’s blame would be morally impermissible because he is a hypocrite, and his blame is therefore morally impermissible’. We thus need some further story about how saying that an agent lacks standing to blame is different to saying that it is simply all-things-considered morally impermissible for that agent to blame. Without such a story, we’d lack the standing to blame whenever blaming would result in terrible consequences overall. Additionally, this would make hypocritical blame philosophically uninteresting. There would be no reason for philosophers to investigate hypocritical blame, as opposed to impermissible blame generally.

If we want our conception of standing to do some explanatory work, our conception of this lack of ‘entitlement’ to blame must refer to something other than all-things-considered moral impermissibility. An intuitively appealing answer is to think about appropriateness. Questions over whether an agent has standing qua entitlement to blame would thus refer to questions about whether an agent’s blame is appropriate. In favour of this approach, we could point to the fact that it does seem apt to describe hypocritical blame as inappropriate, and that following Wallace (1996), much of the philosophical literature on moral responsibility has centred around when blame and the reactive attitudes are appropriate.
Thinking of standingless blame as a kind of inappropriate blame is initially promising, in that this would mean we’d successfully distinguished it from blame which is all-things-considered morally impermissible. Almost all philosophers writing on moral responsibility accept that whether we should blame someone, all-things-considered, is a different question to whether blame is appropriate. In particular, consequentialist considerations can make blaming someone wrong, even if that person is an appropriate target of blame. If blaming someone innocent was the only way avoid some terrible outcome, then you ought to blame them. Blame here would be inappropriate, but all-things-considered justified. Conversely, if someone culpably commits a minor wrong, but blaming them would make them suicidal, then blame would be appropriate, but not all-things-considered justified.

However, another problem arises here. In investigating what it means for blame to be appropriate, philosophers have generally focused on the conditions under which someone is an appropriate target of blame in virtue of their wrongdoing. Here, then, is a challenge to the proposal that questions about standing to blame concern whether it is appropriate for an agent to blame: what reason do we have to think that blaming when one lacks standing is inappropriate in the same way that blaming the innocent is inappropriate? What relevant feature do these differing cases share which makes blame inappropriate in both instances? The fact that an agent lacks standing to blame doesn’t seem like the general kinds of consequentialist considerations identified above which can make blame unjustified. But we can’t conclude from this fact alone that blaming when one lacks standing is therefore inappropriate in the way that blaming the innocent is inappropriate.

Answering this question thus requires that we get a bit clearer about what we mean by ‘appropriate’. Let me consider a number of ways that the term is used within the current moral responsibility literature, to show that none clearly lend themselves to an illuminating account of the sense in which hypocrites lack an entitlement to blame. To say ‘S is an appropriate target of blame’ could mean:

1. It would be fitting to blame S (Bell, 2013).
2. It is intelligible to demand S give a justification for their actions (Smith, 2015).
3. S’s actions are protest-worthy (Hieronymi, 2004; Talbert, 2012).

100 With the exception of response-dependent accounts of moral responsibility, which take whether an agent is an appropriate target of blame to be partially constituted by our actual practices of holding each other responsible (e.g. Shoemaker, 2017; Strawson, 1962).
We have a *pro tanto* reason to blame S.\(^{101}\)

S is a deserving target of blame (Levy, 2005; Nelkin, 2016a; Pereboom, 2014).

It would be fair to blame S (Wallace, 1996).

Consider (1). That a response is fitting is commonly understood to mean that the response correctly represents its target (D’Arms & Jacobson, 2000). If we interpret appropriate blame as simply being fitting blame, then blame is appropriate if its target is, in fact, a culpable wrongdoer. This account doesn’t generate any obvious explanation for why hypocritical blame is inappropriate, since hypocritical blame often targets culpable wrongdoers.

Neither (2) nor (3) seem to lend themselves to an explanation for how or why hypocrites lack standing to blame. It certainly seems intelligible for hypocrites to demand justification from culpable wrongdoers, even if the hypocrite has previously committed similar wrongs. It also seems reasonable for hypocrites to protest against wrongs similar to those that they have committed. If I agree using fossil fuels is wrong, but drive a gas-guzzler, why can’t I walk in protests for stronger policies on climate change? ‘Because that’s hypocritical’ seems to be the explanation we reach for, but that would make our reasoning circular.

Under (4), agents who lack the standing to blame lack a *pro tanto* reason to blame. This seems conceptually possible, but it isn’t immediately clear why hypocrites especially should lack or lose such a reason in virtue of their earlier wrongdoing. If you are in fact a wrongdoer, and this ordinarily gives me a *pro tanto* moral reason to blame you, why do facts about me or my previous wrongdoing undermine or remove this reason altogether?

As I argued in Chapter 1, I endorse (5) as an account of what it means to say an agent is an appropriate target of blame. That is, I believe that S is an appropriate target of blame only if S is deserving of blame. However, thinking of appropriateness in this way doesn’t seem to

\(^{101}\) McKenna (2012) endorses a variant of this, arguing that to say an agent is blameworthy is to say it is *pro tanto* permissible to blame that agent. He argues such an account can be arrived at by appealing to either desert considerations (both basic and non-basic), axiological considerations about the good, or contractualist principles from Lenman (2006) and Scanlon (1998). While I’m not saying his account can’t be modified to explain why hypocrites lack the standing to blame, it doesn’t obviously lend itself to a satisfying explanation for why hypocrites lose this *pro tanto* reason to engage in moral address, since it could be reasonable to agree under contractualist principles that hypocritical blame remains appropriate. Additionally, ordinary violations of contractualist principles don’t seem to generate our distinctive responses to hypocritical blame, which requires explanation. I’m thus inclined to set (4) aside until a more compelling defence of it is found.
give us any explanation for why hypocritical blame is objectionable, since hypocritical blame still gives wrongdoers what they deserve. One could perhaps argue that certain agents are not entitled to mete out deserved responses, by making an analogy with how vigilantes are not entitled to punish lawbreakers. Such an account may be possible, but there are relevant differences between the legal and moral cases. Judges, police and lawmakers are granted their entitlements in virtue of having gone through certain processes and conventions, and upholding the law serves a number of different functions to morality. In contrast, blaming wrongdoers is something that agents in the moral community seem to have a default entitlement to.

Finally, consider (6), which holds that appropriate blame is fair blame. It does seem like hypocrites treat the people they blame unfairly. But there are a number of ways to unfairly blame, without this undermining one’s standing to blame. Suppose I blame you, but fail to blame another colleague for a similar wrong. I treat you unfairly in blaming you, but I haven’t thereby lost my entitlement to blame you altogether. This account thus doesn’t seem capable of explaining why hypocrites lack the standing to blame. One might try to avoid this counterexample by stipulating they understand the ‘fairness’ at issue to be in a non-interpersonal sense, but this would leave us with an account which seems identical to (5).

Given these failed attempts, one might be motivated to interpret ‘appropriateness’ as concerning the ethics of expressing blame, after it has already been determined that an agent is blameworthy. Blame often seems ‘inappropriate’ if it is disproportionately strong compared to the wrong committed, expressed when one lacks sufficient evidence of wrongdoing, applied inconsistently (Friedman, 2013), targets a minor wrong that is ‘none of your business’ (Smith, 2007), or expressed when the wrongdoer has suffered enough already. Numerous factors seem to make expressing blaming inappropriate in some sense, and it could be that hypocritical blame is another factor on this list.

There are a few reasons against trying to explain why hypocrisy undermines standing to blame by thinking about the general ethics of expressing blame. First, this approach doesn’t by itself answer why hypocritical blame is inappropriate. We still need some further story, and it isn’t clear we can find one by looking for some shared property with these other conditions that putatively undermine standing. This is because the inappropriateness of blame in these other conditions seems to be explained by more general pro tanto obligations we have to each other. We generally ought to respect people’s privacy, treat people fairly, and take reasonable precautions against imposing costs on the undeserving.
Second, this approach makes it puzzling why existing philosophical work on standing to blame focuses so much on hypocrisy, and why these other putative means of losing standing to blame aren’t receiving more attention. Third, and relatedly, our moral practices treat hypocrisy as far more significant than other ways in which people blame objectionably. We have religious and cultural admonitions against ‘throwing stones in glass houses’, ‘casting the first stone’, and ‘ignoring the mote in one’s eye while pointing out the speck in others’ eyes’ (Marvin, 1922). People who blame without sufficient evidence display a moral fault, but only hypocrites were given a special place in Dante’s eighth circle of hell (Alighieri, 2003).

Finally, hypocritical blame gives its targets a distinct complaint that other kinds of objectionable blame do not. If Katia accuses Shaun of cheating when she lacks sufficient evidence of this, perhaps she is open to a charge of having given insufficient consideration to Shaun’s interests, since her evidence suggests that she will be blaming someone innocent. But Shaun gains an additional kind of complaint against Katia’s blame if Katia has also been cheating. Blaming when one lacks sufficient evidence is a fault, but it is a very different fault to that displayed by hypocrites, which we react to in a particular way. As noted earlier, philosophers investigating hypocrisy take the ‘Who are you to blame me?’ and ‘Look who’s talking!’ responses to be central to our inquiry (Bell, 2013; Friedman, 2013; Fritz & Miller, 2015; Todd, 2019).

Todd (2019) argues that there is a difference between these responses and our responses to other kinds of ‘inappropriate’ blame by carefully considering some thought experiments involving blame:

Suppose Ian has two friends: Julius and Keegan. We stipulate that this is a case in which Ian would be within his rights to tell either of them to mind their own business. But suppose that, unlike Julius, Keegan himself displays the same objectionable tendencies in his intimate relationships as does Ian. Now, to Julius, Ian may say:

Mind your own business.

But to Keegan, Ian may say:

Mind your own business … and, who are you to blame me for this anyway?

You are in no position to blame me.

And these are importantly different replies … in this sort of case, whereas both Julius and Keegan lack a certain kind of standing with Ian to intervene in his affairs,
Keegan additionally lacks something further – and it is this “something further” that is at issue in this paper.

Todd (2019, p. 349)

I believe there is something particularly inappropriate about hypocritical blame, and Todd is right that there is “something further” to be identified about Keegan that we don’t need to identify about Ian. Unfortunately, getting clearer about this ‘something further’ by referring to existing conceptions of entitlement, appropriateness, fairness, permissibility or desert, or by thinking about the ethics of expressing blame, doesn’t seem fruitful.

Let me summarise my complaints with existing attempting to explain why hypocrites lack the standing to blame, and the puzzle we currently face. Although agents who lack the standing to blame seem to lack an entitlement to blame, they seem to lack this entitlement in a way that is different both to the way that agents lack an entitlement to blame the innocent, and the way they lack an entitlement to blame when consequentialist considerations make blaming a culpable wrongdoer all-things-considered impermissible. If to say that an agent lacks the standing to blame simply means that the agent’s blame is all-things-considered impermissible, then our concept of ‘standing’ doesn’t do any explanatory work, and this approach does not accurately represent the nature of our responses to hypocritical blamers who lack standing. If to say that that an agent lacks the standing to blame simply means that their blame is inappropriate, then we need an argument for why this is so, and none seem forthcoming from existing conceptions of inappropriateness. If to say that an agent lacks the standing to blame means that their blame is objectionable in virtue of some other unique property we are yet to consider, without any further story, this answer is ad hoc.102 And if none of these options are satisfying, then we might have reason to think that blaming without standing is, in fact, not objectionable at all. This would mean that our ‘Who are you to blame me?’ responses aren’t warranted, and that there is nothing philosophically interesting about hypocritical blame worth investigating.

102 Fritz and Miller (2015) do succeed in providing an answer to my challenge by taking this option. In arguing that standing to blame is an alienable right grounded in one’s acceptance of the equality of persons, they explain how standingless blame differs to blaming the innocent, ensuring that our concept of standing does some explanatory work without inventing a third ad hoc category. As noted earlier, I reject their account because of Todd’s terrorist counterexample. One might try to retain thinking of standing to blame as an alienable right, but without some story of what this right amounts to, ‘right’ just looks like a synonym for ‘entitlement’.
7.3 Standing and Normative Powers

Trying to explain why hypocrite lack the standing to blame by thinking about the nature of standing has been unsuccessful. Because of this, I believe we should take a different approach. I think that the lack of a solution thus far stems from the fact that all participants in this debate are thinking of standing to blame \textit{qua} entitlement to blame in exclusively moral terms. Recall that it seemed like standing to blame couldn’t be thought of similarly to how we use the term ‘standing’ in other domains, because those domains concerned an ability (e.g. to sue or forgive), and hypocrites clearly have the ability to blame. I believe that this reasoning should be reconsidered. We should think of standing to blame in the same way that we think of other kinds of standing: as a status which gives agents certain abilities.

My proposed answer comes from noticing two things from the previous section regarding standing. First, notice that talk about metaphysical standing readily lends itself to discussions regarding other kinds of actions, namely consenting, ordering and promising. It sounds intelligible to say that I lack the standing to promise or consent on behalf of most other agents. A notable exception is parents, who are able to consent on behalf of their children, having this entitlement in virtue of their relationship and because we typically take parents to be best placed to act in their children’s interests (Buchanan & Brock, 1989). Parents have the standing to consent to their child being treated in certain ways (within limits) and strangers lack such standing. Parents are also able to order and impose duties on their children. Again, notice that the standing here is an ability that agents can exercise. Strangers simply are not able to consent to having someone else’s child receive invasive surgical treatment if they are not suitably related to that child.

The second thing worth noting is that these actions, which lend themselves to talk of metaphysical standing, are all \textit{normative powers}. Pardoning, consenting, ordering, promising, forgiving, sentencing and apologising are all things we do which can change others’ normative situation in particular ways. When one has been pardoned, one no longer has any obligation to atone, and others no longer have the right to punish. When one consents, one waives or removes certain duties other agents have. When one orders, one imposes obligations on others. When one promises, one imposes obligations on one’s self, and makes it appropriate for others to blame them for violating those obligations. Some philosophers argue that when one forgives, one releases wrongdoers from certain kinds of obligations, and
make certain attitudes towards them impermissible (Nelkin, 2013b; Warmke, 2015). When a judge sentences, they make certain forms of institutional treatment legal. When you apologise, you can give the victim reasons to consider forgiving, or to at least stop blaming. Each of these actions are normative powers that agents can possess or acquire. The stranger who can’t consent on my behalf is simply not able to remove other people’s duties not to treat me in certain ways.

This neat and unified picture of standing across a number of domains suggests that the standing at issue regarding hypocrisy could also concern normative powers. To say that hypocrites lack the ‘standing to blame’ would thus be slightly misleading. Though hypocrites are able to blame, they cannot exercise the normative power that agents ordinarily exercise in blaming.

What might this normative power be? A common refrain amongst philosophers is that blame expresses demands, and that in blaming we are holding wrongdoers answerable to us (Watson, 2004). This fits nicely with my argument thus far, because demands are things which require authority to be made, and authority is a status that can be gained or lost. I believe this is what previous attempts to explain why hypocrites lack standing have been circling around. Identifying the conditions under which an agent lacks the standing to blame is a task of identifying the conditions under which an agent can change others’ moral situation through the issuing of legitimate demands. Philosophers interested in identifying why hypocrites lack this standing should try to identify how hypocrites lose their moral authority. This is a task which has not been taken up in either the literature on moral responsibility, or the literature on hypocrisy.

Given I have previously argued that blame should be thought of as deserved sanction, it might seem illegitimate for me to claim that hypocrites lack the standing to blame in virtue of lacking the authority to make demands in blaming. However, numerous philosophers argue that thinking of blame as a form of desert is compatible with thinking of blame as expressing demands, or as the means by which we hold someone answerable or accountable, though they differ on the details (McKenna, 2012; Shoemaker, 2015). More importantly, Nelkin (2016a), whose account I took as my starting pointing in this thesis, argues that agents who

103 Though this is contentious; Allais (2008), Garrad and McNaughton (2011), Pettigrove (2012) and Russell (2016) all argue one can forgive while continuing to punish a wrongdoer.
104 Though on Owens’s (2012) account forgiveness isn’t strictly a ‘power’ because it may not occur through will alone, he accepts agents have the capacity to forgive and thereby change others’ normative situation (p. 50–56). This is all that is needed for my account.
are deserving of blame and agents who are appropriate targets of moral demands occupy “shared satisfaction conditions” (p. 184) for being targeted with the reactive attitudes. This is even though these two kinds of conditions initially appear to be different, and are generally defended by distinct groups of philosophers. Thinking of blame as expressing demands thus should not be thought of as abandoning the claim that blameworthy agents deserve blame.

In the next section, I will use Darwall’s (2006) account of second-personal authority to explain why hypocrites lack the authority to make the demands that blame addresses.

7.4 Second-Personal Authority
To understand Darwall’s account of authority, we first need to understand his broader account of second-personal reasons and obligations. Darwall argues that there are two different kinds of moral obligations or moral reasons for action. Suppose that you are standing on my foot, hurting me. I could try to give you a reason to remove your foot by getting you to feel sympathy, to see my being in pain as bad. But importantly, on this approach, I wouldn’t be so much addressing the reason to you as merely pointing out that it already existed. The reason I give you by making this kind of address is not primarily about you qua the agent causing me pain. It would only be a general sort of reason, which could be pointed out to anyone in a position to reduce the amount of pain in the world by anyone aware of that pain (2006, p. 6).

The second way that I can get you to take your foot off mine is by addressing to you a (purportedly valid) demand. I can demand this as the person whose foot you are stepping on, or as a member of the moral community, and this demand would concern your relations to me qua the person you are hurting. In making this demand, I am giving you a second-personal reason, a reason “whose validity depends on presupposed authority and accountability relations between persons and, therefore, on the possibility of the reason’s being addressed person-to-person” (p. 8). According to Darwall, these second-personal reasons are at the heart of our moral responsibility practices. When we blame people, we blame them by making second-personal demands of them. Additionally, since we blame people for failing to treat us in certain ways we have a right to expect, the obligations that wrongdoers violate are also second-personal in nature. They are obligations to us (pp. 69–72).

In directing agents to comply with our second-personal demands, we want them to act for certain reasons. We don’t want them to merely act out of fear of punishment. Instead, we
want the agent to recognise the legitimacy of the demand, and direct their will to comply with that demand because they respect our demand is legitimate. Respecting these demands is the way in which the agent treats us as one of Rawls’s (1985) “self-originating sources of valid claims” (p. 242). My status as a being with dignity is what grounds my authority to make demands of you to treat me in certain ways. And in demanding that you treat me a certain way, I am demanding that you do so precisely because I am a being with dignity.

Darwall argues that having the authority to make moral demands is a ‘normative felicity condition’ on such demands being legitimate, analogous to the felicity conditions that Austin (1962) argued a speech act needs to succeed. Normative felicity conditions must be met in order for the demands to be successful; that is, for those demands to in fact create reasons or obligations for addressee of those demands. That the blamer has the authority to make these demands is the main condition needed for such demands to be successful. Note that this authority is not a de facto authority granted by convention, but a de jure authority granted by morality. If I demand that you, a stranger, give me ten push ups for instance, then my demand is infelicitous because I do not have the authority to make such a demand. If I demand for you to take your foot off mine, my demand is successful because I do possess such an authority and so you have no right to object to my demand.

Additional normative felicity conditions also exist. A key insight from Darwall is that when we hold someone accountable by making second-personal demands, we presuppose and are thereby committed to a number of things. First, we presuppose that the other person is free and rational – they have the capacity to recognise and comply with such demands qua demands. We presuppose that they have the capacity to direct their will in the required way, and to hold themselves responsible for wrongdoing through self-blame. In blaming, we also presuppose that these demands are the kinds of demands that we have the authority to make, because we are beings with dignity. Note that we make these presuppositions, and are committed to them being the case, even if we never actually have explicit thoughts like ‘this person can direct their will in response to my demands’.

These presuppositions are not merely things the addressee is committed to. They are normative felicity conditions which need to be met in order for demands to be genuinely second-personal and in fact create reasons for the addressee. One simply cannot make a second-personal demand to someone if they do not believe that person is capable of understanding and complying with that demand. One also cannot make this kind of demand
if they are aiming to coerce the other person, or if one does not see themselves as a source of self-originating claims.

However, in presupposing that the other person has the capacity to direct their will in recognition of my second-personal demands, I have thereby presupposed that they are a kind of being with dignity too. I presuppose that they are a source of self-originating claims which I need to respect. Second-personal claims presuppose “a common competence, authority, and, therefore, responsibility as free and rational, a mutual second-personality that addressee shares and that is appropriately recognized reciprocally” (p. 21).

In presupposing that both of us are beings with dignity which needs to be respected, and that both of us are members of the moral community who can hold others morally accountable, we are also committed to there being certain standards which we can hold each other morally responsible for failing to meet. What we have the authority to demand depends on which account of morality is correct. Darwall thinks that this account naturally lends itself to contractualism. Specifically, what actions are right or wrong largely depend on what sort of principles agents couldn’t reasonably reject, with morality’s content deriving itself (at least partly) from our equal dignity. Watson (2007) and Wallace (2007) dispute whether Darwall succeeds in arguing from the structure of second-personal reasons to the structure of morality, but for my purposes we can set this debate aside. What matters for my account is that our commitment to there being shared normative standards is also a normative felicity condition. In order to succeed in making a second-personal demand to you, I must presuppose that we have a shared authority qua members of the moral community, and that you are able to hold me to account for failing to meet those shared standards. In short, I must recognise that shared standards are in place, and I must see these standards as authoritative on me.

Recall that we are investigating why hypocrites lack the standing to blame. Here is what happens when we put all of the pieces of Darwall’s account together and examine hypocritical blame. In blaming, the hypocrite makes a second-personal demand. They make a demand as a member of the moral community. The hypocrite thereby holds the wrongdoer accountable to them, calling on the wrongdoer to answer to them and to comply with their moral demands. The hypocrite necessarily purports to have the authority to make this demand, and presupposes that the wrongdoer is an authoritative source of second-personal demands too. The hypocrite is thus purporting to address the wrongdoer from the shared normative perspective of the moral community. And they expect the wrongdoer to accept
that second-personal demand, thereby accepting the hypocrite’s authority to make such a demand.

However, the hypocrite is someone who has themselves committed wrongdoing similar to that which they are blaming the other person for committing. They have done something which is inconsistent with the shared normative standards they are purportedly committed to when blaming the wrongdoer. Of course, mere inconsistency with the moral demands that we make of others doesn’t make one a hypocrite, or prevent such demands from being genuinely second-personal. But the kind of inconsistency evinced by paradigmatic hypocrites is notable, because it demonstrates that they do not accept certain kinds of obligations as authoritative on them. The hypocrite is someone who tries to make certain kinds of second-personal demands on others, while failing to accept the authority of others to make the same kinds of second-personal demands on them. Wrongdoers cannot be expected to accept the demands of hypocrites, because doing so would accept the hypocrite’s authority to make such a demand, while that hypocrite presently does not accept others’ authority to make the same demand on them. In short, the hypocrite is not relating to the wrongdoer as an equal, and their demand is not one that could be accepted by a free and rational agent among equals.

On this account, inconsistency between one’s second-personal demands and behaviour doesn’t undermine standing when the source of the inconsistency is not proof of a failure to accept others’ second-personal authority. If I fail to comply with others’ demands because I am coerced, or have an irresistible desire, or a reduced capacity to comply with those demands, then there is no evidence that I fail to accept others’ demands on me. If, on the other hand, I have the capacity to easily comply with those demands but simply choose not to, then I likely do not treat those demands as authoritative.

7.5 Strengths Over Competing Accounts

In the previous chapter, I used Todd’s account of hypocrisy and standing to blame to identify the circumstances under which agents lack the standing to blame. The aim of this chapter was to explain why these agents lack the standing to blame in these circumstances. In examining normative powers, demands, authority and Darwall’s account of second-personal obligations, it thus seems like I’ve ended up with a competing account of hypocritical blame.

This is not a problem, because my account can incorporate all the strengths of Todd’s account, while avoiding its weaknesses. Recall in the previous chapter, we arrived at thinking
sufficient commitment to relevant values was necessary for standing to blame through Todd’s analysis of four thought experiments. If those thought experiments support my account equally as well as Todd’s, then we can just swap out his account for mine. My Darwallian account will remain equally persuasive. It may turn out that the class of agents who treat certain kinds of second-personal obligations as authoritative just is the class of agents who are sufficiently committed the kinds values which would condemn violating those obligations, and so Todd and I are talking about the same class of people, but I won’t try to demonstrate this. Here were the thought experiments used to support Todd’s account:

**Diet:** Charlie knows that Linus, who has a weakness for sweets, is trying to lose weight. Nevertheless, he takes Linus to a place for dinner that he knows is located next to an incredible ice cream shop. Quite predictably, after dinner Linus visits the shop next door and has some ice cream. Charlie blames Linus.

**Nazis:** Jonas and Thomas are Nazi commanders in a WWII death camp. Jonas orders Thomas to investigate the fence and sound the alarm if anyone is trying to escape. Thomas investigates, sees someone escaping and sounds the alarm. However, Jonas is actually a double agent, working to sabotage the camp. He only gave that order to keep up appearances, and he picked Thomas to investigate because he thought Thomas would be the most likely to have mercy and not sound the alarm, allowing the prisoners to escape. Jonas blames Thomas.

**Abandoned:** Paul, when he is 25, abandons his wife and newborn infant to go touring around the world and “find himself”. His wife and child suffer terribly as a result. Later in life, at the age of 45, he comes to realize the virtues and obligations of fidelity and becomes a committed husband and father to his wife and child, with whom he is somehow reconciled. He now holds values, very deeply, which condemn his past actions. Paul hears of another 25-year-old, Peter, who has abandoned his wife and child to explore the world just as he once did, and blames him.
Mistaken Blame: Samuel is a shop owner who has himself never tried to shoplift – but not for want of wishing to do so. Samuel simply lacks the means to try – a fact he regrets. He has no vehicle to travel to other stores, and there are no suitable targets within walking distance. If there were, he would try to shoplift from those stores. One day, Samuel gets a call: someone has been shoplifting from his own store. On hearing this, Samuel burns with rage and incredulity and a desire to confront the criminal. However, in fact the call has been made to the wrong number, and no-one has stolen anything from Samuel’s store.

Recall that according to Todd, Samuel and Charlie seemed to lack the standing to blame, while Paul and Jonas possessed standing to blame. It seems to me that Jonas treats his second-personal obligations to the Jewish prisoners as authoritative, and Paul treats his second-personal obligations to his family as authoritative. It also seems like Samuel fails to treat the second-personal obligations of other shop-owners as authoritative on him when blaming the non-existent thief. Three out of the four thought experiments thus straightforwardly support my account equally as well as Todd’s.

Diet seems under-described for our purposes. King’s (2015) original presentation of the case, which Todd shortened, describes Charlie’s action as permissible and not ill-intentioned. King also says that Charlie has a reasonable belief that Linus will resist the treats. My intuition is that when described this way, Charlie does have the standing to blame Linus. Todd too admits any blame Charlie would be entitled to is “extremely mild” (p. 356), so I don’t think this case supports either my account or Todd’s over the other. If Charlie gleefully invited Linus out, expecting him to fail, then I would be inclined to say he lacks the standing to blame Linus. But he would also not be treating Linus with second-personal respect.

I thus conclude that my Darwallian account of second-personal demands provides an equally acceptable explanation of our intuitions in these thought experiments as Todd’s account. Again, it could be that the class of agents who treat a certain kind of second-personal obligations as authoritative just is the class of agents who are sufficiently committed the kinds values which would condemn violating those obligations. But even if one thinks these classes of agents are different, they clearly have substantial overlap, as evidenced by the fact that they give the same verdict on standing to blame in these thought experiments.

If one is still tempted by Todd’s account, it is worth noting that my account has a number of strengths which his lacks. First, Todd couldn’t explain why sufficient commitment to
values was necessary for standing to blame, or what the ‘something further’ was that hypocrites lacked which undermined their standing to blame. My account can answer both of these questions: The ‘something further’ is the authority to make certain kinds of second-personal demands, and hypocrites lack this because they do not accept others’ authority to make those same kinds of demands. Second, my account renders our conception of ‘standing to blame’ consistent with our conception of ‘standing’ in many other domains. Standing is an ability to exercise a normative power. Third, in explaining why standingless blame is inappropriate, I’ve done so in such a way that makes it consistent with our ordinary understanding of why blaming someone who is innocent or excused is inappropriate, while still having the concept do some explanatory work. Both blaming when one lacks the standing to do so and blaming an innocent person violate the normative felicity conditions required for second-personal demands to, in fact, give other agents moral reasons for action. Blaming someone who has done nothing wrong violates the requirement that the target of our blame must have violated a legitimate demand, and blaming when one lacks the standing to do so violates the requirement that prospective blamers accept others’ authority to make demands on them too.

It may seem like I’ve pulled a bait-and-switch, since my previous chapter relied on Todd’s account and here I’ve ended up adopting a different account. But there are a number of reasons why this approach was best. First, as already identified, Todd’s and my accounts have substantial agreement regarding which kinds of agents lack standing to blame, so there shouldn’t be any concern that I’ve undone my previous chapter’s argument. Second, Todd’s account is a much more recent contribution to the literature than Darwall’s, and engages directly with recent debates on hypocrisy. Temporarily adopting Todd’s account allowed me to situate my argument regarding Milgram’s subjects within this debate. Third, Todd’s account was much simpler than Darwall’s. It would have been difficult to situate Darwall’s nuanced account amongst previous work on hypocrisy, and only later investigate whether we possess the standing to blame the obedient subjects. Fourth, some people might reject Darwall’s account of second-personal authority, and had I presented Darwall’s account first, this might lead them to conclude that there are no other grounds for thinking hypocrites lack the standing to blame. The previous chapter’s argument was neutral regarding which is the correct account of blame and moral responsibility, and could be accepted by many people who do not accept Darwall’s account of second-personal reasons.
7.6 Objections

My Darwalian explanation for why hypocrites lack the standing to blame has a number of strengths. Let me now consider some objections and clarify some aspects of the account.

One objection to this account concerns private blame. A common objection to the claim that blame expresses demands is that private blame doesn’t seem to express or communicate anything (Driver, 2016; Macnamara, 2013). Similar worries apply to blaming the dead or the absent. I don’t think that this objection is substantial, because whether or not private blame expresses demands, I am happy to accept the result. If private blame does, in fact, express demands, then I am happy to accept that hypocrites can also lack the standing to blame privately. We might also think that hypocritical blamers are excused for their blame in those instances where they lack any control over their private blaming attitudes. If, on the other hand, private blame does not express demands, then I am happy to accept that hypocrites do possess the standing to blame privately. We could decide that, in this instance, private blame is simply an emotion, with no further moral consequences, or say that this person still possesses an objectionable attitude towards the person they blame, or that this person is an appropriate target of certain kinds of attributability assessments.

Additionally, many different accounts of moral responsibility maintain that blame expresses demands, and so I am in good company (Hieronymi, 2004; McGeer, 2012; Shoemaker, 2007, 2011; Strawson, 1962; Wallace, 1996, 2011; Watson, 2004; cf. Macnamara 2013). I suspect that what is pumping the intuition that private hypocritical blame can be problematic, even if it doesn’t express demands, are cases where agents choose not to express blame out of self-interest. If someone experiences private blame, but recognises that expressing this would be inappropriate, then their blame may be something which ‘just happens’ to them and which we can’t reasonably rebuke them for having. In contrast, if they choose not to express blame because they want to deceive the other wrongdoer, then while they do the right thing, they fail to do it for the right reasons. They retain an objectionable attitude towards the wrongdoer while not recognising the authority of certain kinds of second-personal reasons, and retain a disposition to behave problematically in certain situations. These things can make agents fitting targets of negative moral assessment. But I don’t think I need to show that this person is thereby treating the target of their blame inappropriately, although my account can accept they are.
Another objection to my account concerns its application to the situationist experiments. It seems that most of us do accept the authority of the second-personal obligations to not electrocute people, even when we are ordered to by a scientist. Were we to take part in the experiment and obey, we would most likely self-blame, and take ourselves to be appropriate targets of blame from others in the moral community. This seems like evidence that we, in fact, accept the authority of the relevant second-personal demands, and thus have the authority to make second-personal demands ourselves.

It is important to note that believing that one is not abusing demands or violating normative felicity conditions doesn’t entail that one is, in fact, not doing so. The fact that one would self-blame does not imply that one has standing to blame. My argument is that, if we were to blame one of the obedient subjects, we would also be abusing the demands our blame expresses because the fact that we would act the same way is proof that we do not accept their second-personal demands as authoritative on us. Although we might ‘accept’ these demands in the way one ‘accepts’ a certain proposition is true by believing it, the relevant sense of accept is much stronger than this. Todd argued that in order to have standing to blame, agents must endorse a moral value and be sufficiently committed to acting in accordance with that moral value. In the previous chapter, I argued that what counts as sufficient motivation will be determined by the wrong under consideration, and how difficult it was for the agent to avoid wrongdoing. Those arguments apply just as forcefully when we adopt my Darwallian account. In order to accept a certain class of second-personal obligations, agents must have a certain degree of motivation to comply with them. Recall again that in this chapter, we are hypothetically assuming that if we had been in Obedience, we would have committed egregious wrongdoing, we would be blameworthy for our actions, and we could have easily avoided doing so.

Taking an agent’s standing to blame to be determined by whether they accept others’ second-personal authority allows us to explain why there seems to be some relationship between the seriousness of a wrong under consideration, and the number of transgressions an agent must commit before they lose their standing to blame others for similar transgressions. It also explains why we are sometimes uncertain whether agents possess standing to blame when their wrongdoing is the result of weakness of will. The vegetarian who refrains from eating meat because of its contribution to global warming, but eats steak one time, seems to retain their standing to blame regular meat-eaters. But a single instance of infidelity can cause one to lack the standing to blame adulterers. Though this vegetarian and
adulterer both commit a single transgression, the difference in their standing to blame is a product of the seriousness of their wrongs, and the ease with which they can avoid wrongdoing. Cheating on one’s partner is more serious than eating meat on one occasion, and much easier to avoid, given we all need to eat regularly and vegetarian options are not always readily available. Because of this, we take the adulterer’s actions to be strong evidence that they do not accept their partner’s authority to demand they not cheat, while the vegetarian’s actions do not necessarily show they fail to accept others’ second-personal demands that they not contribute to global warming.

In contrast, a ‘vegetarian’ who knows that they should avoid eating meat and accepts blame for eating meat, but who eats meat daily, lacks the standing to blame people who eat meat. This is because their continued meat-eating is evidence that they do not truly accept others’ demands to not contribute to global warming as authoritative. An agent who is ready to self-blame and accept criticism for their wrongs, but who keeps doing the wrong thing despite having the means to reform their habits, is too complacent with their moral failings. They lack the standing to blame agents who commit similar wrongs, because given their capacity and opportunity to avoid wrongdoing, their decision to not exercise this capacity expresses a lack of sufficient concern with the wrongness of their actions. Arguing that such agents retain the standing to blame simply because they accept blame for similar wrongs risks treating blame like a cost of admission to wrongdoing, rather than an authoritative demand which ought to be respected. Of course, some agents’ routine failure to do the right thing is caused by factors that do not express a failure to accept others’ second-personal demands as authoritative (e.g. addiction, mental illness). Such agents will retain their standing to blame.

Darwall explicitly looks at the concept of acceptance when examining how his second-personal account compares to The Golden Rule:

A moment’s reflection shows, I think, that the most natural way to take ‘accept’ in “don’t do to others what you wouldn’t or couldn’t accept their doing to you,” is in terms of the second-personal circle of concepts we identified at the outset of this book. What I can accept in these terms is what I would not object to or claim or demand otherwise … One should not act in ways that one demands or expects (or would demand or expect) that others not act, or equivalently, in ways that one would resent or object to. If I am going to object to others’ stepping on my feet (and thus demand that they not do so), then I must not step on their feet either. Lincoln’s famous remark, “As I would
not be a slave, so I would not be a master” (1989: 484) is along the same lines: As I would not allow others a master’s claim on me so should I not make a master’s claim on others.


In order to have the standing to address second-personal demands to others, it is not enough to simply believe that one recognises others’ second-personal authority, nor to simply agree one is (or would be) blameworthy for violating others’ demands. When we blame the obedient subjects for electrocuting a stranger to the point of heart failure, the subjects are licensed to expect that we would not treat them in that way either. But the empirical evidence strongly suggests that, had the obedient subject been in the learner’s seat, and we in the teacher’s position, this expectation would be violated. We thus lack the authority to make certain second-personal demands towards the obedient subjects, and therefore lack the standing to blame the obedient subjects, even if they are blameworthy.

7.7 Conclusion

In this chapter, I asked why some agents lack the standing to blame. First, I argued that previous attempts to analyse standing to blame are problematic, and thinking of blame without standing as inappropriate blame doesn’t give any clear answer to why hypocrites lack standing to blame. I then argued that we should reject thinking of standing in exclusively moral terms. Considering our conception of standing in other contexts suggested that standing to blame might refer to an ability agents have. Using Darwall’s account of second-personal obligations, I argued that hypocrites lack the standing to blame because they lack the authority to make the demands agents typically make in blaming. This account had a number of strengths over Todd’s account. I then applied this argument to *Obedience*, and argued that in blaming the obedient subjects, we purport to accept certain kinds of demands as authoritative on us. However, the fact that we would have culpably done the same thing in their situation, while retaining the capacity to disobey, shows that we in fact don’t accept these demands as authoritative, even though we believe we do and would likely self-blame if we had ever been placed in that situation. This account explains our reluctance to blame the obedient subjects, even if we strongly believe that they are blameworthy for their wrongdoing.
Conclusion

This thesis has sought to make sense of our conflicting attitudes towards the subjects who commit wrongdoing in the situationist experiments, particularly those in Milgram’s ‘Obedience to Authority’ studies. The behaviour of the subjects in these settings sits uneasily with us for a number of reasons. The wrongdoing itself is concerning, because the subjects are distressed about what is happening, no-one is coercing them into administering the shocks, and yet, they continue to obey. The fact that so many subjects in these settings commit wrongdoing is concerning, because it conflicts with our perception of most people as minimally decent. And the fact that the subjects are so ordinary is concerning, because it suggests we might have done the same in their shoes, which sits in tension with our self-image as people who know and care about the difference between right and wrong.

How we reconcile these tensions depends on why we think these subjects behave as they do, and when we think agents are blameworthy for wrongdoing. Most philosophers who have examined these experiments have done so in the context of arguing that these studies aren’t evidence that global character traits don’t exist. Focusing on this debate has meant much less attention has been paid to explaining why subjects act as they do. It was only by thinking about why particular changes to subjects’ environments result in particular changes to their behaviour, that we could then begin to assess subjects’ moral responsibility.

I have presented two very different arguments for the conclusion that it is inappropriate to blame the obedient subjects. Both can help us make sense of our ambivalence towards the subjects, despite giving opposite assessments of the subjects’ blameworthiness. According to the first argument, the subjects’ behaviour is so surprising, and deviates so dramatically from how they normally act, that we have considerable reason to think they were not interpreting their situation accurately. By looking at evidence from a wide range of studies, I have highlighted how we often rely on numerous cues from our environment to interpret our situation. In unfamiliar, stressful situations with multiple factors competing for our attention, it is possible for us to become confused and for our interpretations to be misled. If other agents, who we take to be trustworthy, then take advantage of this and direct us towards certain actions and considerations, we can find ourselves committing wrongs that are extremely out of character. It is easy enough to overcome this distorting influence when one knows that the other agents have nefarious motives, or when one has had practice with recognising when certain features of our environment should be ignored. But we need to
guard against assuming that doing the right thing when one doesn’t know these things, or hasn’t had practice, is only marginally more difficult, simply because it seems intuitively obvious to us that certain features of that environment should be ignored. We need to examine a range of settings in which people succeed or fail at doing the right thing, and think about what best explains their failures. When subjects’ failures to do the right thing do not track a recognisable pattern of reasons, it becomes less plausible that these failures are the result of some criticisable motivational defect, and more likely that certain factors are reducing their capacity to avoid wrongdoing. These considerations show that many of the subjects in the situationist experiments are excused for their wrongdoing, and thus are not appropriate targets of blame.

Though I stand by this argument, there are additional considerations which can make us feel uncomfortable with blaming the subjects, even if they are morally responsible and blameworthy for their actions. The fact that the subjects are so ordinary, and so much like us, seems to matter in some way. If the obedient subjects possessed the capacity to avoid wrongdoing, but simply didn’t exercise it, and if many of us would have done the same thing in their shoes, then there seems to be something off about blaming them. Such sentiments have been expressed for centuries in sentiments like ‘Let ye without sin cast the first stone’. I’ve argued that most of us lack the standing to blame the obedient subjects, because our disposition to act in the same way they did shows that we are not sufficiently committed to the kinds of values we implicitly defend when blaming, or, alternatively, shows that we do not accept the authority of others to make second-personal demands on us.

Regardless of which argument one endorses, a remaining problem is reconciling the thought that blaming the obedient subjects is inappropriate with the thought that any victims of the subjects’ wrongdoing are entitled to blame. The fact that victims are entitled to blame their perpetrators seems to be a regular feature of our moral practices, and supports the argument that the obedient subjects really are culpable. I have argued that there is nothing inconsistent with claiming that excused wrongdoers are undeserving of blame, while also maintaining they can incur many of the duties that culpable wrongdoers incur. Our intentional actions matter greatly; they are the central means by which we exercise our agency. But we also sometimes interact with others when our rational capacities are compromised, and the importance of these interactions should not be neglected in our theory of moral responsibility.
Bibliography


Court Martial of Ivan L. Frederick, II, 4, American Civil Liberties Union Database 490–613 (Headquarters, III Court, Baghdad and Victory Base, Iraq 2004).


doi:10.1002/ejsp.170


Wisnewski, J. J. (2009). What We Owe the Dead. *Journal of Applied Philosophy, 26*(1), 54–70.