2013

Towards a Revisionist Account of Moral Responsibility

Kelly Anne McCormick

Abstract: Revisionism is the view that we would do well to distinguish between what we think about moral responsibility and what we ought to think about it, that the former is in some important sense implausible and conflicts with the latter, and so we should revise our concept of moral responsibility accordingly. There are three main challenges for a successful revisionist account of moral responsibility: (i) it must meet the *diagnostic challenge* of identifying our folk concept and provide good reason to think that significant features of this concept are implausible, (ii) it must meet the *motivational challenge* and explain why, in light of this implausibility, our folk concept ought to be revised rather than eliminated, and (iii) it must meet the *prescriptive challenge* and provide an account of how, all things considered, we ought to revise our thinking about moral responsibility. In order to meet (iii) revisionism must provide a prescriptive account of responsibility that is free of the problematic features of our folk concept identified in meeting the diagnostic challenge, is naturalistically plausible, normatively adequate, and justifies our continued participation in the practice of moral praising and blaming. So, while the first of these three challenges is primarily concerned with the nature of our concepts, the latter two move to questions about whether or not, to use Dennett's terms, we can defend and accept an account of moral responsibility "worth wanting."

In my dissertation I raise a new problem for revisionism, the *normativity-anchoring problem*. The heart of this problem is that the methodological commitments used to motivate revisionism and distinguish the view from conventional theorizing about moral responsibility make it uniquely difficult for revisionists to justify our continued participation in the practice of moral praising and blaming. Following Manuel Vargas, who has thus far developed and defended the view most rigorously, revisionists endorse the following skeptical claim: it is possible that our

intuitions fail to inform us about what responsibility is, and furthermore we lack good epistemic reasons for thinking that they ever do. For conventional theorists, the fact that a particular account of responsibility best aligns with our refined intuitions, beliefs, and theoretical commitments is reason enough, *ceteris paribus*, to endorse that view. But revisionists who endorse the skeptical claim must find some alternative method for arguing that their prescriptive account is one that we should in fact endorse. One alternative, suggested by Vargas himself, is to show that the prescriptive account in question justifies our continued participation in the practice of moral praising and blaming, and preserves the "work of the concept." However, I argue that Vargas' own claim that the prescriptive account he offers promotes an independently valuable form of agency fails to bridge the gap between axiological claims about value and normative claims about how we should treat responsible agents. Moreover, bridging this gap looks to be a serious problem for any form of revisionism which shares the methodological commitments used to motivate the view thus far, and so further development of revisionism requires having a solution to the normativity-anchoring problem in hand.

I go on to develop a new revisionist strategy that avoids the normativity-anchoring problem. I propose and defend a new methodological assumption (hereafter referred to as MAP) that I argue revisionists can and should accept, capable of preserving the skeptical spirit of revisionism while identifying a particular class of intuitions about moral responsibility as having a privileged epistemic status. In particular, I argue that revisionists can and should accept that widespread judgments about responsibility generated by concrete cases which elicit a strong affective response in the person making the judgment have a privileged epistemic status in our responsibility theorizing. My arguments in support of this assumption depend on an analogy

between the responsibility judgments in question and the kinds of paradigmatic judgments which constrain our ethical theorizing more generally. Having established this analogy I then offer a series of companions in guilt style arguments for the claim that the epistemic status of these two kinds of judgments should stand and fall together. I conclude that the responsibility judgments in question should ultimately share the privileged status of the paradigmatic ethical judgments in question, and thus play an evidentiary role in our theorizing about moral responsibility. If these arguments are successful then acceptance of the methodological assumption I defend allows revisionists to avoid the normativity-anchoring problem while preserving the unique methodological spirit which motivates revisionism and sets it apart from conventional theorizing about moral responsibility.

TOWARDS A REVISIONIST ACCOUNT OF MORAL RESPONSIBILITY

Kelly McCormick
B.A. Colgate University 2006

DISSERTATION
Submitted in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Philosophy

Syracuse University
June 2013

Acknowledgment

There are a great many mentors, colleagues, and friends without whom what follows would not be what it is (or perhaps ever have come into existence in the first place). First and foremost I would like to thank my advisor, André Gallois for the countless hours spent discussing what must have at first blush seemed like a number of very crazy ideas, for the wealth of invaluable feedback he provided on first, second, third, *ad infinitum* drafts, and for his constant support and encouragement. This project and my development as a philosopher have both benefited in ways impossible to do credit to here from the time he invested in them. Next, a big thanks to Manuel Vargas for the helpful and insightful feedback provided on drafts of this project at all of its many stages. Thanks to Ken Baynes, Ben Bradley, Mark Heller, and Derk Pereboom for their excellent comments and questions in my dissertation defense, and for an all-around engaging and productive discussion of this project. The faculty and graduate students of the Syracuse University philosophy department provided helpful feedback on early drafts of parts of this work (especially Chapter Three) in both ABD workshops and informal discussions. Special thanks are due to the following colleagues and dear friends in particular: Kirsten Egerstrom, Jake Greenblum, Matthew Koehler, Amy Massoud, Rachel McKinney, and Aaron Wolf. Finally, thanks to my loving and supportive family for never once asking, "Shouldn't you think about getting a real job," and for keeping me going every step of the way.

Table of Contents

Chapter One

Revisionism: Goals and Challenges

Introduction

As the "new kid on the free will block," revisionism does not benefit from the same degree of familiarity as many of its competitors (Vargas 2009, 46). Nor is it always clear exactly how best to characterize where the view is situated dialectically in the contemporary free will debate. In order to assess the strengths and weaknesses of revisionism and to attempt to defend and develop the view further it is therefore helpful to first clarify what revisionism is, and where it stands in relation to the other influential views considered to be the main players in this debate.

I begin this chapter with a brief overview of the contemporary philosophical discussion of free will and moral responsibility. This overview is not intended to be comprehensive, nor does it provide an exhaustive account of the different views one might hold on the topic. Rather, it is a characterization that I take to be the most instructive in mapping out where revisionism stands in relation to some of its most prominent competitors. In Section 2 I turn to revisionism itself. I first discuss how revisionism differs from other traditional accounts of free will, and the various ways one might motivate revisionism. I then identify three challenges that any revisionist view must meet: the *diagnostic*, *motivational*, and *prescriptive* challenges. The majority of Chapters 1-3 will be devoted to discussion of these three challenges and the potential for revisionists to meet them. In Section 4 I discuss some further motivating considerations on behalf of revisionism, and in Section 5 I provide a brief outline of the chapters to follow.

The discussion of revisionism in this chapter (and Chapters 1-3) borrows heavily from recent work by Manuel Vargas (2005, 2009, 2011, 2013). While strands of revisionism about

free will prior to Vargas' work can be found in the literature[1], Vargas has been the first to systematically develop and defend a revisionist account of moral responsibility and his brand of revisionism has thus far generated the most attention for the view by far. Vargas-style revisionism is therefore the basis for much of the discussion of the basic contours of the view in this chapter and the next, and as such I will often use 'revisionism' to refer to Vargas' brand in particular. In my view revisionists do well to accept many of Vargas' arguments for meeting the diagnostic and motivational challenges to revisionism, his work to carve out a space for revisionism in the larger debate, and many of his methodological suggestions for revisionist prescriptive theory construction. However, there are some points at which I will diverge from Vargas. When I do so I hope to make explicit the distinction between what I take Vargas, as opposed to revisionists in general, to be committed to. This distinction will be particularly relevant to the discussion of the considerations that motivate revisionism in this chapter, which will again become salient in Chapter 3 and Chapter 6.

Lastly, it is important to emphasize that what both Vargas and myself are concerned with is a view about *moral responsibility*. As such the notion of 'free will' will be relevant only to the extent that it is whatever freedom condition is required for moral responsibility. This reflects the contemporary shift in the literature away from the language of 'free will' and towards explicit accounts of 'moral responsibility',[2] though as a matter of habit or convenience there is sometimes a slide back and forth between talk of free will and talk of responsibility. This slide is unfortunate and can sometimes lead to confusion. However, given the long history of this particular philosophical debate it seems unlikely that 'free will' will be replaced entirely by

[1] For precursors to Vargas' revisionism see Smart (1961), Heller (1996), and Hurley (2000). Nichols (2007) has also argued for a view that makes a distinction between descriptive and prescriptive accounts of responsibility, which I take to be a defining feature of revisionism. This distinction will be discussed in much further detail later in the chapter.

[2] For a voice of dissent, see van Inwagen (2008).

'moral responsibility' anytime soon. Those working in this area must therefore do the best they can given the historical tradition they find themselves engaged in, and make explicit which of these two intimately related concepts they are interested in (often with a footnote or brief paragraph just like this one). Likewise, hereafter whenever I use the term 'revisionism' I intend to refer to revisionism about *moral responsibility*, though for ease of exposition I will refer to the contemporary debate on free will and moral responsibility in the following section as *the free will debate*.

1 The state of the debate

One standard way of characterizing the free will debate is in terms of where a variety of views stand in regards to what many have called the *compatibility question*: is free will compatible with determinism? Compatibilist views provide an affirmative answer to this question, and incompatibilist views answer it negatively. In the twentieth century the majority of prominent accounts of free will could be classified as belonging to one of these two categories. However, the emergence of a variety of views in the last few decades complicates current attempts at such simple classification. I will discuss some of these views shortly. First, it is helpful to say a bit more about some important distinctions between traditional compatibilist and incompatibilist views.

One popular way of categorizing compatibilist accounts of free will – those which answer the compatibility question in the affirmative – is in terms of *Real Self* views versus *Reasons* views.[3] Both posit some condition or set of conditions necessary[4] for free will and responsibility which can be met even if determinism is true. Real Self views focus on the idea that the agent

[3] I borrow this terminology from Vargas (2013, Chapter 5).
[4] These conditions should also be jointly sufficient, or the view in question will not ultimately provide an affirmative answer to the compatibility question.

herself must play a certain role in bringing about an action in order to be responsible for it.[5] The kind of role that the agent must play will differ depending on the view in question. Vargas provides a helpful summary of some of the various key requirements for differing Real Self views:

> Contemporary versions have variously emphasized that the agent needs to "identify" with the motives that lead to the act, or the act has to be expressive of a "Real Self" or expressive of the agent's values, or the action has to be an expression of the regard in which the agent holds others.[6] (2013, 136)

The shared feature of these views is that certain psychological states of the agent (for example, their desires, motives, or intentions) must stand in the right sort of relation to an action in order for the agent to be responsible for it.

On the other hand, compatibilist Reasons views focus less on the expression of the agent's character and identity but rather on whether or not the agent possesses a particular power to identify, assess, and respond to reasons. Different Reasons views provide different accounts of how this power is characterized. For example, according to Fischer and Ravizza (1998) the mechanism that issues in the action must be moderately reasons responsive, where moderate reasons responsiveness requires that, at least sometimes, if there were good reasons for acting otherwise the agent would recognize those reasons and act accordingly.[7] A further distinction can also be made regarding whether or not a particular Reasons view is symmetrical. While Fischer

---

[5] This follows Susan Wolf's terminology (1990).

[6] For views that develop around varieties of these central conditions see especially Frankfurt (1971), and Watson (1975).

[7] This is of course a very coarse description of Fischer and Ravizza's view. They provide a detailed discussion of the distinctions between different degrees of reasons responsiveness, and also build in additional conditions including the requirement that the agent must be sensitive to *moral* reasons. Fischer and Ravizza also propose a second necessary condition for responsibility: that the agent must view the action as their own. While their view is in many ways a paradigm of a compatibilist Reasons view this second condition also makes the view reminiscent of a Real Self view. Fischer and Ravizza's view therefore highlights the fact that there is a great deal of room for overlap between compatibilist Real Self and Reasons views.

and Ravizza claim that the same conditions for moral responsibility hold for both praiseworthy and blameworthy actions, there are some who disagree. Susan Wolf (1990) and Dana Nelkin (2008) are perhaps the most prominent defenders of asymmetrical views, arguing that while only a reasons responsiveness condition(s) must be satisfied in order for an agent to be praiseworthy for a particular action, in order to be blameworthy she must also have the ability to do otherwise.

There is a further distinction that cuts across both Reasons and Real Self compatibilist views: the degree to which they are either *historical* or *snapshot* views.[8] Historical views are those that take the quality of the chain of causation leading up to an action to be important in determining whether or not an agent is responsible for that action, while snapshot views focus on features of the agent at the time of the action. Given the nature of Real Self and Reasons views, Real Self views tend toward an historical approach, while Reasons views tend toward a snapshot approach. However, there is again a great deal of room for overlap. The moderate reasons responsiveness condition of Fischer and Ravizza's view, for example, is a snapshot condition, but they also posit an additional necessary condition. According to Fischer and Ravizza an agent must also view the action as her own in order to be responsible, and this additional necessary condition gives their account an historical flavor.

Before turning to incompatibilist views it is important to note the absence of certain compatibilist views from the above discussion. First, I make no mention of c*onditional analysis* versions of compatibilism. By conditional analysis I mean those views that attempt to provide a successful compatibilist treatment of the ability to do otherwise.[9] While such views dominated much of the discussion of compatibilism in the mid-twentieth century they have fallen largely

[8] Again, I borrow this terminology from Vargas (2013, Chapter 5).
[9] See G.E. Moore (1898).

out of favor in the contemporary debate, and so I leave them off the map here.[10] Nor have I mentioned Strawsonian compatibilism. This brand of compatibilism attempts to sidestep any robust metaphysical treatment of the compatibility question because, as Strawson (1968) famously argued, it is psychologically impossible to give up the reactive attitudes that result from our judgments of praise and blame. I will discuss this view in further detail shortly, but will turn first to views that give a negative answer to the compatibility question.

Incompatibilist views can be further categorized in terms of the answer they provide to the *free will question*: given that determinism and free will are incompatible, are agents ever free and responsible (or can they be)? If the answer to this question is affirmative, then the incompatibilist view in question is libertarian. There are at least three main categories that can then be used to classify the diverse array of libertarian views: *non-causal*, *event causal*, and *agent causal* libertarianism. According to non-causal libertarian views, argued for and defended by Ginet (1990, 2002, 2007), McCann (1998), and Pink (2004), unless an action is wholly determined by causes that lie outside of the agent she is responsible simply in virtue of the fact that the choice or volition that brings the action about occurs in and belongs to her. According to event causal libertarians, such as Nozick (1981), Balagaur (1999, 2004), and Mele (2006), an agent can be responsible for actions that are undetermined and caused by her reasons or some other deliberative event. Robert Kane (1996) also offers a unique brand of event-causal libertarianism, according to which efforts of will – cases where one's will is genuinely conflicted and the outcome of one's choice is not only undetermined but indeterminate (the possible outcomes have no distinct probability of occurring) – are necessary in order for an agent to be

[10] One might take issue with this claim in light of the recent emergence of what has been coined *new disponsitionalism* (McKenna 2009b), a version of compatibilism which attempts to offer a positive compatibilist treatment of the ability to do otherwise. I will not discuss these views in detail here, but proponents of this brand of compatibilism include Smith (2003), Vihvelin (2004), and Fara (2008).

autonomous and capable of being responsible for later choices and actions.[11] Finally, according to agent causal libertarianism, defended by Chisholm (1966) and more recently Randolph Clarke (1993, 1996, 2003) and Timothy O'Conner (1995, 2000, 2005, 2009, 2011), an agent is free and responsible for her action when she herself brings it about by acting as an agent cause. Explaining the nature of agent causation is no easy task and the details differ depending on one's preferred view, but one shared feature of these accounts is that when an agent acts as an agent cause the cause of the action in question cannot be explained in terms of any other events (psychological, neurophysical, or otherwise) occurring in the agent prior to the choice or action. Or, to put the point more simply, the agent must be the uncaused cause of her action.

According to older terminology, incompatibilists who answered the free will question negatively were referred to as hard determinists. According to hard determinism, determinism is true (or close enough to being true to rule out freedom and responsibility), and so we are not free and responsible agents.[12] Like conditional analysis compatibilist views, hard determinism has fallen largely out of favor in the contemporary literature. However, *hard incompatibilism*, a view that plays a very prominent role in the contemporary debate, captures much of the spirit of hard determinism. Hard incompatibilism, developed and defended rigorously in the last decade by Derk Pereboom (2001, 2009a, 2009b, and Fischer et al. 2007) is a view that remains agnostic regarding the truth or falsity of determinism. Regardless, it is highly unlikely that human beings are free and responsible agents *either way*. Pereboom argues that free will and responsibility are incompatible with determinism (and so compatibilist views fail), *and* most kinds of indeterminism (and so the bulk of libertarian views fail). While he grants that we could be free and responsible if agent-causal libertarianism were true, our best scientific theories give us no

[11] For further development of this view see also Kane (Fischer et al. 2007).
[12] For a more recent defense of hard determinism see Wegner (2002).

reason to think that we are in fact capable of acting as agent causes, and may even provide reason to think that we do not have such a power. So, in light of this we ought to give up the idea that we are free and responsible agents and examine which of our reactive attitudes and practices can be preserved upon jettisoning this belief. Pereboom argues that much of what we care about regarding responsibility can be maintained in the face of hard incompatibilism.[13]

Before summing up this discussion of incompatibilist views it is helpful to make a quick note about a shift in the literature that Pereboom (2001) himself makes explicit. This is the recent shift in focus from *leeway* to *sourcehood* conditions for moral responsibility. Leeway incompatibilists are those who hold that the ability to do otherwise is what is essential to being free and responsible. Sourcehood incompatibilists are those who hold that it is not the ability to do otherwise that matters most, but whether or not an agent is the *ultimate source* of her action. The latter has also been referred to as an *ultimacy* or *origination* condition. A detailed discussion of what sourcehood requires is well beyond my current purposes. Here I wish only to point out that questions about the ability to do otherwise, which dominated much of the literature in the second half of the twentieth century, enjoy far less attention today. While there is still a great deal of discussion about what the ability to do otherwise amounts to and whether or not it is required for moral responsibility, much of this debate is largely a result of questions regarding the relationship between this condition and some form of sourcehood. For many contemporary incompatibilists it is the sourcehood condition, not the ability to do otherwise, that has taken center stage.

This discussion completes a brief taxonomy of contemporary views categorized according to their response to the compatibility question. However, not all contemporary views

[13] In particular, Pereboom (2001, 2009b) argues that giving up the belief that we are free and responsible need not negatively impact our emotions and would in fact allow us to preserve positive affective attitudes like love, while providing reason to jettison potentially harmful ones such as moral resentment and anger.

can be easily classified in this way. For example, Strawsonian compatibilism largely sidesteps metaphysical questions about the compatibility of free will and the various proposed conditions for free will and responsibility entirely. Rather, Strawson argues that because it is impossible for us to give up our reactive attitudes and the practice of holding each other responsible these metaphysical questions are irrelevant and we *must* take free will and determinism to be compatible.[14] While our individual reactive attitudes and attributions of praise and blame can be justified internally on a case by case basis and our standards for judging them can be refined, the overall practice of attributing praise and blame and holding agents responsible cannot be given up. So, no external justification of the practice can be provided, and furthermore none is required.

Strawson's is not the only view that does not easily fit the traditional compatibility question taxonomy. Many have argued for variants of what have been called "no free will either way" or *eliminativist* views. Galen Strawson (1993), for example, argues for *impossibilism*. According to Strawson, free will and determinism are incompatible, but we cannot be free and responsible agents regardless of the truth of determinism. According to Strawson, it is essential to the concept of free will that one must act as an uncaused cause which, he claims, is logically impossible. Other proponents of views that deny that it is possible for human beings to be free and responsible agents regardless of whether determinism is true or false include Saul Smilansky's *illusionism* (2000) and Richard Double's *non-realism* (1991, 1996). Finally, Pereboom's hard incompatibilism might also be classified as a version of eliminativism. While he grants that it is possible for agents do be free and responsible, he takes it to be highly unlikely

[14] See Strawson (1968). For additional views continuing in the Strawsonian tradition, see also Watson (1987) and Wallace (1994).

that they ever in fact are. So, we might also take hard incompatibilism to be a "*probably* no free will either way" brand of eliminativism.

This concludes a brief tour of the terrain of the contemporary philosophical discussion of free will and moral responsibility. I have of course not been able to do due diligence to many (if not most) of the interesting questions that have taken a prominent role in this discussion. Here my goal has simply been to provide a lay of the land in order to more clearly explicate where revisionism stands in relation to some of these views. According to the above taxonomy revisionism falls under the category of unconventional views that do not take the compatibility question to be central. Like these views, revisionism approaches the issue from a different direction. And, like Strawsonian compatibilism and unlike eliminativist views, revisionism is a success theory – according to revisionists we can be and sometimes are responsible agents.

2 What is revisionism?

One helpful way of understanding revisionism is to identify how it differs from the views discussed above. For starters all of these views share a particular method of approaching the issue. Vargas points out that according to this method one comes to a conclusion about free will and moral responsibility primarily

> …via reflection on our concepts as we find them, and we test proposals by checking to make sure they do not run afoul of our intuitions about cases….Throughout, the governing presupposition is that our metaphysics of free will can be read off of our beliefs about free will, our intuitions about cases and principles, and what these imply. (2011, 459)

The main point here is that this method of generating a theory of moral responsibility seems to be based squarely within the framework of descriptive metaphysics, and this methodology does not allow a theory of free will and responsibility much departure from our actual beliefs,

theoretical commitments, and intuitions. Revisionism offers a radical departure from this method in making explicit the distinction between what we *do* think about moral responsibility and what we *ought* to think about it.[15]

With this distinction in mind there are two different types of an account one might provide. First, like the majority of the views discussed in the previous section, one might provide a *diagnostic* or descriptive account of moral responsibility. The goal for such an account is to provide the most accurate account of our refined and systematized beliefs, theoretical commitments, and intuitions regarding moral responsibility as they stand. Such an account is therefore closely related to the idea of concept mapping (or, if you like, conceptual analysis) that has traditionally taken a prominent role in descriptive metaphysics.[16] Alternatively, one might provide a *prescriptive* account of moral responsibility. Here the goal is to provide an account of what we *should*, all things considered, take moral responsibility to be. Unlike diagnostic accounts, a prescriptive account is not wedded to our present beliefs, intuitions, and theoretical commitments regarding moral responsibility. However, it is in theory possible for there to be little difference between our best diagnostic and prescriptive accounts of moral responsibility.

Given this distinction between prescription and diagnosis one can make a further distinction between conventional and revisionist theories of moral responsibility. Vargas makes this distinction as follows:

> "Conventional" accounts entail consistency between prescription and diagnosis….
>
> "Revisionist" views are those on which the proposed prescriptive account conflicts with the diagnostic account….What sets revisionist accounts apart from their conventional counterparts is

[15] I credit Vargas (2005) with first making this distinction explicit in the responsibility literature. However, Shaun Nichols (2006, 2007) has also made this distinction and given it significant attention in structuring his own account of responsibility.

[16] For further discussion see Jackson (1998).

> the contention that we should abandon some of the commitments that constitute our ordinary way of thinking about free will. (2011, 460)

So, on a conventional account our best prescriptive account of what we ought to take moral responsibility to be does not conflict with our actual folk concept as it stands. Revisionist accounts, on the other hand, are those that maintain that our best prescriptive account of responsibility is not only different from, but also conflicts with, our best diagnostic account.[17] So, we ought to abandon some of our actual beliefs and theoretical commitments regarding moral responsibility and revise our concept. We ought to accept that responsibility is in some important way different from what we currently take it to be. The important point here is that there is at least one major methodological difference between Vargas and most, if not all, conventional responsibility theorists. Revisionists motivate a distinction between diagnostic and prescriptive account of responsibility via the following skeptical claim about what our intuitions can be said to reveal about moral responsibility:

> **The skeptical claim:** it is possible that our intuitions fail to inform us about what responsibility is, and furthermore we lack good epistemic reasons for thinking that they ever do.

For most conventional responsibility theorists, providing the best account of free will and moral responsibility just *is* getting clear about our concept. The best account, or the one that we should ultimately endorse, is the one that best aligns with our theoretically refined and systematized intuitions about principles and cases. But Vargas makes the interesting and somewhat radical claim that there are few reasons (and even less argumentation) in support of the claim that these

[17] It is unclear whether or not Vargas in particular wants to go so far as to say that a revisionist account is one in which the *correct* prescriptive account differs from the *correct* diagnostic account. This is an interesting question and one that I will return to in later chapters.

intuitions tell us much, if anything about the nature of responsibility. He puts the point as follows[18]:

> Revisionists are not bound by intuitions in the same way as compatibilists; revisionists are prepared to acknowledge a difference between what we believe and what we should believe and traditional compatibilists are not. For traditional compatibilists, if the theory gets the intuitions right, and if the theory provides some guidance on handling new or borderline cases, then it has done its work….Revisionists, however, cannot always appeal to intuitions, for revisionists disavow those intuitions rooted in our putatively error-ridden folk concepts. (Fischer et al. 2007, 216)

Commitment to the skeptical claim above is what best motivates Vargas' revisionist claim that what we think about responsibility does not necessarily tell us much, if anything, about what we ought to think about responsibility, and that we should treat descriptive and prescriptive questions about responsibility separately. Not only is it possible that these intuitions fail to get things right, but we lack any good epistemic considerations in favor of thinking that sometimes they *do* get things right. And this means that they simply do not have adequate epistemic standing to play the evidential role in revisionist theorizing about responsibility that they do for conventional theorists.

I introduce the skeptical claim here because in various parts of his work Vargas presents this departure from traditional theorizing about free will and responsibility as one of the defining features of revisionism, one that does a great deal of work in motivating the idea that revisionism is a unique and interesting view. However, in Chapter 3 I argue that this feature leaves revisionism open to a serious objection, and so should be divorced from the view more generally. One of the goals of my overall project will be to show that if revisionism is to succeed (which I

[18] The following quote refers explicitly to "traditional compatibilists," but can be extended to conventional responsibility theorists more generally.

think it can), it can and should abandon the skeptical claim, at least as it has been presented here. In Chapter 6 I will argue that revisionists can and should accept a *qualified* methodological assumption about the proper role of intuition in our responsibility theorizing, one that recognizes a distinction between the epistemic status of some of our intuitions about responsibility and others. I argue that this qualified methodological assumption can preserve the much of the motivation for revisionism while avoiding the difficulties that arise from commitment to the skeptical claim presented above. I will discuss these issues in much greater detail in Chapter 3 and Chapter 6. Here I intend only to identify the skeptical claim, and make explicit that commitment to this claim is a defining, motivating feature of Vargas-style revisionism. Ultimately I will argue that it need not – and more importantly *should* not – be viewed as a defining feature of revisionism more generally.

To sum up, revisionism is the view that we would do well to distinguish between what we think about moral responsibility and what we ought to think about it, that the former is in some important sense implausible and conflicts with the latter, and so we should revise our concept of moral responsibility accordingly. I turn now to what I take to be the three main challenges for a successful revisionist account of moral responsibility.

3 Three challenges for revisionism: diagnostic, motivational, and prescriptive

A successful revisionist theory must meet at least three challenges. First, it must address the descriptive task of identifying our folk concept of responsibility, and provide good reason to think that the claim that this concept has application is, in some important sense, implausible. I will hereafter refer to this as *the diagnostic challenge* for revisionism. Second, it must motivate why, in light of this implausibility, our folk concept ought to be revised rather than eliminated. I

will hereafter refer to this second task as *the motivational challenge* for revisionism. Finally, it must address the prescriptive task of providing an account of how, all things considered, we ought to revise our concept. I will hereafter refer to this as *the prescriptive challenge* for revisionism.

At first glance it may seem like revisionists are faced with a daunting task. Like conventional theorists, revisionists must do some descriptive work and provide an account of our folk concept of moral responsibility.[19] But unlike most conventional theorists, upon meeting the diagnostic challenge the revisionist's work is not done.[20] Once revisionists have provided an account of our folk concept, in order to motivate revision they are faced with an additional two-part challenge. First, they must provide reason for thinking that the concept is deeply problematic by identifying significant features of the concept that are either incoherent or implausible.[21] Second, they must provide arguments for why, in light of this incoherence or implausibility, the concept ought to be revised rather than eliminated entirely. So, revisionists face the dual motivational challenge of showing that there is something seriously wrong with our folk concept, but not so hopelessly wrong that it ought to be eliminated entirely.

In order for the task of motivating revision over elimination to even get off the ground it must also be possible for revisionists to provide a tenable prescriptive account of moral responsibility. What is needed is an account of how we ought to revise our concept of moral

---

[19] One might ask who the "our" here refers to: all human beings? Only those who reside in modern Western societies? This is an interesting question and it is not at all clear that there is any such thing as a single, shared, global concept of moral responsibility (for recent empirical work on this issue see Sarkissian et al. (2010)). And Vargas (2013) himself admits the difficulty of responding to this question. It may be the case that what revisionists (and conventional theorists doing primarily descriptive work on mapping our concept) have to say about responsibility may not generalize globally across radically different cultures, and may be restricted to a modern Westernized concept of responsibility.

[20] This is true of some conventional theorists as well, in particular those who defend eliminativist views. Because such views maintain that we are not free and responsible they must say something about what is to become of our reactive attitudes and practice of moral praising and blaming in light of this.

[21] I will discuss some potential worries having to do with concept individuation in further detail in Chapter 3.

responsibility. The most basic constraint on such an account is that it must respect *the work of the concept* (Vargas 2011, 2013). Vargas identifies the work of the concept of moral responsibility as the following:

> …I propose that we understand the work of the concept of moral responsibility as having to do with regulating inferences about differential moral praiseworthiness and blameworthiness, marking those who are praiseworthy or blameworthy and those who are not. (2013, 100)

One of the key challenges for revisionists is to ensure that their prescriptive account is in fact an account of genuine *responsibility*, or at least something close enough to it that it might appropriately be called responsibility. In order to avoid charges of changing the subject[22], revisionists must tie their prescriptive account closely to the responsibility practices, attitudes, judgments, and inferences involved in making assessments and attributions of moral praise and blame. This prescriptive account must be an account of whatever it is we care about and intend to refer to when we talk about 'responsibility' in the first place. It must further avoid whatever problematic features of our folk concept are identified in meeting the diagnostic challenge. Assessing whether or not revisionism has the tools to adequately meet the prescriptive challenge will be the focus of Chapter 3.

4 So, why be a revisionist?

In light of the three challenges outlined above, one might wonder why anyone might opt to defend revisionism in the first place. It may initially seem that the diagnostic challenge entails that revisionism will be faced with many of the same well established problems that plague conventional theories of moral responsibility. The fact that revisionism must also shoulder the

[22] McKenna (2009a) and Pereboom (2009a) both raise versions of this charge. These objections will be discussed in further detail in Chapter 3.

burden of the motivational and prescriptive challenges might therefore appear reason enough to avoid recommending the view at the outset.

However, there are many features of revisionism which make the view appealing and worth pursuing. First and foremost is the potential for revisionism to reframe the current philosophical debate. While a great deal of progress has been made on many fine-grained issues in the literature on free will and responsibility, many have expressed their frustration at the apparent intractability of this debate more generally. The revisionist distinction between diagnostic and prescriptive accounts of moral responsibility in particular has the potential to initiate a paradigm shift in philosophical thinking about these issues, a shift that might allow the debate to move forward in new and interesting ways.

It seems to me that this potential is itself enough to recommend revisionism. But for those in need of a further push, it might be found in the recent explosion of experimental data on free will and moral responsibility. It is not at all clear what conventional theorists ought to do with this data. If they accept it, it provides a bulk of evidence in favor of the claim that our beliefs, theoretical commitments, and intuitions about moral responsibility are far more diverse, fragmented, and perhaps even contradictory than conventional theorists have thus far assumed. And because the conventional theorist's methodology is heavily grounded in mapping our concept of responsibility, this raises a serious problem. Whatever one's conventional theory of responsibility, it looks as though there will be a great deal of explaining to do regarding conflicting empirical data about the intuitions, doxastic, and conceptual commitments of the folk. For example, conventional compatibilists must explain why there is evidence that under certain conditions people consistently demonstrate incompatibilist commitments.[23] And conventional incompatibilists must explain why there is evidence that under certain conditions people

[23] See Nichols & Knobe (2007).

consistently demonstrate compatibilist commitments.[24] This is no easy task for a theory that is intended to provide the best account of our beliefs, theoretical commitments, and intuitions about moral responsibility as they stand.[25]

Revisionism, on the other hand, is uniquely equipped to accept this data as it comes (warts and all) and deem some of the particular beliefs and commitments that it identifies as unworthy of our continued rational commitment without any widespread attribution of error. For revisionists, it is not that these problematic beliefs and commitments *get things wrong* in some deep metaphysical sense, but rather that we *ought not to rationally go on maintaining them*. And this makes revisionism very appealing indeed.

5 Where we're going

Before concluding this chapter it is helpful to provide a brief summary of the overall goals of the project to follow, and how I plan to address them.

First I raise a new problem for revisionism, *the normativity-anchoring problem*. The heart of this problem is that the methodological commitments (in particular, commitment to the skeptical claim) used to motivate revisionism and distinguish the view from conventional theorizing about moral responsibility make it uniquely difficult for revisionists to justify our continued participation in the practice of moral praising and blaming. Solving this problem is a necessary step in laying the groundwork for any successful revisionist account.

Second, I argue that revisionists can ultimately avoid the normativity-anchoring problem via appeal to a principled difference between the epistemic status of some of our judgments about moral responsibility and others. More specifically, appeal to this principled difference can

[24] See Nahmias (2006, 2011), Nahmias et al. (2005, 2006), and Nahmias et al. (2007).
[25] The data generated by recent experimental work in this area and its relevance to the overall revisionist project will be discussed in much greater detail in Chapter 4.

be used to ground a qualified methodological assumption that I argue revisionists can and should accept, an assumption that allows revisionists to preserve the overall motivation for the view while avoiding the normativity-anchoring problem. I conclude that revisionism is a live option in the philosophical debate, and provides a new and potentially fruitful methodological approach to theorizing about moral responsibility. In making the normativity-anchoring problem explicit and showing that it does not constitute an insurmountable problem for revisionism, I also hope to have dispelled many misguided worries and confusions about the view, and ultimately to have left it better situated in the overall debate.

I address these goals as follows. In the next two chapters I focus on clearly articulating what revisionism is, what a successful revisionist view requires, and how the normativity-anchoring problem arises. In Chapter 2 I provide a detailed discussion of the three main challenges for revisionism mentioned above – the diagnostic, motivational, and prescriptive challenges – and what meeting each of these challenges requires. I argue that Vargas' brand of revisionism successfully meets the first two of these challenges, but not the third. At the end of Chapter 2 I discuss in particular Vargas' proposed criteria for revisionist prescriptive theory construction. In Chapter 3 argue that Vargas' own brand of revisionism fails to generate a prescriptive account capable of meeting these criteria, leaving his view open to the normativity-anchoring problem.

In Chapters 4 and 5 I lay the groundwork for diffusing the normativity-anchoring problem. Because I take this problem to arise largely out of commitment to the skeptical claim, I defend a qualified methodological assumption that I argue revisionists can and should accept (hereafter referred to as MAP). This assumption is based on identifying a principled difference between some of our judgments about moral responsibility and others. Chapter 4 provides a

survey of the wealth of recent empirical work on laypersons' judgments about moral responsibility which suggests that a variety of individual factors influence these judgments. In Chapter 5 I focus on the influence of one of these factors in particular, concreteness. At the outset of Chapter 6 I argue that we have good reason to think that concreteness plays an enabling role in generating competent judgments about moral responsibility. In the remainder of Chapter 6 I argue, via a series of companions-in-guilt style arguments, that we are justified in accepting MAP. I then present and asses some potential objections to my arguments, and conclude that none of them bar our acceptance of MAP. Finally, I make explicit how acceptance of MAP allows revisionists to avoid the normativity-anchoring problem and take stock of where this project leaves revisionism in comparison to its conventional competitors.

Conclusion

In this chapter I first provided a brief taxonomy of prominent views in the contemporary free will debate traditionally classified in terms of responses to the compatibility question, and discuss where revisionism stands in regards to them. Revisionism is like eliminativist (or "no free will either way") views in that it largely sidesteps the compatibility question. However, unlike these views revisionism is a success theory that provides an account of how beings like us can, at least sometimes, be morally responsible for our actions. In Section 2 I present the basic contours of revisionism. One of the defining features of revisionism is that it makes a distinction between diagnostic and prescriptive accounts of moral responsibility, a distinction motivated by a commitment to the skeptical claim about the proper role of our intuitions in responsibility theorizing.

In Section 3 I briefly discuss what I take to be the three main challenges for revisionism: the diagnostic, motivational, and prescriptive challenges. A successful revisionist theory of moral responsibility must meet all three of these challenges. In light of the fact that meeting them all may seem a daunting task, in Section 4 I sketch two important considerations that recommend revisionism. I take these considerations alone make revisionism worth pursuing, though there are likely many others that I do not mention here. Finally, in Section 5 I provide an overview of what is to come in the following chapters, and outline the primary goals of my project as a whole.

In the next chapter I turn to discussion of how revisionism can meet the diagnostic, motivational, and prescriptive challenges. I will ultimately argue that these challenges can be met, but that success depends on divorcing revisionism from a commitment to the skeptical claim.

Chapter Two

Meeting the Three Challenges

Introduction

This chapter focuses on the details of Vargas' revisionist account of moral responsibility and how it deals with the three challenges laid out in the previous chapter. I will begin with the diagnostic challenge, and discuss Vargas' arguments for the claim that our folk concept of moral responsibility is significantly libertarian, and that these libertarian features are deeply problematic. I then turn to the motivational challenge of showing that revision rather than elimination is called for. This discussion focuses on two distinct paths to elimination, which I will call strong and weak elimination. I will use Galen Strawson's (1993b) impossibilist view as a model for the former and Derk Pereboom's (2001) hard incompatibilism as a model for the latter. I appeal to Susan Hurley's (2000) powerful argument against strong eliminativism and to Vargas' (2013) arguments for revision over weak elimination. However, the success of these arguments in motivating revision over elimination depends in part on whether or not the prescriptive challenge can be met. In Section 3 I discuss the details of Vargas' proposed desiderata for a successful prescriptive account of moral responsibility, as well as his suggested methodology for revisionist prescriptive theory construction.

While I accept the arguments presented in Section 1 and Section 2 and take Vargas' brand of revisionism to be well equipped to handle the diagnostic and motivational challenges I will argue in Chapter 3 that, as it stands, it cannot meet the prescriptive challenge.

1 Meeting the diagnostic challenge

In order to assess whether our folk concept of moral responsibility is in need of revision we must first have some idea of what our folk concept of moral responsibility *is*. Here it is helpful to say something about what it means to talk about our 'folk concept' of responsibility in the first place. Vargas characterizes it as follows:

> ...depending on your favorite view of concepts, we are talking about the broad overlap of semantic, representational, causal, or inferential structures, structures which themselves permit of an array of tokenings whose precise content is more and less elaborated. (2009, 50)

Vargas himself avoids taking a strong stance on both the nature of concepts and whether or not our folk concept of moral responsibility is, for example, univocal or fragmented.[26] And he does not attempt to provide anything like a comprehensive diagnostic account of the concept of moral responsibility. Instead, he restricts his diagnostic project to determining whether or not there are important features of the concept that are deeply problematic, and concludes that there are. In particular, Vargas cites what he takes to be strong evidence that our concept of moral responsibility is significantly incompatibilist and, moreover, libertarian. He argues that these libertarian features are deeply problematic, and should therefore be jettisoned from our conceptual commitments.

In Section 1.1 I discuss the various considerations which Vargas argues provide strong evidence that our folk concept of moral responsibility is significantly libertarian. In Section 1.2 I discuss arguments for the claim that these libertarian features of our folk concept are deeply problematic, and thus should be given up. In Section 1.3 I present and assess some objections to this attempt to meet the diagnostic challenge, and conclude that it is ultimately successful.

---

[26] Though he admits that recent empirical data suggests that it may in fact be fragmented (2009, 48). I will discuss the question of whether or not Vargas can really avoid taking a stance on these issues in greater detail in Chapter 3. For arguments that the concept is univocal see Nelkin (2005). For arguments that it is fragmented, see Doris, Knobe, & Woolfolk (2007) and Feltz & Cokely (2009).

1.1 Vargas' arguments for three strands of libertarian folk thinking

Vargas (2009, 2013, Fischer et al. 2007) cites three distinct kinds of consideration in favor of a significantly libertarian diagnostic account of free will. While each provides only defeasible evidence that our folk concept is libertarian, when taken together these considerations make a libertarian diagnosis look quite plausible.

Here it is important to note that this is not to say that Vargas is claiming that the considerations he cites show that our folk concept is *entirely* libertarian. For Vargas and revisionists more generally the aim of the initial diagnostic project is to provide good reason for thinking that a *significant* part of our shared concept of moral responsibility is problematic and that these features can and should be jettisoned, while leaving enough of the overall concept intact to justify our continued use of the term 'moral responsibility'. This point again raises questions about concept individuation. I will discuss whether or not Vargas' brand of revisionism is saddled with any particular view of concept individuation and what I take to be the implications of such a commitment in greater detail in Chapter 3. For now I will simply further echo Vargas' own attempt to clarify what he intends in his use of the term 'folk concept' as it pertains to meeting the diagnostic challenge:

> Of course, if some or another commitment is just an element of one person's thinking and no one else, that commitment is poorly suited to make a claim on being our shared concept. But if you can easily get lots of people—or university undergraduates, at any rate—to vigorously agree with you under suitable conditions, then it looks like you've got a candidate for a real part of one's conceptual commitments. Alternately, we might say such a process yields data about prima facie contributors to semantic content. (2009, 49)

What Vargas seems to be interested in here is identifying particularly problematic features of our widely shared conceptual commitments about moral responsibility. If such features can be

identified then we will have done the work required to move from the diagnostic to the motivational challenge, and can then examine whether the concept in question should be revised or eliminated. Vargas himself identifies what he considers widely shared *libertarian* conceptual commitments, and argues that these commitments are problematic enough to move us to either revision or elimination.

The first type of consideration that Vargas appeals to is data from recent experimental work on ordinary attributions of free will and moral responsibility.[27] The data generated from this empirical work is far from conclusive, and does not appear to support any general claims about whether or not our folk concept of free will is entirely and always compatibilist or incompatibilist. But there is some consensus that this data does indicate that abstract, general cases involving agents and their actions tend to elicit predominantly incompatibilist responses from ordinary subjects.[28] Furthermore, when asked whether or not our universe is deterministic or indeterministic, an overwhelming majority of subjects say that our universe is indeterministic (Nichols & Knobe 2007, 669).[29] So, Vargas concludes that this experimental data provides evidence that at least a significant portion of our folk thinking about free will and responsibility is libertarian. This evidence may be defeasible, but it nonetheless provides support for the claim that at least some of our conceptual commitments regarding responsibility are in fact libertarian.

The second kind of consideration in favor of thinking that our folk concept is significantly libertarian is the naturalness and intuitiveness of various philosophical arguments in favor of libertarianism, in particular van Inwagen's (1983) Consequence Argument and Pereboom's (2001) Four Case Manipulation Argument. While Vargas does not claim that such

[27] For a more detailed discussion of this data and its relevance to the overall revisionist project see Chapter 4.
[28] See, for example, Nichols & Knobe (2007), and Nahmias, Coates, & Kvaran (2007).
[29] According to their data, over ninety percent of participants judged that a minimally described indeterministic universe was more similar to our own than a minimally described deterministic universe.

arguments necessarily show us anything about the metaphysical structure of the world, their intuitive force suggests that they capture an important aspect of our conceptual commitments regarding free will and moral responsibility. So, they provide at least some evidence (though, again, defeasible) that our folk concept is significantly libertarian.

The third kind of consideration Vargas appeals to is cultural and historical (2009, Fischer et al. 2007). Popular Western religious commitments employ a conception of free will that relies on the ability for agents to do otherwise. In particular, the Christian conception of free will historically employed to explain the existence of evil relies on this ability. In addition, the long historical tradition of dualism supports a conception of agency not connected to determinism. When these historical considerations are combined with the phenomenology of decision-making, Vargas argues that the idea that we would, over time, have formed a libertarian conception of free will is at least reasonable. While it is of course possible that our commonsense concept could have developed independent of these cultural and historical threads, he claims that it would be "naïve and wholly unrealistic to think that they have" (Fischer et al. 2007, 139).

Taken together these considerations provide strong, though defeasible evidence that our folk concept of moral responsibility is, to some extent, libertarian. However, this is not to say that it is *entirely* libertarian. None of the above considerations show that significant strands of compatibilist thinking do not also play a role in our folk thinking about moral responsibility. It may be the case, as mentioned at the start of this section, that our folk concept is fragmented, and may have contradictory compatibilist and incompatibilist elements. Or, it may be the case that our concept of moral responsibility is contextual, and has different applications depending on particular situational differences.[30] However, all that is needed to motivate revision is the identification of some significant feature of the concept that is deeply problematic and should be

[30] I will discuss these possibilities in detail in Chapter 4.

given up. And the family of considerations cited by Vargas provides good prima facie reason for accepting the presence of libertarian elements in our folk thinking. Going forward I will assume that Vargas' arguments for the claim that there are such elements are persuasive, and that our folk concept of moral responsibility is (at least in part) libertarian.[31]

1.2 The naturalistic implausibility of libertarianism & the argument from fairness

Vargas argues that the fact that our folk concept of moral responsibility appears to be significantly libertarian is problematic for two related reasons. First, these libertarian elements are naturalistically implausible. Second, this implausibility raises serious worries regarding fairness.

I will begin with the standard of naturalistic plausibility. In broad strokes, Vargas takes a significantly libertarian concept of moral responsibility to be naturalistically implausible because it does not appear to provide a picture of human agency compatible with a broadly scientific view of the world (Vargas 2009, 51). More will be said about naturalistic plausibility in Section 4, but it is helpful to make a few remarks on what it amounts to here.[32] The naturalistic plausibility of a particular view of moral responsibility depends directly on the kinds of demands that the view places on the results of future science. For example, a specific libertarian account might require that indeterminacies show up at specific times and locations in the brain. This places a demand on the results of future neuroscience.[33] If we currently have no scientific evidence for such indeterminacies, and no good reason to think that we will discover them as neuroscience progresses, then the requirement that they must exist places a burden on a view that

[31] For those who disagree I will return to the question of whether or not we should accept Vargas' diagnosis at the end of this section.
[32] Here, again, I borrow from Vargas (2013, Fischer et al. 2007).
[33] Pereboom (2001) argues that Kane's (1996, 2007) event-causal libertarian view makes this kind of demand.

posits them which reduces its naturalistic plausibility. All else being equal, the more scientifically demanding the view, *ceterus paribus*, the less naturalistically plausible it is. This is not to say that a current lack of scientific evidence in support of a view *necessarily* counts against it. Only that, if we value naturalistic plausibility, where there is no definitive scientific evidence for or against a view we should assess the degree and specificity to which it requires future science to turn out a certain way and take this into consideration when assessing the merits and costs of the view.

Vargas argues that, at the very least, current neuroscience fails to provide evidence in support of an indeterministic picture of the brain (Fischer et al. 2007, 145). And, even if this lack of evidence is taken to be indicative only of current shortcomings in the science, he cites a tendency among neuroscientists to think that any picture of the brain that is indeterministic to the extent capable of grounding libertarian free will would be problematic.[34] So, there is reason to think that libertarian free will does not give us a picture of agency that is consistent with a naturalistic view of the world.

Furthermore, libertarianism is far more metaphysically demanding than its competitors. Specific libertarian accounts of free will give rise to a variety of striking demands. For example, agent causal theories require that agents act as physically uncaused causes, and so posit a unique form of causation. On the other hand, many event causal theories require as yet undiscovered indeterminacies to show up in the brain at just the right times and in just the right places.[35] It would be miraculous if we were to discover that these indeterminacies *do* show up just where

[34] For further discussion of this point see Vargas in Fischer et al. (2007). Here I admit a large degree of ignorance about the relevant scientific literature, and intend only to point out that Vargas makes this claim. However, whether or not he is right about this seems to be more or less irrelevant to his overall argument. This argument relies only on the claim that there is currently no *positive* neuroscientific evidence supporting libertarianism and that, in combination with the demands of the theory, this makes libertarianism naturalistically implausible.

[35] See Kane (1996, 2007).

these libertarian theorists hoped they would.[36] So, while a libertarian view may be *possible* and coherent, it is far more metaphysically demanding and thus more *implausible* than its competitors. When combined with the fact that there is at the very least no good scientific evidence to support the claim that we are free in the sense required by libertarianism, these further considerations provide good reason to think that if the correct diagnostic account of our concept of moral responsibility is significantly libertarian, then it is deeply problematic.

Before moving on it may be helpful to make a few additional remarks about the standard of naturalistic plausibility as it is presented by Vargas.[37] First, it should be contrasted with a more widely accepted standard of naturalistic *compatibility*, which requires only that a theory avoid contradicting our best scientific worldview (Vargas 2013, 46). The standard of naturalistic plausibility that Vargas as in mind, on the other hand, requires that (all else being equal) there is something that *speaks in favor* of a theory beyond its mere coherence with this worldview. Second, the type of thing that might count as "speaking in favor of" a theory should be restricted to truth-relevant considerations, in particular those that have a direct connection to whether or not the theory is correct (2013, 46). So, things like emotional appeal or coherence with other theories one happens to value will not speak in favor a theory in the sense relevant to naturalistic plausibility. Third, this standard might be interpreted in either a *threshold* or *scalar* way (2013, 47). On the first understanding of naturalistic plausibility we might stipulate a particular measure that the theory must meet in order to count as naturalistically plausible. For example, we might say that if there are *any* truth-relevant considerations that speak in favor of a theory then it will count as naturalistically plausible. This threshold notion of naturalistic plausibility

[36] Vargas focuses much of his discussion on event causal theories like Kane's. Pereboom (2001) makes the related point that it would be miraculous if the analogous scientific discoveries necessary to support agent causal libertarianism were made (for example that our motivating reasons and desires *just happened* to coincide with our indeterministic free choices).

[37] Vargas (2013) discusses this standard at length in Chapter 2.

can be contrasted with a *scalar* notion, where a particular theory will fall somewhere on a spectrum of plausibility. This understanding of naturalistic plausibility lends itself best to comparative assessments of plausibility between particular theories of responsibility. According to Vargas, libertarianism fails to meet the standard of naturalistic plausibility on both the threshold and scalar notions.[38]

Vargas' second line of argument that the libertarian elements of our concept of moral responsibility are deeply problematic is related to the first and appeals to the notion of fairness. Even if there are arguments to be made against the claim that it is implausible to think we are free in the sense required by libertarianism, the fact that we currently have no good scientific evidence to support the claim that we are free in this way raises a further worry regarding our current practice of moral praising and blaming. It looks as though the justification for this practice presupposes a particular kind of agency. Attributions of praise and blame are only appropriate in regards to subjects that are in fact free and responsible. So, if our concept of moral responsibility is such that *we have no good evidence for thinking that we are ever really responsible*, it seems we are left merely hoping that our practice of praising and blaming is justified (Vargas 2009, 52). And this leaves open the possibility that this practice results in radical and widespread unfairness. Most importantly, it leaves open the possibility that we often undeservedly punish and inflict *harm* upon those who are not in fact free and responsible. So, our libertarian folk concept of moral responsibility leaves us – at best! – hoping that we are really responsible and thus that our practices are justified. And this provides good reason to jettison the libertarian features of the concept.

To sum up, there are several reasons for thinking that a concept of moral responsibility that is significantly libertarian is deeply problematic and in need of either revision or elimination.

[38] For detailed arguments see Vargas (2013), Chapter 2.

First, it is naturalistically implausible. There is at best no good evidence in favor of the claim that we are free and responsible in this way, that it is likely that we are, or even that it is *possible* for us to be. Furthermore, even if this lack of evidence is indicative only of current neuroscientific shortcomings, the metaphysical demands that a libertarian concept of moral responsibility places on the results of future science are comparatively high. So, again, a libertarian conception is naturalistically implausible. And if these arguments are correct then the naturalistic implausibility of the libertarian features of our folk concept also raises a further worry regarding fairness. If our practice of moral praising and blaming is based on a libertarian concept of moral responsibility that requires powers or abilities we have no good reason to think that we have or could ever have, then there is a good chance that this practice is unjustified. Given that the practice of praising and blaming often results in punishment and harm, this conclusion strongly recommends that we either give up our libertarian conceptual commitments, or eliminate the concept of moral responsibility entirely.

1.3 Objection – the wrong diagnosis?

Before moving on to the motivational challenge one might wish to raise an obvious objection to the above arguments: what if this diagnosis gets things wrong?[39] At the very least there has been a recent explosion of empirical data on this topic, some of which casts doubt on the intuitiveness of incompatibilism.[40] Vargas admits that overwhelming evidence that our folk thinking about moral responsibility is not significantly libertarian would be a serious problem for the *revisionist* nature of his account. However, it need not constitute a serious threat to the overall project:

[39] This type of objection to revisionism has been raised by Michael McKenna (2009a).
[40] In particular, see Nahmias et al. (2006), Nahmias et al. (2007) and Nahmias (2011).

> Still this concern may prove to be only superficial. For any self-described revisionist account that proves to not be revisionist, we would still have a substantive prescriptive account that merits consideration on its own terms. If the best self-described revisionist compatibilist account turns out to be nothing more than an excellent conventional compatibilist account, this would, I suspect, hardly dismay the account's proponent. (Vargas 2011, 469)

Here Vargas' point seems to be that this objection is merely terminological. If the diagnostic account outlined in this section gets things wrong and our folk concept is *not at all* libertarian this need not affect the overall success of the so-called revisionist project of tackling the motivational and prescriptive challenges. The only difference is that this project should proceed under the title of conventional compatibilism rather than revisionism.[41] Furthermore, if our folk concept is not at all libertarian, then Vargas' arguments for the implausibility of libertarianism are arguments against holding certain beliefs and theoretical commitments about moral responsibility that we (luckily) do not hold currently. While such arguments would surely be less provocative, they still might prove valuable in articulating why *prescriptively* we ought not to allow libertarian elements to creep into our folk thinking about responsibility, and why we ought to be skeptical about any practices that presuppose such elements.

I agree with Vargas in thinking that this objection threatens only what we should call the view in question, not the view itself. So, I think that revisionists can set it aside and proceed. This is in part because a complete diagnostic account of our shared concept of responsibility (to the extent that there is one) as it stands is ultimately an *empirical* project. It is not a matter of getting the right theory, but of identifying what our actual conceptual commitments are. As such, expecting revisionists to provide knock-down arguments for the claim that our concept of

[41] For example, the revisionist-turned-conventional-compatibilist might still argue that libertarianism is implausible, that eliminativist views are under-motivated, and that our best, all things considered prescriptive account of moral responsibility is a compatibilist one.

moral responsibility is significantly libertarian would impose an unreasonable burden on the view. It seems that the best we can do here is to make use of the existent empirical data, and provide independently good reasons for thinking that there are such libertarian strands in our folk thinking. And the considerations which Vargas identifies, in combination with the data, look to adequately meet this task. Perhaps future empirical work will show that these claims are wrong, and it is in fact a mistake to think that our concept of moral responsibility is at all libertarian. The jury is, after all, still out. However, it is not likely that empirical work on this subject will yield the kind of comprehensive results needed to show this anytime soon.[42] And even if it did Vargas' rejoinder that the objection then faced by revisionists would merely be a terminological one stands.

Furthermore, I do not put much stock in even the terminological worry. I find the minimal claim that Vargas is making here – that *to some degree* our folk concept of moral responsibility is libertarian – to be relatively obvious. The complexity of recent empirical work on this topic alone lends a great deal of support to the idea that there are a diverse number of elements that influence our folk thinking about moral responsibility, and that at least some of them are libertarian. In short, it seems that a lot of people a lot of the time tend to think like libertarians. Many philosophers writing on free will and moral responsibility will agree with this claim without much argument at all. Most incompatibilists, for example, already believe it. But even those who don't (those who are firm in their compatibilist convictions) often attempt to explain why this way of thinking is mistaken, rather than deny that we ever think this way in the first place. So, while revisionists can easily respond to the objection that they have provided a mistaken diagnostic account by granting that what they are defending is really a conventional

[42] See Chapter 4 for further discussion of various conflicts in the data generated by empirical work on our commonsense thinking about moral responsibility

compatibilist (rather than a genuinely revisionist) view, the fact that Vargas' particular diagnostic account is relatively uncontroversial makes a need for such concession seem unlikely.

There is, however, a stronger objection to the diagnostic account outlined in this section. One might reject Vargas' claim that libertarianism is implausible. There are after all many widely respected libertarian views. While these views face many serious objections (luck, incoherence, etc.), a great deal of impressive philosophical work has been devoted to responding to them. In light of this fact, there is a much bigger dialectical problem lurking for Vargas' diagnostic account. While it seems likely that revisionism might gain traction with some conventional compatibilists, particularly those who wish to maintain their compatibilist accounts of moral responsibility while avoiding widespread attributions of error to those with libertarian intuitions, it seems highly unlikely that revisionism will have much pull for libertarians. A libertarian is far more likely to point to some error in the revisionist's diagnostic account and their arguments that libertarianism is implausible. After all, many philosophers have taken on the project of arguing that libertarianism is implausible, and these arguments have as yet failed to be decisive.[43]

Revisionists have a quick response to this objection available. The arguments for the implausibility of libertarianism in Section 1.2 do not *depend* on the much discussed objections and arguments against libertarianism already on the table. The force of such arguments is of course an additional resource that revisionists might appeal to in arguing that libertarianism is implausible, but it is not their only resource. Vargas' claims regarding the naturalistic implausibility of libertarianism and the resultant fairness problem constitute a new and very serious worry for libertarianism. Currently there is no adequate libertarian response to these

[43] Many opponents of libertarianism will of course disagree here. Such disagreement is all the better for revisionism.

worries. Is such a response possible? Perhaps, but in lieu of any such response revisionists are warranted in concluding that the libertarian strands of our folk thinking about moral responsibility are deeply problematic and should thus be given up.

2 Meeting the motivational challenge

If the arguments above are correct then the best diagnostic account of our folk concept of moral responsibility is that it is significantly libertarian, and that this is deeply problematic. This conclusion leaves us with two options. First, we might attempt to eliminate the concept of moral responsibility from our beliefs and theoretical commitments, along with any practices that presuppose that we sometimes act as genuinely responsible agents. Alternatively, we might revise our concept of moral responsibility in a way that jettisons the problematic libertarian elements while leaving the concept itself intact. So, in order to motivate their position revisionists are in need of arguments for why revision rather than elimination is appropriate.

This section focuses on providing these arguments. In Section 2.1 I distinguish between two distinct paths to elimination: strong and weak eliminativism. In Section 2.2 I discuss a powerful argument against the former raised by Susan Hurley (2000). In Section 2.3 I discuss Vargas' (2013) arguments against the latter. I conclude that these arguments are sufficient to motivate revision over elimination, though this conclusion comes with a caveat. In order to fully meet the motivational challenge it must be possible for revisionism to meet the prescriptive challenge and provide a tenable prescriptive account of moral responsibility. The details of the methodology for providing such an account will be discussed in Section 3, and making explicit the difficulties involved in meeting this challenge will be the subject of Chapter 3.

2.1 Strong v. weak eliminativism: impossibilism and hard incompatibilism

As discussed in Chapter 1, there are a variety of eliminativist or "no free will either way" views on moral responsibility. Here it is helpful to distinguish between two prominent paths to eliminativism. First, strong eliminativist views maintain that we should eliminate moral responsibility from our set of beliefs and theoretical commitments because it is *impossible* for beings like us to ever act as morally responsible agents. Galen Strawson's (1993b) impossibilism is a prominent example of this kind of view. Weak eliminativist views, on the other hand, are those that either remain agnostic about or grant that it is theoretically possible for beings like us to sometimes be responsible for our actions, but that it is highly unlikely that we ever actually are. In light of this doubt and the fact that our reactive attitudes and the practice of moral praising and blaming are warranted by the assumption that we are in fact sometimes responsible for our actions, we ought to eliminate our concept of moral responsibility and determine which of these attitudes and practices can be maintained without appeal to responsibility. Derk Pereboom (2001, 2009a) most rigorously defends this kind of eliminativist view in his defense of hard incompatibilism.

Revisionists must show that revision is preferable to both kinds of eliminativism, and this requires both positive and negative arguments. On the negative side, revisionists must explain why we should not be either strong or weak eliminativists. On the positive side, they must show that revisionism is a tenable alternative to elimination, despite the fact that our folk concept of moral responsibility is significantly libertarian and thus deeply problematic. Meeting this latter requirement depends on whether or not revisionists can meet the prescriptive challenge and provide a tenable prescriptive account of moral responsibility free of these problematic libertarian features. I discuss the revisionist tools for addressing this challenge in Section 3. In

this section I focus on the negative aspect of meeting the motivational challenge. I will look first at a specific argument against impossibilism raised by Susan Hurley (2000). I will then turn to arguments against accepting hard incompatibilism raised by Vargas, which appeal to a methodological principle of conservatism.

2.2 Against strong elimination

Susan Hurley (2000) addresses the general question of when elimination rather than revision of a particular entity or property is warranted, and examines the connection between this question and particular views about essence.[44] She concludes that while some of these views are less hospitable to elimination than others, eliminativist views that appeal to impossible essences face serious problems regardless of one's general theory. These problems are made clear in regards to Strawson's impossibilist view of moral responsibility in particular.

Hurley distinguishes between *context-driven* and *theory-driven* accounts of essence. Whether an account is context-driven or theory-driven depends on whether the essence of the thing in question is determined by the contexts in which it has been applied along with its causal history, or by some theoretical or conceptual role assigned to it (2000, 230).[45] So, if the essence is context-driven:

[44] If you dislike talk of essences an analogue to Hurley's argument can also be run in terms of semantic internalism and externalism. Perhaps it is more charitable to interpret Strawson's impossibilism as a view about what we think and mean when we use the term 'moral responsibility,' rather than what is *essential* to responsibility. On the one hand, if what we think and mean when we use the term 'responsibility' is determined to some extent by facts about our environment and the outside world (and not just intrinsic facts about us) then it is not at all clear how we could come to think and mean that the term refers to a logically impossible property. On the other hand, if what we think and mean is all "in the head" it is also unclear how we would come to use a term with a logically impossible extension in the first place. In short, it seems that many of the arguments that follow here could also be run in terms of internalist and externalist semantic theories, rather than context and theory-driven views of essence.

[45] Whether an account is context or theory-driven can (and likely will!) differ for different kinds (Hurley 2000, 230).

> …we can be very wrong in our theoretical descriptions of a given entity or kind. We can discover surprising things about what is essential to that stuff we've been talking about. (2000, 230)

On the other hand, if the essence is theory-driven:

> …we can be very wrong in our applications of a term. We can discover to our surprise that nothing occupies the theoretical role essential to the entity or kind in question, that there is no such thing as what we took ourselves to be talking about. (2000, 230)

At first glance, context-driven accounts of essence appear to be more hospitable to revision than to elimination. When the essence of an entity or property is fixed by our use of the term that denotes it and its causal history our very use of the term makes the claim that we ought to eliminate that entity or property from our ontology puzzling when we *do* in fact use it. Likewise, theory-driven accounts appear initially to be more hospitable to elimination than revision. If the essence of an entity or property is fixed by the theoretical or conceptual role that it plays for us, then it is unclear how one might argue that we ought to continue talking about that thing if we discover that there's nothing that plays the role in question. So, while revisionist views often tend to assume, implicitly or explicitly, that a context-driven account is appropriate, eliminativist views often tend to assume that a theory-driven account is.

Hurley does not intend the distinction between context- and theory-driven accounts to be exhaustive, and acknowledges that the connection between an account of essence and whether or not elimination or revision is appropriate for a particular entity or property is complex. While there is an initial appearance that context-driven accounts are more hospitable to revision and theory-driven accounts are more hospitable to elimination, it is possible to argue for revision or elimination against the background of either theory. However, Hurley argues that this is not so

for particular brands of elimination, in particular strong eliminativist views that posit impossible essences for the kind, entity, or property in question.

In developing this argument Hurley targets a specific case for elimination by way of appeal to an impossible essence, Galen Strawson's impossibilism about free will and moral responsibility (1993b). Strawson's view can be understood as a strong eliminativist position as it goes beyond the claim that human beings like us never instantiate the property (or properties) essential to moral responsibility, making the stronger claim that it is *impossible* for us to ever instantiate them. Strawson argues that a regression condition is essential to responsibility, and that in order to be responsible one must be responsible not only for the proximate cause of one's action, but also for *its* causes and so on, "all the way back" (Hurley 2000, 247). According to Strawson, meeting this condition requires an infinite regress of self-determination which is logically impossible. So, beings like us are never responsible, nor could we be. This argument, which Strawson calls the "Basic Argument," goes as follows:

> (1) Nothing can be *causa sui* – nothing can be the cause of itself. (2) In order to be truly morally responsible for one's actions one would have to be *causa sui*, at least in certain crucial mental respects. (3) Therefore nothing can be truly morally responsible. (1993b, 5)

The key premise in the above argument is (2), the claim that being *causa sui* is a necessary condition for moral responsibility. In support of this claim Strawson states that this feature of moral responsibility "has for a long time been central to the Western religious, moral, and cultural tradition," and that it is "a natural part of the human moral-conceptual repertoire"

(1993b, 8-9). Strawson concludes that, because meeting this condition is impossible, elimination of moral responsibility from our "conceptual repertoire" is appropriate.[46]

Hurley argues that it is difficult to see how a call for elimination based on appeal to an impossible essence could be motivated on any account of essence, because it is unclear how appeal to an impossible essence might get off the ground in the first place. First, if one is operating within a context-driven view of essence then essences have explanatory depth and must do the relevant explanatory work in relation to our contexts of use (2000, 236). So, when it comes to responsibility, on a context-driven view the essence of responsibility must go some way towards explaining our actual attributions of moral responsibility, and why in some cases we hold people responsible and in others we do not. But if the essence of moral responsibility is impossible to instantiate, then it simply cannot do the explanatory work required. A necessary condition for responsibility that we do not and *cannot* ever satisfy cannot explain these attributions. As such, this condition is not a plausible candidate for the essence of responsibility on a context-driven account. Here one might object that this argument should not trouble the eliminativist, because context-driven accounts of essence actually make elimination impossible, and so the eliminativist will have already assumed a theory-driven account. However, it is not the case that elimination is impossible given a context-driven account. Hurley points out one possible way in which a context-driven account might still recommend elimination: if the essence of the kind in question turns out be grue-like, with no casual or functional unity (2000, 237).

[46] Though Strawson elsewhere argues that we may be *unable* to eliminate responsibility from our conceptual apparatus, and addresses the question of whether or not doing so would have serious negative consequences if we could (1993a).

But then how does elimination by way of appeal to an impossible essence fare on a theory-driven account? According to Hurley, theory-driven accounts of essence also require explanatory depth, though in relation to the theory itself rather than the contexts of application:

> Explanatory depth within a theory-driven account would relate to the theory itself. It would have a coherentist character. A subset of the properties the theoretical role assigns to the kind *F* may do better than any other subset at preserving the internal coherence and point of the theory. Such explanatory depth has a theory-internal normative and justificatory dimension. (2000, 239)

But it is clear that an impossible essence cannot do this explanatory work either. If the point of a theory of moral responsibility is to provide an account of when attributions of praise and blame are warranted, then a property that is logically impossible for us to possess clearly will not do the best job of preserving the internal coherence of that theory. Again, the point here is not that elimination could never be motivated on a theory-driven account. It could turn out that the property or properties that best explain our responsibility system are never actually instantiated, in which case responsibility would turn out to be much like phlogiston. The problem for elimination driven by appeal to an impossible essence is that it is not clear how, if it really is essential to responsibility that something must be *causa sui*, our theory of responsibility ever got off the ground in the first place. How, for example, might the phlogiston theorist have ever come up with a theory of phlogiston in the first place if they considered it essential to phlogiston that it instantiate some logically impossible property, for example the property of being both a substance and not a substance?

So, it looks as though neither a context- nor theory-driven account of essence is hospitable to eliminativist views motivated by appeal to an impossible essence. Hurley acknowledges, however, that the distinction between context- and theory-driven accounts is not exhaustive. In addition to a straightforwardly context- or theory-driven view one might also take

a meaning-driven approach to essence, or something akin to a reflective equilibrium approach that requires trade-offs between context- and theory-driven considerations (2000, 234). Hurley sets aside the tenability of a meaning driven approach in light of worries regarding disagreement and skepticism about the analytic/synthetic distinction (2000, 233). And it is unclear how adopting a reflective equilibrium approach might avoid inheriting the same problems for impossible essences that plague both context- and theory-driven approaches. Therefore, in lieu of some alternative approach to essence, attempts to motivate elimination by appeal to impossible essences looks deeply problematic on any picture of essence. So, Strawson's argument for impossibilism does not support elimination over revision, and when it comes to moral responsibility attempts to motive strong elimination fail.

Before moving on to arguments against weak elimination, it is helpful to address one potential objection to Hurley's argument against strong elimination. The main thrust of Hurley's argument is that *impossible essences can't do the relevant explanatory work* on any view which requires that essences have explanatory depth (2000, 242). It is important to make clear that the kind of impossibility at issue here is either metaphysical or logical. A property that is merely nomologically impossible to instantiate might still be considered a *possible* essence, capable of doing the relevant explanatory work. Hurley illustrates this point with the following example: let's say we are interested wizards, and think that it is essential to being a wizard that one has magical powers. The fact that it is impossible, in a world like ours, for beings to have magical powers does not mean that it is logically or metaphysically impossible for beings to have such powers. We can imagine possible worlds in which they do. If, in such worlds, having magical powers does the relevant explanatory work regarding who counts as a wizard or to a theory of wizard-kind, then having magical powers is a possible essence for wizards. But we cannot

assess the relevant counterfactuals for being *causa sui* because being *causa sui* is *logically* impossible. As such there are no possible worlds we can imagine where we might assess whether something being the cause of itself does the relevant explanatory work for our theory or attributions of responsibility, because there are no possible worlds where we can imagine beings that are *causa sui*.

Here one might take issue with an assumption that Hurley seems to be making about impossible worlds, namely that reasoning about such worlds is not theoretically useful because all claims about them are trivially true. But some, for example Daniel Nolan (1997), have argued that we can make sense of the idea that claims about impossible worlds are not just trivially true. In fact, Nolan argues that they can be quite useful when reasoning about possibility. If this is correct then why think that we cannot reason fruitfully about what would be the case in logically impossible worlds in which beings are *causa sui*? While a detailed response to this objection is well beyond my current purposes, I think it is helpful to briefly outline a potential response on behalf of Hurley: even those who take impossible worlds to be theoretically useful do not go so far as to claim that *all* impossible worlds are theoretically useful. In particular, while we may be capable of principled reasoning about comparatively "close" impossible worlds, this does not mean that we are capable of such reasoning when it comes to those that are especially distant. Perhaps there is some fact of the matter about what would be the case in worlds where *I* was born in the 18th rather than the 20th century.[47] But intuitions about what would be the case in a world like this are far clearer than intuitions about what would be the case in a world in which all the logic books are false.[48] The point here is that Hurley can grant that in some cases appeal to impossible worlds can be theoretically useful. But

[47] Of course, you would only take this world to be metaphysically impossible if you accept Kripke's (1980) view of the necessity of origin.
[48] If we accept something like Lewisian (1979) similarity conditions between possible worlds.

it seems reasonable and relatively uncontroversial that their usefulness declines as one moves from "nearby" to very distant impossible worlds. And *logically impossible* worlds, like those in which there are entities acting as self-caused causes, are a long way off indeed. Thus any appeal to facts of the matter about what would be the case in such worlds should be, at best, viewed with a healthy amount of skepticism. In particular, there seems little reason to think that our judgments about whether or not particular properties do the relevant explanatory work in such worlds are at all reliable or informative.[49]

This concludes discussion of Hurley's arguments against strong elimination. To sum up, the prospects for motivating strong elimination over revision in this way look grim.

2.3 Against weak elimination

Vargas' arguments against weak elimination are based largely on appeal to what he calls a *principle of philosophical conservation*:

> **PPC:** we ought to abandon our standing commitments only as a last resort. When we do abandon our commitments, there is pressure to minimize the consequences, limiting the scope of revision or elimination. (2013, 62)

Vargas asserts that this principle is widely accepted (implicitly or explicitly), and that whatever rational authority it might have derives from the principles that govern our actual mechanisms for belief formation and retention (2013, 64). The idea is that, for finite beings like us, widespread revisions in our beliefs will likely have a large impact on the stability of our overall doxastic commitments. And the stability of our overall doxastic commitments is important for a number of reasons. So, without some large degree of pressure to revise those commitments they have a kind of "doxastic inertia" (2013, 64). According to Vargas, PPC is likely why many view

[49] For further discussion of these issues see Williamson (2007).

eliminativist philosophical views as radical positions in the first place. If PPC is correct then it is appropriate to view them with a healthy amount of skepticism.

Eliminativism about moral responsibility in particular - even weak eliminativism - will likely have a substantial impact on our doxastic commitments. It would force us to jettison a number of our beliefs, beliefs about not just the status of agents we had previously judged responsible, but also about the justification for the overall practice of praising, blaming, and punishing. Take, for example, the potential impact of accepting Derk Pereboom's brand of weak eliminativism, hard incompatibilism. According to Pereboom only agent-causal versions of libertarianism can provide a tenable account of moral responsibility under which agents are ever genuinely responsible. But there are strong reasons for thinking that our actions are not agent-caused events. So, there is good reason to think that we are never morally responsible for our actions (Pereboom 2001, 128). Given the acceptance of this conclusion, there are many doxastic commitments that hard incompatibilists might be forced to give up. For example, they may be forced to jettison the natural view of ourselves as agent causes, the belief that people are (at least sometimes) praiseworthy when they perform morally good actions and blameworthy when they perform morally wrong actions, our general acceptance of the principle 'ought implies can,' and retributivist forms of punishment.[50] These examples are by no means exhaustive. Hard incompatibilism might also raise difficulties for non-consequentialist moral commitments, our view of ourselves as rational deliberators, and certain conceptions of achievement and self-worth. While Pereboom argues that hard incompatibilists can preserve many of these commitments, the issue turns largely on whether or not the commitment in question requires that

[50] Pereboom discusses the potential consequences of adopting a hard incompatibilist position in much greater detail in Chapters 5-7 of *Living without Free Will*. While he makes a compelling case that many of these commitments can in fact be preserved, he also grants that this is not true for all of them, in particular the belief that agents sometimes genuinely deserve moral praise and blame.

we view ourselves as agent causes or that we take others to genuinely deserve moral praise and blame. Determining which of our doxastic commitments depend on these beliefs is a difficult question to assess, as is determining the overall cost of giving up the particular commitments that do.[51] I will not discuss these issues further here, but mentioning them should suffice to make clear that, regardless of how the details get sorted out, hard incompatibilism will likely require some widespread revision of our overall web of doxastic commitments. At the very least, it will require jettisoning all those that depend on the belief that we sometimes genuinely deserve moral praise and blame. So, in light of PPC even a weak eliminativist view about moral responsibility ought to be adopted only as a last resort.

This is not to say that a weak eliminativist view of moral responsibility like hard incompatibilism is *wrong* or untenable, merely that the changes to our overall web of doxastic commitments that it would require recommends that we ought to pursue all other options before accepting this kind of view. And revisionism is one such option. Revisionists can accept many of Pereboom's arguments, especially his arguments that libertarianism is implausible. But before these arguments move us to eliminate moral responsibility altogether, PPC recommends that we first explore whether or not revision is possible. So, here the prescriptive challenge looms large.

To sum up, given the diagnostic account of moral responsibility discussed in Section 2, we are left with two options. First, we might adopt an eliminativist view of moral responsibility and attempt to give up our concept of moral responsibility entirely, along with any practices that depend on it. Second, we might try to retain the overall concept by jettisoning only the libertarian features that are naturalistically implausible and problematic. There are two ways that one might motivate the first option. They might pursue the strong eliminativist route, and claim

[51] For recent discussion of the cost of giving up the reactive attitude associated with blameworthiness – resentment or moral anger – see Nichols (2007) and Pereboom (2009b).

that it is impossible for agents to ever be responsible. Hurley's arguments provide good reason for thinking that this path to elimination fails. Alternatively, one might pursue the weak eliminativist route and argue that while it is possible that creatures like us could sometimes be morally responsible for our actions, there is good reason to think that we never are. We should therefore eliminate our concept of moral responsibility and explore which of our related attitudes and practices can be preserved without it. But, if we accept that PPC or some principle like it underlies a great deal of our philosophical theorizing, then we have good reason to view weak eliminativism with a healthy amount of skepticism. We should pursue whether or not there are tenable theoretical options available that do less violence to our overall web of doxastic commitments before endorsing elimination. Revisionism is one such option, and so we must turn to what a tenable revisionist account of moral responsibility might look like.

A great deal therefore hangs on whether or not revisionism can meet the prescriptive challenge. I turn next to Vargas' account of how this might be done, focusing in particular on the desiderata he proposes for a successful prescriptive account and the details of his methodology for revisionist prescriptive theory construction.

3 Meeting the prescriptive challenge

Without a prescriptive account of moral responsibility revisionists cannot successfully motivate their view. So, if the prescriptive challenge cannot be met revisionism cannot get off the ground. Thus far only Vargas has tackled the challenge of proposing a clear set of criteria for a revisionist prescriptive account, as well as a methodology for revisionist prescriptive theory construction. This section outlines the details of his proposal, and of the specific prescriptive account that he offers.

3.1 Theory construction & methodology: systematic revision[52]

There are several basic criteria that must be met in order for a revisionist prescriptive account of moral responsibility to count as tenable. Vargas suggests the following[53]:

(C1) It must provide justification for the responsibility system.[54]

(C2) It must be naturalistically plausible.

(C3) It must be normatively adequate.

First, meeting (C1) is best understood as requiring an answer to two questions. The first involves the justification of the responsibility system external to facts about the system itself. I will call this question the *external question*:

> Is there anything that would, in general, justify our participation in practices of moral praising and blaming? (Vargas 2013, 128)

The second question regards the practices, attitudes, and judgments that make up the responsibility system as we find it. I will call this question the *internal question*:

> Can we explain our customary patterns of assessment in ways that make it plausible that they are tracking normatively relevant features of agents in the world? (Vargas 2013, 128)

In order to meet (C1) a revisionist prescriptive account of responsibility must provide answers to each of these questions. In doing so, this account will thus license our ongoing participation in the responsibility system, as well as tie the system to whatever normatively relevant features of agents ground our interests in it in the first place.

---

[52] For discussion of the distinction between this kind of revision and what Vargas calls "revisionism on the cheap" or "repurposing revisionism" see Fischer et al. (2007, 152) and Vargas (2013, 92).

[53] These criteria are first laid out explicitly in Fischer et al. (2007, 153-155) and are developed in much greater detail in Vargas (2013, Chapter 4).

[54] Vargas defines this system as follows

> Taken as a whole, the responsibility norms and their attendant social practices, characteristic attitudes, and paradigmatic judgments constitute what we can call *the responsibility system.* (Fischer et al. 2007, 154)

Here I will skip over any in depth discussion of (C2), which can be found above in Section 1. In order to assess (C3) and determine the degree to which an account is normatively adequate Vargas (2013, 95-96) makes explicit five additional sub-requirements:

> (S1) *Integration*: the account should explain the relationship between a theory of moral responsibility and broader philosophical accounts of morality.
>
> (S2) *Distinctiveness*: the account should say something about the distinctive normative structure of responsibility.
>
> (S3) *Justification*: the account should explain how the responsibility system can be justified.[55]
>
> (S4) *Relevance*: the account should explain how the features of agency and the aspects of the responsibility practices invoked have something to do with what's at stake in our attributions of responsibility.
>
> (S5) *Conservatism*: other things being equal, we should favor accounts that do a minimum of violence to other independently plausible normative notions.[56]

It is important to point out that it is not clear what the relationship (if any) between these sub-requirements is, or if they are to be ranked in any particular order of importance. Here I assume that they are not, and that assessing the normative adequacy of a prescriptive account of responsibility requires attending to whether or not the theory in question meets all of these requirements taken as a whole, to some intuitively sufficient degree. For example, insofar as a particular theory provides some prima facie explanation of the relationship between responsibility and some of our most basic normative notions (such as fairness, rightness, wrongness, etc.), of how responsibility is distinct from these notions, of why we care and should continue to care about responsibility, of why we take the agential features it marks out as

[55] It is unclear whether or not this sub-requirement is also met if (C1) is, or whether it places some further explanatory burden on the account of justification required.

[56] For example, the principle 'ought implies can.' For further discussion of the relationship between this principle and responsibility see Haji (1999).

essential to responsibility to themselves be valuable, while remaining compatible with other independently plausible normative notions, then such a theory would count as normatively adequate. This is of course just one way in which a theory of responsibility might count as normatively adequate. These sub-requirements do not rule out the possibility that, all things considered, the overwhelming appeal of a particular account's potential for meeting, say, the integration and justification sub-requirements might allow that the theory is normatively adequate, even though it violates the conservatism sub-requirement and requires us to abandon one or two of our other independently plausible normative notions. I take these sub-requirements to offer something more akin to prima facie guidelines than independent, necessary features of a normatively adequate account of responsibility.

With these guidelines in mind, how might one go about actually constructing a revisionist prescriptive account of responsibility? Here Vargas is also instructive. First, we must identify what the responsibility system actually is. We must get clear about what it is we are looking to provide an account *of*. Second, we must look to the "internal logic and structure" of that system (Vargas 2013, 94). In Vargas' terms, this requires identifying the *work of the concept* of responsibility. He most recently defines the work of the concept as:

> …the characteristic roles played by the collection of beliefs, commitments, and distinction-making characteristic of moral responsibility. (2013, 99)

With an idea of the work of the concept in mind, we then have two choices. First, we might decide to overturn the framework of our responsibility system. Vargas argues that we would only have good reason to choose this option if the work of the concept itself looks problematic for some very powerful, independent reasons (if, for example, one has good reason to think that the correct metaethical theory stands in direct conflict with this concept). In lieu of such reasons, there is a second option. We can provisionally accept the framework of our responsibility

system and our identification of the work of the concept, and provide a prescriptive account of moral responsibility capable of grounding and explaining it. And we do this by providing answers to the internal and external questions discussed above, answers that are naturalistically plausible and normatively adequate.

3.2 Vargas' prescriptive account

Before turning to what I take to be a serious problem for Vargas' brand of revisionism in the next chapter it will be helpful to provide at least a rough sketch of some of the distinguishing features of the particular prescriptive account that he argues can meet all of the criteria outlined above. First, Vargas defends an *agency cultivation model* of moral responsibility:

> What would justify our responsibility characteristic practices, those that emerge in our holding one another responsible, is if these practices fostered a distinctive form of agency in us, a kind of agency sensitive to and governed by moral considerations. (2013, 177)

What does most of the justificatory work for Vargas' own prescriptive account is that our responsibility practices promote a particular kind of independently valuable agency: *moral reasons-responsive agency.* As such he borrows from a long (though, as he admits, somewhat sordid) historical traditional and defends a view that might be classified as a *moral influence theory* (2013, 168). One defining feature of this kind of view is appeal to the idea that praise and blame get creatures like us to behave in certain desirable ways. What justifies the responsibility system on a moral influence theory is the fact that this system has certain desirable consequences. And it is precisely this feature that gives Vargas' prescriptive account the power

to explain why we care about moral responsibility and why we are justified in continuing to participate in the practice of moral praising and blaming (2013, 168).[57]

While the details of Vargas' agency cultivation model are not relevant to the primary objection I will be raising against his brand of revisionism in the next chapter, it may be helpful to identify a few additional important features of the view. First, Vargas' view can be characterized as a Reasons account according to which effective self-directed agency and the capacity to recognize and respond to moral reasons are jointly sufficient for moral responsibility (2013, 206). Vargas largely sets aside the question of what counts as effective self-directed agency and says only that:

> Among the features of agency I will take as implicated in self-directed (but not yet responsible) agency are such things as beliefs, desires, means-ends reasoning, the ability to formulate and execute action plans, and the presence of ordinary epistemic abilities, including a general capacity for some degree of foresight regarding the consequences of actions. (2013, 207)

He focuses instead on the capacity to recognize and respond to moral reasons, ultimately defining what is required for moral responsibility as follows:

> An agent *S* is a responsible agent with respect to considerations of type *M* in circumstances *C* if *S* possesses a suite of basic agential capacities implicated in effective self-directed agency (including, for example, beliefs, desires, intentions, instrumental reasoning, and generally reliable beliefs about the world and the consequences of action) and is also possessed of the relevant capacity for (A) detection of suitable moral considerations *M* in *C* and (B) self-governance with respect to *M* in *C*. Conditions (A) and (B) are to be understood in the following ways:
>
> (A) the capacity for detection of the relevant moral considerations obtains when:

[57] However, this feature of the account also leaves him with the burden of explaining why he is not committed to the unpalatable consequence that praise and blame are merely forward-looking attempts to influence the behavior of others often attributed to moral influence theories of responsibility. Vargas (2013, Chapter 6) takes great care to address this worry, identify the merits of this kind of view, and diffuse several prominent objections. However, because these arguments are not relevant to my overall purpose, I will not discuss them in further detail here.

(i) *S* actually detects moral consideration of type *M* in *C* that are pertinent to actions available to *S* or

(ii) in those possible worlds where *S* is in a context relevantly similar to *C*, and moral considerations of type *M* are present in those contexts, in a suitable number of those worlds *S* successfully detects those considerations.

(B) the capacity for volitional control, or self-governance with respect to the relevant moral considerations *M* in circumstances *C* obtains when either

(i) *S* is, in light of awareness of *M* in *C*, motivated to accordingly pursue courses of action for which *M* counts in favor, and to avoid courses of action disfavored by *M* or

(a) *S* detects moral considerations of type *M*, and

(b) in virtue of detecting *M* considerations, *S* acquires the motivation to act accordingly, and

(c) *S* successfully acts accordingly. (2013, 221-222)

Furthermore, he intends "suitability" and "relevant similarity" in the above definition to be cashed out in the following way:

> …the notions of suitability and relevant similarity invoked in (A.ii) and (B.ii) are given by the standards an ideal, fully-informed, rational, observer in the actual world would select as at least co-optimal for the cultivation of our moral reasons-responsive agency, holding fixed a range of general facts about our current customary psychologies, the cultural and social circumstances of our agency, our interest in resisting counterfactuals we regard as deliberatively irrelevant, and given the existence of genuine moral considerations, and the need of agents to internalize norms of action for moral considerations at a level of granularity that is useful in ordinary deliberative

> and practical circumstances. Lastly, the ideal observer's determination is structured by the following ordering of preferences:
>
> (1) that agents recognize moral considerations and govern themselves accordingly in ordinary contexts of action in the actual world
>
> (2) that agents have a wider rather than narrower range of contexts of action and deliberation in which agents recognize and respond to moral considerations. (2013, 222-223)

Again, the details of Vargas particular account of responsible agency are not relevant to the normativity-anchoring problem, and so here I intend only to identify some of the most important features of this view. I take these features to be the following: it is a moral-influence version of a Reasons account of responsibility according to which self-directed agency and the capacity to recognize and respond to moral reasons (as defined above) are necessary and jointly sufficient for responsible agency.

Conclusion

The goal of this chapter has been to show how revisionism might meet the diagnostic, motivational, and prescriptive challenges. Looking forward, I take the arguments in Section 2 and Section 3 to provide good reason for thinking that revisionism can successfully meet the diagnostic challenge, and the negative requirement for meeting the motivational challenge. Whether or not the positive requirement can be met depends on whether or not it is possible to provide a tenable prescriptive account of moral responsibility. I present Vargas' proposed criteria for such an account as well as his proposed methodology for revisionist prescriptive theory construction in Section 4. However, in the next chapter I argue that Vargas' own brand of

revisionism fails to meet these criteria. In particular, commitment to the skeptical claim prevents his prescriptive account from meeting the external aspect of (C1).

Chapter Three

Anchoring a Revisionist Account of Responsibility

Introduction

This chapter focuses on two potential problems for revisionist attempts to meet the prescriptive challenge: *the reference-anchoring problem* and *the normativity-anchoring problem*. I take anchoring objections to revisionism discussed recently in the literature to be versions of the reference-anchoring problem. In Section 1.1 I discuss these objections, why I take them to be versions of this particular anchoring problem, and outline what I take to be a successful revisionist response. However, this discussion raises questions about whether or not revisionists must take on costly theoretical burdens regarding concept individuation. In Section 1.2 I argue that they do not.

Focus on the reference-anchoring problem and objections to revisionism regarding concept individuation have thus far allowed what I take to be a far more serious worry for revisionism, the normativity-anchoring problem, to go largely unnoticed. In Section 2.1 I distinguish the normativity-anchoring problem from the reference-anchoring problem, and argue that the former raises serious worries for revisionism as it has been formulated thus far. In Section 2.2 I discuss why this problem is unique to revisionism, outline some potential revisionist strategies for responding to it, and why these strategies are not available to revisionism as it has been formulated thus far. I conclude that any form of revisionism which follows Vargas in his commitment to the skeptical claim is in deep trouble. And without a response to the normativity-anchoring problem the potential for revisionism to meet the prescriptive challenge looks grim. Furthermore, if revisionists cannot meet the prescriptive

challenge then they cannot motivate revision over elimination, and the view looks to be a nonstarter.

1.1 The reference-anchoring problem

Versions of what I call the reference-anchoring problem have been raised in the literature by both Michael McKenna (2009a) and Derk Pereboom (2009a). Broadly, this problem might be understood as the worry that there is no guarantee that a revisionist prescriptive account of 'moral responsibility' will in fact be a genuine account of responsibility. There are two different ways of characterizing this worry. First, one might argue that the reference of 'responsibility' is and can only be picked out by our folk concept.[58] As such, any alteration of the concept to the extent that revisionism is likely to prescribe will entail that we are no longer talking about genuine responsibility. So, in providing an account of revisionist-responsibility and calling it 'responsibility' the revisionist is changing the subject. Second, even if the revisionist is clear about precisely what they mean by 'responsibility' there is a further worry that what the revisionist means is not what we actually care about when we talk about responsibility. So, the revisionist is not only changing the subject, but is not justified in doing so.

McKenna (2009a), for example, points out that comparisons Vargas attempts to draw between responsibility and other more widely accepted examples of conceptual revision are not analogous. Unlike these widely accepted historical examples (for example, our conceptual revision of water), it is not clear that we can *discover* facts about responsibility in the same way that our discovery of the atomic structure of water led us to revise that concept.[59] To suppose

[58] This claim need not depend on an assumption about the correct theory of reference more generally, but only on the assumption that an internal theory is appropriate to 'responsibility' in particular.

[59] Here one might object to the claim that we should categorize the latter as a genuine case of conceptual revision, and that this is an unfair comparison. If this is correct, then so much the better for revisionism. However, there is a

that we can revise our concept of responsibility is to assume a realist position about responsibility, and suppose that there is something that responsibility *is* beyond our concept of it. But it is not clear that responsibility is like this. And even if we were to grant that there is some real nature of moral responsibility beyond our concept, the fact that this real nature does not appear to be discoverable in the same way that the atomic structure of water is makes it difficult for revisionists to explain why we are licensed to call this thing (rather than properties and agential features picked out by our folk concept) 'responsibility.' In changing the reference of 'responsibility' it looks as though revisionists are changing the subject, and it is not clear that they are licensed to do so.[60]

Pereboom (2009a), on the other hand, grants that it might be possible to revise our concept of responsibility. And, if the revisionist can accomplish this then they need not be accused of changing the subject. However, even if it makes sense to talk about revising our folk concept of responsibility Pereboom presses the question of whether or not the revised concept will be

> near enough to the folk's to count as a natural extension of it, one that can do enough of the work the folk conception does in adjudicating questions of moral responsibility and punishment, and in governing our attitudes to the actions of those around us? (Pereboom 2009a, 25)

---

related, deeper worry that any examples of conceptual revision that revisionists use to help motivate their view might best be understood as examples of changes in our widely shared beliefs about these things, not genuine revision of our shared concepts. Here I again provide the reader with a promissory note that I will discuss these issues in further detail in the next section.

[60] Here I use the example of the purported conceptual revision of water only because it is the example that McKenna himself uses to motivate his arguments against revisionism. But not much hangs on this particular example or the fact that water has empirically "discoverable" features which responsibility seems clearly to lack. Vargas uses many other examples (such as the apparent revision of our concept of marriage), and one might run a version of McKenna's argument against revisionism using an example that has a more *a priori* flavor. For example, the concept of continuity was revised due to *a priori* mathematical considerations, and it is not immediately obvious what the analogous *a priori* considerations in favor of revising our concept of responsibility might be either. Thanks to Kris McDaniel for suggesting this example.

If the answer to this question is negative, then it looks as though the revisionist has gone beyond altering merely our conception, and the resulting view calls for us to use the same term to stand for distinct concepts. But this threatens to result in serious confusion and miscommunication, and the same worries that McKenna raises about changing the subject become salient.

However, if the answer to Pereboom's question is affirmative, then it looks as though the revisionist has succeeded in altering only our conception while maintaining the overall concept, giving them license to use the term 'responsibility.' So, a revisionist response to the reference-anchoring problem depends on their ability to answer this question affirmatively, and tie their prescriptive account of responsibility near enough to the folk's. This would not only address Pereboom's worries, but likely McKenna's as well. McKenna's primary concern is that it seems that:

> …what moral responsibility is cannot come apart from the concept in such a way that there is, so to speak, something for moral responsibility to be beyond our concept of it. (2009a, 11)

But if the revisionist can successfully tie their prescriptive account of responsibility near enough to the folk's then it looks as though revisionism need not require that responsibility *come apart* from our concept after all. In jettisoning problematic features of the concept we need not change the subject entirely. We might still retain enough of the important features of that concept to warrant the claim that we are still talking about responsibility, that is, whatever it is that grounds our attributions of moral praise and blame.

The essential question, then, that the reference-anchoring problem raises for revisionists is the following: is their prescriptive account of moral responsibility tied, in some way, closely enough to the folk concept to license their use of the term 'responsibility'?

There is obviously some ambiguity here regarding what will count as 'closely enough.' However, the contours of revisionist theory construction outlined in Chapter 2 provide strong

reason for thinking that a revisionist prescriptive account of responsibility will in fact be tied closely enough to the folk concept. Recall that a prescriptive account of responsibility must, according to the criteria for constructing such a theory, explain how the theory is relevant to what we care about and take to be at stake in our attributions of responsibility. In constructing a prescriptive account of responsibility revisionists begin by identifying the responsibility system as we find it, and the work of the concept. In earlier work Vargas provides some further detail in characterizing the work of the concept of responsibility:

> The most useful initial characterization of the conceptual role for moral responsibility is as something that plays an important role in our organization, coordination, and justification of differential treatment of one another. In particular, it is connected to praiseworthiness and blameworthiness. In turn, judgments of praiseworthiness and blameworthiness underwrite a web of emotional reactions, judgments, and social practices that can include (but are not limited to) reward and punishment. (Fischer et al. 2007, 154)

This account of the work of the concept sounds quite similar, if not identical, to the features of the folk concept that Pereboom claims revisionists must tie their view closely to. The work of the concept is, at least in part, to regulate our judgments about when agents deserve praise and blame. So, it looks as though raising the reference-anchoring problem for revisionism in the first place misses the point of the overall revisionist project. Not only will revisionists who accept Vargas' characterization of the work of the concept tie their prescriptive account of responsibility to our folk thinking about moral responsibility, they will accept it as one of the most basic features of and starting point for generating their account.

Vargas, for example, provisionally accepts the framework of the responsibility system that the work of the concept grounds, and takes the overall revisionist prescriptive project to be that of justifying our participation in that system and explaining the customary patterns of

inference which make up the system itself. The project of justifying and explaining may result in jettisoning some particular features of our folk concept as it currently stands, but not those that are essential or constitutive. If this is the correct picture of the best revisionist methodology for theory construction, then it is unclear how one might uphold the charge that revisionists are changing the subject when it seems that an affirmative answer to Pereboom's question is built into their methodology.

Finally, Vargas at least is willing to grant that his brand of revisionism might in fact shift the reference of 'moral responsibility,' but argues that if this is so the resulting dispute is merely terminological. He distinguishes between two possible kinds of revisionism, *connotational* and *denotational* revision (2011, 462). According to the former, the beliefs associated with moral responsibility that revisionism prescribes jettisoning "do no substantive work in designating some property in the world," and so do not affect the reference of the term (2011, 462). According to the latter, revisionism does prescribe giving up "some reference-fixing content" (2011, 462). However, so long as there is "some nearby property" very much like responsibility which

> …preserves the primary inferential roles we take to organize our beliefs…regiments our practices and characteristic attitudes in familiar ways….weighs in our deliberation in just the same way….

and preserves the same "normative import," then this shift in reference is warranted (Vargas 2011, 462). So, even if one turns out to be a denotational revisionist the charge of changing the subject will be merely terminological, as whatever the denotational revisionist is referring to after this shift will respect the most fundamental work of the concept. In light of this, the denotational revisionist might even counter reference-anchoring objections by challenging those who raise this kind of objection to explain why we ought to care about what we *have* been referring to, rather than the nearby *moral responsibility**, at all.

So, it seems that revisionists can respond to versions of the reference-anchoring problem like those raised by McKenna and Pereboom. So long as they adopt Vargas' methodology for constructing a prescriptive account of moral responsibility and accept his characterization of the work of the concept, it seems they get this response for free.

1.2 Revisionism and concept individuation

By now it should be apparent that the reference-anchoring problem raises a series of issues regarding how concepts are to be individuated. The heart of the problem is the charge that the revisionist is changing the subject and that what they are talking about is not genuine moral responsibility, but something else entirely. Definitively settling the question of whether or not this charge is warranted will likely depend on a worked out theory of concept individuation. Unfortunately, there is little to no consensus on what such a theory might look like in the contemporary concepts literature.[61]

The *classical theory* – that the structure of concepts is definitional – has been more or less abandoned, due largely to the fact that the theory looks to be inconsistent with a large body of empirical data in psychology.[62] In its place at least three competing theories have emerged. First, the *prototype theory* is the view that lexical concepts have probabilistic structure. Second, according to the *theory theory* concepts stand in relation to one another in the same way that terms in scientific theory do, and so are inter-defined and are individuated at least in part by their conceptual role.[63] Finally, *conceptual atomism* stands in stark contrast to all of these views in

[61] For a helpful survey of this literature see Laurence & Margolis (1999).
[62] In particular, it conflicts with the data on typicality. For a survey of this data, see Murphy (2002).
[63] For further discussion of a recent version of the theory theory view see Carey (2009).

that concepts are individuated by their relation to the world alone (not to other concepts).[64] This is by no means an exhaustive list of the views one might hold on the nature and structure of concepts (for example, one could also be an eliminativist or pluralist), but is a helpful, if coarse-grained way of dividing up the terrain.

There are at least three relatively uncontroversial desiderata for any theory of concepts, and none of the above views do a particularly good job of meeting them all. Ideally, a view of concepts should be able to account for the following phenomena:

(1) Concepts are sharable.

(2) Concepts are appropriately fine-grained.

(3) Concepts allow for compositionality.

A successful view of concepts should have the tools to explain how different individuals (and the same individual over time) are able to share the same concepts. It must make sense of the fact that when I use the concept CAT, and a distinct individual uses the concept CAT, we are interested in the same thing. A successful view of concepts should also be appropriately fine-grained in the sense that it can account for the strong intuition that HESPERUS and PHOSPHORUS or WATER and H2O are distinct concepts, even though they refer to the same thing. So, it must allow for intensional differences between terms with the same referent. Finally, a successful theory of concepts should allow for and explain how we compose complex concepts out of simpler constituent concepts. If it does not then the theory will face serious difficulties explaining how concept acquisition and conceptual development are possible.[65]

---

[64] For example, according to Fodor (1990, 1998) a concept stands in a lawful relation to the property it expresses, where other lawful relations involving the concept are symmetrically dependent on the relation between the concept and the property it expresses. Other lawful relations *depend* on the lawful relation between the concept and the property it expresses, and would not hold without that relation.

[65] One might object that (3) makes conceptual atomism a nonstarter. However, (1)-(3) are intended only as *desiderata* for a theory of concepts, not necessary requirements for a successful theory. In general, we want a theory of concepts to allow for compositionality because we want the theory to account for concept acquisition and

While the prototype, theory theory, and conceptual atomist views described above all do a particularly good job of meeting one or more of these desiderata, they each do a notoriously bad job of meeting them all. Versions of the prototype theory can account for sharability and fine-grainedness, but fail at compositionality. Versions of the theory theory can handle fine-grainedness and compositionality well, but face serious worries regarding sharability because of their commitment to *holism*, the idea that concepts must be individuated by attending to their inferential role in its entirety.[66] Conceptual atomists can account nicely for sharability, but are saddled with providing an account of concept acquisition that does not appeal to compositionality, and seem committed to an unsatisfactorily coarse-grained account.[67] Given these tradeoffs it is not at all clear which theory of concepts might win the day, and so there is little hope of settling whether or not the above response to versions of the reference-anchoring problem succeeds via appeal to any consensus in the concepts literature.

Vargas himself is aware of this difficulty, and attempts to avoid taking a stance on the nature of concepts altogether. In order to see how, we might here extend his appeal to the distinction between connotational and denotational revision. If one's particular view of concept individuation entails that what revisionists are really interested in is RESPONSIBILITY* rather than RESPONSIBILITY, then it will be true that revisionists and conventional theorists are not talking about the same shared concept. If this is the case, then revisionists must take great care

---

learning. That the conceptual atomist is in a difficult position regarding this particular desideratum does not rule the view out automatically, given that its main competitors are in an equally difficult position with regards to either (1) or (2). Furthermore, the conceptual atomist will likely appeal to a combination of some degree of nativism about concepts, and an alternative story about concept acquisition involving particular sustaining mechanisms and mind-world relations in order to explain why they need not allow for compositionality. For an example of a particular positive conceptual atomist account of concept acquisition see Margolis (1999, 559-564).

[66] Proponents of the theory theory need not embrace holism, but might opt instead for some form of *molecularism* (the view that only some *part* of the inferential role of a concept is relevant to its individuation). However, theory theorists who go this route must make appeal to some version of an analytic/synthetic distinction, which most readers will immediately reject as hopeless.

[67] For example, this kind of view looks to yield the counterintuitive result that the concepts of HESPERUS and PHOSPHORUS are the same. In short, it is not clear that conceptual atomism leaves room for intensionality.

not to overlook it. Furthermore, under these circumstances it might also be misleading to call the view in question *revisionist*. Rather, if the prescriptive account provided by the revisionist is actually an account of RESPONSIBILITY* then the view might more accurately be categorized as a version of *eliminativism*. However, this would not trivialize revisionism. The view would still stand in sharp contrast to other eliminativist views such as Pereboom's (2001, 2009a) and Galen Strawson's (1993) in its prescriptive recommendation that we *replace* our concept of RESPONSIBILITY rather than simply give it up. So, again, the question of whether or not revisionism prescribes conceptual change or conceptual revision according to one's particular theory of concepts amounts to merely a terminological one.

However, there is a further question about whether or not Vargas is really entitled to sidestep any commitment to a particular view of concepts. In various discussions of the work of the concept Vargas seems to use this term and the "conceptual role" of moral responsibility interchangeably.[68] For Vargas, it looks as though the conceptual role of responsibility is what determines the widely shared semantic content of the concept RESPONSIBILITY, and what ensures that we are talking about the same thing, or something relevantly similar, when we use the term 'responsibility'. If preserving the work of the concept is primarily what Vargas is beholden to in constructing his prescriptive account and this is how the work of the concept is to be characterized, then it looks as though he is therefore committed to a view about concepts that takes inferential role to be at least partially constitutive. So, it looks as though he is committed to some kind of theory theory view of concepts.

At first glance this may seem like a cost to revisionism. There is little to no consensus that the theory theory is correct, and it rules out a number of competing views that do not take

[68] For example, see Vargas (2013, 99-101) and in Fischer et al. (2007, 154),

inferential role to be constitutive. However, there are at least four considerations that count against thinking that there is a significant cost to this commitment.

First, if commitment to some version of a theory theory view of concepts is a cost to revisionism then it counts equally against conventional theories of responsibility. The inferences that Vargas points to as constitutive of the concept of moral responsibility are the same inferences that conventional theorists pick out. Contemporary compatibilists, incompatibilists, and even hard incompatibilists point to precisely the same basic features of the inferential role of the concept RESPONSIBILITY (whatever does the work of warranting our attributions of moral praise and blame) to ground the claim that they are not talking past one another.

Second, use of the term 'inferential role' here is misleading. In the concepts literature talk of inferential role is often associated with a commitment to holism, the view that the massive web of *all* of the causal and inferential relations that the concept is embedded in determines how that concept is to be individuated. But this is not what Vargas or his conventional competitors are appealing to. What they are identifying is instead what they take to be the most basic, or essential role of the concept, that of regulating our judgments about when an agent deserves moral praise and blame.

Third, while commitment to a theory theory view rules out other views according to which inferential role plays no constitutive role in concept individuation such as straightforward classical theories, prototype theories, or conceptual atomism, each of these views face their own serious difficulties. As such it is not clear that there is any unique burden on Vargas to justify an appeal to some version of the theory theory beyond the mere desire to avoid unnecessary theoretical commitments wherever possible. Given the relatively controversial status of all of

these views, the theory theory does not saddle Vargas or revisionism more generally with any particularly heavy theoretical burdens.

Finally, even if Vargas is tacitly assuming a theory theory view of concept individuation and it turns out that the theory theory is incorrect, this need not trouble the revisionist. In this case the view of responsibility in question might best be characterized as a version of eliminativism, not revisionism. But again, this point is merely terminological, and revisionism would still be a unique and interesting position in its prescriptive recommendation that the eliminated concept can successfully be *replaced*.

Here I wish only to point out that I am skeptical that the reference-anchoring problem for revisionism can be settled by appeal to a particular view of concept individuation. Nor am I persuaded by Vargas' own assertions that revisionism can avoid any specific commitments regarding the nature of concepts and how they are individuated. Vargas instead appears tacitly committed to some version of the theory theory, but this commitment should not be considered a significant cost to revisionism.

This concludes discussion of the reference-anchoring problem. I turn now to what I take to be a far more serious problem for revisionism, the normativity-anchoring problem.

2 The normativity-anchoring problem

I have argued that revisionism can overcome the reference-anchoring problem, but this problem has a far more troubling normative cousin. The heart of the problem is this: even if it is built into revisionists' methodology that what they are providing a prescriptive account of is something that we *do* deeply care about, there is a further question about whether or not we *should*. We might grant that their prescriptive account captures the work of our concept of

responsibility and the features of responsibility that we actually take to be worth wanting (while departing from our folk concept in some significant way), but why think that this is an account of responsibility that we ought to endorse? What grounds the normativity of this prescriptive account, and does the actual *prescribing*? This question is closely tied to the external question discussed in Chapter 2, the question of whether or not a prescriptive account can justify our continued participation in the responsibility system and general practice of moral praising and blaming. It does not look as though Vargas' own prescriptive account can, and his acceptance of the skeptical claim discussed in Chapter 1 makes it uniquely difficult for revisionists to provide such justification.

According to Vargas, the prescriptive account of moral responsibility he offers satisfies the external aspect of (C1) because, on this account, moral responsibility contributes to the cultivation of a particularly valuable kind of agency: moral-considerations responsive agency (2009, 52-53). In particular, what justifies our continued participation in the responsibility system is that

> these practices foster a distinctive form of agency in us, a kind of agency sensitive to and governed by moral considerations. (2013, 177)

But if this is correct then it looks as though the value of our actual responsibility system depends on the value of this kind of agency, and Vargas' prescriptive account might best be interpreted as a kind of *buck-passing* account of moral responsibility. While his view tracks normatively relevant features of agents in the world (as it must if it is to satisfy the internal aspect of (C1)), it is not *facts about responsibility* that provide us with reason to continue to participate in the practice of attributing moral praise and blame, but *facts about whether or not this practice promotes something that is independently valuable*, moral-considerations responsive agency.

But what is wrong with a buck-passing prescriptive account of responsibility? Moral-considerations responsive agency is clearly a unique kind of agency that we do value greatly. However, the fact that we do value this particular kind of agency does not entail that we should continue to participate in any particular practices that promote it, including the practice of moral praising and blaming. In order to see why, it is helpful to distinguish between three distinct claims that might mistakenly be run together:

(1) *Psychological*: we value moral-considerations responsive agency.

(2) *Axiological*: moral-considerations responsive agency is valuable.

(3) *Normative*: we should continue to participate in any practice that promotes moral-considerations responsive agency.

The psychological claim is obviously true. We do indeed value moral-considerations responsive agency. However, Vargas seems to assume that the psychological claim entails the axiological claim, and that the axiological claim entails the normative claim. And because the practice of praising and blaming (and the responsibility system more generally) promotes moral-considerations responsive agency, our continued participation in this practice is justified. But it is not obvious that *any* of these entailments hold, and Vargas is in a particularly difficult position when it comes to arguing that they do.

Take first the connection between the psychological and axiological claims. It is clearly possible to value something that is not in fact valuable. Human beings make mistakes about what is valuable all the time, and not just on an individual level. Might we be mistaken about the value of moral-considerations responsive agency? Perhaps, for example, the correct metaethical view is some kind of error-theory, and so the kinds of considerations that our moral-considerations responsive agency is responsive to are systematically false. It is not at all clear

that under these circumstances this kind of agency would be genuinely valuable, regardless of whether or not we actually value it. At the very least, the move from the psychological claim to the axiological claim requires further argument.

In light of this, Vargas might argue that reasons independent of the psychological claim justify our acceptance of the axiological claim. But what sort of reasons? Vargas cannot appeal to the fact that the axiological claim is intuitive or obvious, since his commitment to the skeptical claim blocks him from reading normative claims off of descriptive facts about our intuitions.[69] For Vargas, the fact that it is possible that our intuitions get things wrong, and that we have no good epistemic reason for thinking that they get things right, is enough to block appeal to them as evidence, at least when it comes to matters normative. But then it is not clear what alternative options are available for supporting the claim that we are epistemically justified in accepting the axiological claim that moral-considerations responsive agency is genuinely valuable.

However, even if we grant that Vargas can offer successful arguments to support the axiological claim, it is support for the normative claim that is required in order to meet the external aspect of (C1). The truth of the axiological claim does provide *pro tanto* reason to accept the normative claim, but that is not enough for the task at hand. What Vargas needs to meet the external requirement of (C1) is to show that our participation in the responsibility system should, *all things considered*, continue. The fact that moral-considerations responsive agency is genuinely valuable and that the responsibility system promotes this kind of agency fails to establish this. One might think that we could devise a new system that could promote the cultivation of moral-consideration responsive agency that made no reference to praise and blame

[69] Vargas makes this explicit in regards to moral responsibility, and it is not at all clear that there is any principled difference between responsibility and other normative concepts capable of providing revisionists more generally with a reason to maintain the skeptical claim in regards to the former but not the latter.

at all. Perhaps, for example, we simply rewarded behavior that promotes this kind of agency and left praise and blame out of the picture entirely.[70] This certainly seems possible, and if it is then continued participation in the responsibility system is not the *only* way to promote this valuable kind of agency.[71] And if the responsibility system is not the only way to promote this kind of agency, then appeal to its value fails to show that we should, all things considered, continue to participate in the responsibility system.

So, appeal to the value of moral-considerations responsive agency fails to do the normative work necessary to successfully meet the external aspect of (C1). The psychological fact that we do value this kind of agency does not entail the axiological claim that this kind of agency is actually valuable. And Vargas' commitment to the skeptical claim places him in a particularly difficult methodological position when it comes to providing independent support for the axiological claim. Furthermore, even if we grant for the sake of argument the truth of the axiological claim, this claim does not entail the normative claim. In particular, the fact that moral-considerations responsive agency is genuinely valuable provides only pro tanto support for the normative claim that we should continue to participate in any practice that promotes this kind of agency. If the practice of moral praising and blaming is not the only way to promote this kind of agency (and it is not) then appeal to the value of moral-considerations responsive agency fails to establish that we should, all things considered, continue to participate in this practice.

Finally, little hangs on Vargas' appeal to the value of moral-considerations responsive agency in particular. The normativity-anchoring problem poses a serious worry for any brand of

---

[70] Pereboom suggests something along these lines in Kane (2011, 417), and argues that moral admonition and encouragement are sufficient to communicate a sense of right and wrong and achieve effective moral education and improvement. However, Pereboom's arguments assume that giving up on praiseworthiness and blameworthiness need not entail giving up on rightness and wrongness, which has been challenged at length by Haji (1998, 2002).

[71] Of course, we could *call* the behavior in question 'responsible' behavior, but to do so would again raise versions of the reference-anchoring problem discussed in Section 1. In this case I am inclined to say that we *would* here be changing the subject.

revisionism that accepts the skeptical claim and offers a buck-passing prescriptive account of responsibility. Commitment to the skeptical claim will make it uniquely difficult for revisionists to offer independent support for the relevant axiological claim, whatever that particular claim may be. Appeals to the intuitiveness of such claims will leave revisionists in, at best, the uncomfortable position of requiring an explanation for why we ought to be skeptical about the epistemic status of our intuitions about moral responsibility, but not other normative concepts. And even if we are willing to grant the truth of the relevant axiological claim this, again, will not do the work necessary to meet the external requirement of (C1). Revisionists must show that the practice of moral praising and blaming is the *only* way to promote the kind of value referenced in the preferred axiological claim they appeal to. Vargas' own attempt to meet (C1) fails because there are in fact other ways to promote moral-considerations responsive agency. And it is not at all clear what alternative facts about independent sources of value revisionists might appeal to such that the practice of moral praising and blaming will be the only way to promote this kind of value. It looks as though the methodological commitments used to motivate revisionism therefore block revisionists from providing a tenable prescriptive account of moral responsibility capable of satisfying the external requirement of (C1).

3 Conclusion

It is important to note that the normativity-anchoring problem is a unique problem for revisionism. Conventional responsibility theorists sidestep this problem entirely by way of their own methodological commitments. According to the conventional theorist, the best account of responsibility (the one that we ought to endorse) just is the one that best captures what we actually think about moral responsibility and what we take to be most important and valuable

about it. If one accepts this assumption then there is no reason to ask why we ought to endorse an account of responsibility beyond the fact that it best aligns with our intuitions about particular cases and our systematically refined beliefs and theoretical commitments as we find them. In granting this assumption one grants that these things are in some sense tracking the truth, and so a theory of moral responsibility generated by this methodology will not need to tell a further story about why it is normatively anchored. The theory will be normatively anchored because, given our standing methodological commitments, it is the account that we have the best reason to think is *correct*.

Revisionists committed to the skeptical claim cannot appeal to these considerations. But commitment to the skeptical claim is precisely what makes revisionism so interesting. It is what motivates the distinction between descriptive and prescriptive accounts of responsibility and distinguishes revisionism from conventional theorizing about moral responsibility. So, it seems that revisionists are in a uniquely difficult position in regards to the normativity-anchoring problem. They must find some alternative way to ground the normativity of their prescriptive account of responsibility, and explain why the account they generate is one that we ought to endorse. However, if the arguments in Section 2 are correct then it is not at all clear that they can.

If the arguments presented in this chapter are correct then revisionism, at least as the view has been formulated thus far, is in serious trouble. While revisionists are well equipped to deal with the reference-anchoring problem, they do not currently have the tools for responding to the normativity-anchoring problem. In the remaining three chapters I will take on the task of providing these tools. In what follows I develop a positive proposal on behalf of revisionists that

I argue can ultimately allow them to avoid the normativity-anchoring problem while preserving the underlying motivation for the view.

Chapter Four

Ordinary Judgments about Moral Responsibility: Experimental Philosophy, Empirical Data, and Individual Influences

Introduction

If the arguments raised in Chapter 3 are correct then the normativity-anchoring problem poses a serious threat to revisionism, and it looks as though a tenable response to this problem is blocked by commitment to the skeptical claim discussed in Chapter 1. But, need revisionists follow Vargas in his commitment to this claim in order to adequately motivate their view? In Chapter 6 I will argue that revisionists can avoid the normativity-anchoring problem by accepting a qualified methodological assumption, one that preserves the spirit of the skeptical claim while acknowledging a principled difference between the epistemic status of some of our judgments about moral responsibility and others. However, before turning to these arguments a great deal of groundwork must first be laid in order to motivate the claim that this principled difference exists. Laying that groundwork will be the focus of this chapter and much of Chapter 5.

Making the case for some epistemically privileged class of judgments about moral responsibility will depend in part on appeal to empirical data. There has been an explosion of empirical data on our ordinary responsibility judgments in recent work in experimental philosophy. I take this data to be not only relevant, but invaluable to those interested in defending a revisionist account of responsibility. In this chapter I provide a survey of some of the most central work on moral responsibility in this area.

In Section 1 I provide some background on experimental philosophy, its methodology, and the underlying motivations for this empirical approach to philosophical issues. Responses to

this approach have varied widely. Regardless of one's initial attitude towards experimental philosophy, I will argue that a large portion of the data it has produced thus far is both relevant and useful to the overall revisionist project. In Section 2 I canvass the main dialectic in experimental philosophy regarding moral responsibility centered on the compatibility question. Interestingly enough, this work has generated conflicting results about whether ordinary folk thinking about moral responsibility is predominantly compatibilist or incompatibilist. In Sections 3 and 4 I discuss several different attempts to make sense of these conflicting results. These attempts range from arguments that the conflict in question is merely apparent, to attempts to interpret these results when taken at face value.

Finally, in Sections 5 I shift focus from the compatibility question to the question of what individual factors influence our judgments about responsibility. I will canvass data suggesting that a variety of individual factors seem to influence these judgments in surprising ways. These influences include the way that cases and vignettes are described, the moral character of the action being assessed, and the potential influence of stable features of the subject making the judgment.

This discussion is intended to lay the groundwork for Chapter 5, where I turn my attention to one particular factor that a large body of empirical evidence suggests influences our responsibility judgments: concreteness. In Chapter 5 I argue that revisionists can appeal to the influence of concreteness on our responsibility judgments in support of the claim that there is a principled difference between the epistemic status of some of our intuitive judgments about moral responsibility and others.

1 Experimental philosophy – goals and methodology

There has been much discussion in the recent literature about the goals and proper role of what many have termed *experimental philosophy*. Here I outline some of the primary motivations for this philosophical approach, as well as the methodological differences that set this work apart in the discipline more generally.

There is some disagreement about how this particular approach to philosophical issues ought to be characterized. It is notoriously difficult to attempt to define experimental philosophy in terms of its subject matter, and attempts to distinguish this movement from other parts of the discipline often focus on differences in goals and methodology. For example, Tamler Sommers provides the following helpful description of experimental philosophy as

> …a movement which employs the methods of empirical science to shed light on philosophical debates. Most commonly, experimental philosophers attempt to probe ordinary intuitions about a particular case or question in hopes of learning about the psychological processes that underlie these intuitions. (Sommers 2010, 199).

Unsurprisingly, this methodology is better suited to some philosophical topics than others. For example, it stands to be particularly useful when it comes to debates that rely heavily on fundamental assumptions or generalizations about the intuitions of the folk or ordinary language users. Such assumptions and generalizations are obviously widespread in philosophy as a whole, but recent work in experimental philosophy has, as a matter of contingent fact, focused on several issues in particular: free will, personal identity, knowledge, and morality (Knobe & Nichols 2008).

Why approach these issues using empirical methods? The goals of experimental philosophy, like the more traditional philosophical method of conceptual analysis, have much to do with attaining a more accurate understanding of our concepts, and the way that we *think* about

things like free will, personal identity, knowledge, and morality. However, rather than attempting to provide an analysis of the concepts themselves, and to do so from the armchair, experimental philosophers are primarily interested in providing an explanation of *why* we have the concepts that we do. The goal here is not merely to collect data about patterns in people's actual intuitions about the appropriate application of certain concepts. Rather, it is to

> …provide a deeper explanation of why the intuitions come out this way….And ultimately, the hope is that one will be able to arrive at a more fundamental understanding of people's thinking in the relevant domain. (Knobe & Nichols 2008, 5)

Of course, collecting empirical data on what ordinary language users think about some of the most vexing questions in philosophy might also itself be of interest and have some intrinsic value. After all, it has long been common practice for many philosophers to assume that their own intuitions about these things are widely shared. So, determining whether or not these assumptions are warranted might be of great use in many contemporary philosophical debates. But collecting data on what our intuitions in fact are is only the first step in the larger project of achieving a greater degree of explanatory depth regarding our concepts.

So, what does the use of empirical methods to address these issues amount to, and how might we characterize the overall goals and methodology of experimental philosophy? This is a surprisingly difficult question, and to answer it I will borrow heavily from recent attempts by Joshua Knobe and Shaun Nichols (2008), and Jesse Prinz (2008).

First, experimental philosophy seems to share a starting point with the more traditional methodology of conceptual analysis, in that both begin with questions about ordinary people's intuitions about cases. However, from there they share little else. While conceptual analysts analyze concepts by looking in large part at intuitions about the necessary and sufficient conditions for their application, experimental philosophers are far more concerned with

providing an account of the factors that *influence* those applications (Knobe & Nichols, 2008). In particular, what experimental philosophers are interested in are the *"internal psychological processes* that underlie such applications" (Knobe & Nichols 2008, 5). Here progress is measured:

> …not in terms of the precision with which one can characterize the actual patterns of people's intuitions but in terms of the degree to which one can achieve explanatory depth. (Knobe & Nichols 2008, 5)

So, while experimental philosophers take data on the application conditions for a concept to be their starting point, providing an account of those conditions is not the ultimate goal. Rather, what they aim to provide is a better understanding of *why* our intuitions turn out the way that they do, and a deeper understanding of the *way* that people think about the issue in question.

For example, Knobe and Nichols characterize the methodology of experimental philosophy as a three step process. First, experimental philosophers gather empirical data on people's intuitions about a particular concept, and attempt to identify certain patterns in these intuitions. This first step may involve drawing from a body of existent data, or constructing one's own experimental study in order to collect new data. If the latter, this new experiment will be designed around a clear research question or hypothesis.[72] For example, one might begin with the following hypothesis:

> (a) Most people think that free will and moral responsibility are compatible with determinism, but incompatible with fatalism.[73]

[72]For further discussion of whether or not experimental philosophy satisfies the most basic demands of experimental design and data analyses, see Bernstein (2007).
[73] Nahmias et al. (2007) design a study to test just this hypothesis.

With a particular hypothesis like this one in mind, the experimental philosopher then goes on to gather empirical evidence either confirming or disconfirming the hypothesis, usually by way of surveying groups of non-philosophers.

However, a potential concern for this first step in the overall methodology of experimental philosophy is that one's initial hypothesis can often be problematically general. For example, a great deal of the initial empirical work on free will and moral responsibility has focused largely on the traditional compatibility question. So, the closest approximation to a research question or hypothesis suggested by some of this early work is something like one of the following:

(b) People's intuitions about free will are largely compatibilist.

(c) People's intuitions about free will are largely incompatibilist.

Unfortunately, the generality of these hypotheses make any data they generate very difficult to interpret. And what experimental philosophers are interested in is not the data itself, but what that data suggests about how and why people have the intuitions that they do. But there are a host of considerations that might influence people's intuitions about something as general as compatibilism or incompatibilism, many of which will be discussed later in this chapter. The main point here is that, ideally, experimental philosophers who begin their project by collecting new data via their own experimental design will begin with a specific, clearly formulated research question more akin to (a).

Second, once a pattern (or lack thereof) in people's intuitions has been identified, experimental philosophers attempt to "provide a deeper explanation of why the intuitions come out this way" (Knobe & Nichols 2008, 5). How they go about doing so might best be explained

with an example. Eddy Nahmias and colleagues offer the following explanation for data in support of hypothesis (a) above:

> We think these results are best explained in terms of the psychological processes that regulate whether people engage in the mechanistic stance toward other agents. By defualt, humans take the *participant stance* toward other agents who behave in purposeful ways. When people adopt this stance toward an agent, they tend to assume that ascriptions of [free will] and [moral responsibility] are appropriate unless and until they have perceived factors that are paradigmatically excusing or exempting….However, when people adopt the mechanistic stance toward an agent (for instance, when primed by a description of decision-making in terms of neural processes), then they tend to disengage from the participant stance….Thus, being viewed as a mechanistic system does act as an exempting condition. (Nahmias et al. 2007, 233)

This explanation can be taken as a paradigm for the type of explanation that experimental philosophers attempt to offer once a particular pattern in our folk intuitions has been identified for a certain body of data. Above all, this kind of explanation should attempt to identify the specific processes (or other conditions, such as belief states) that generate the pattern in question. Given this explanation there are at least two options for how to proceed. First, the relevant explanation might suggest additional, more highly specified research questions and hypotheses to be tested with further experiments. For example, given the above explanation one might wish to test whether particular features of vignettes encourage individuals to take on the participant versus the mechanistic stance, and refine our understanding of these stances and the underlying mechanisms that generate them. Second, if one is satisfied with the explanation at hand, and confident that further data will continue to support it, one might move on to the third step in the experimental philosopher's overall methodology.

This third step involves incorporating the relevant explanation into a larger *theory* for the domain in question. Here there are several questions one might ask, depending on the particular domain. Taking again the topic of moral responsibility and Nahmias and colleagues' work as an example, one might ask the following:

1. How does this explanation bear on the current philosophical discussion, and the best theories available?
2. Does this explanation stand at odds with or directly contradict any particular theories, or any of the fundamental assumptions underlying them?
3. Does this explanation entail any particular *prescriptive* recommendations? Does it provide reason to think that our intuitions are getting things right, or that they are in error?

This is of course not an exhaustive list of the relevant questions that experimental philosophers might be interested in once they have a tenable explanation for why we apply a particular concept in the way that we do, but they are sufficient to highlight at least some of the final goals of experimental philosophy.

Consider again the example of Nahmias et al.'s work in regards to moral responsibility. One might suggest the following answers to the above questions. First, one might argue that the explanation provided gives rise to a healthy degree of skepticism about whether or not the correct way to frame the philosophical debate on free will and responsibility is in terms of the compatibility question. If Nahmias and his colleagues are right then our intuitions about the compatibility of free will and determinism are not settled merely by facts about free will and determinism, but also by the different psychological stances generated by different *descriptions* of determinism. In regards to the second question, one might argue that there is reason to doubt

traditional incompatibilist arguments that appeal to the idea that incompatibilism is more intuitively appealing than compatibilism. If Nahmias et al.'s results are correct, then the issue here is at least far more complex than previously supposed. Finally, an answer to the third question will likely depend on appeal to a variety of independent considerations. Is there any independent reason to think that judgments about responsibility generated while occupying the participant stance are more reliable, or likely to get things right than those generated while occupying the mechanistic stance? Here one might draw on a combination of philosophical and empirical considerations involving justification, epistemological warrant, as well as data in psychology and neuroscience. However, this brief sketch should be sufficient to show that, if there are good reasons to think that the data and explanation in question show us that there is some *error* in our current thinking, this looks to motivate prescriptive views about ways in which we ought to change it.

2 The compatibility question

As discussed above, one of the main issues that experimental philosophers have taken interest in is the topic of free will and moral responsibility. As a result, a great deal of data on free will and responsibility has already been generated, and continues to grow at a rapid rate. Given the diversity of this data, and the way that the research questions that these experiments set out to assess have evolved in only a few short years, it is helpful to provide a survey of the main dialectic of this work. Providing such a survey is the purpose of Sections 2-4.

Like the traditional philosophical debate, much of the early experimental work on free will and moral responsibility has focused on the compatibility question, the question of whether or not moral responsibility and determinism are compatible. In light of conflicting data produced

by this early work, experimental work on responsibility has since branched off in a number of different directions. In this section I will focus on several influential studies that take the compatibility question, or something very much like it, as their primary research question. In Sections 3 and 4 I will turn my attention to different strategies for interpreting the conflicting results generated by early work on the compatibility question. One such strategy is to argue that the conflict is merely apparent, and to explain why people *seem* to have both compatibilist and incompatibilist intuitions. Another is to accept that the conflict is genuine, and pursue one of a variety of different options for explaining this conflict. One such option is to appeal to a variant or contextualist view of our ordinary folk concept. Another is to attempt to reconcile our conflicting responsibility judgments with an invariant, unificationist view. Finally, one might argue that conflicting judgments are best explained by the fact that there is actually a multiplicity of folk concepts of moral responsibility.

Examining these different strategies for interpreting the existent data requires first having a clear picture of the data itself. In Section 2.1 I present a variety of studies that some have argued support compatibilism, and in Section 2.2 I present a variety of studies that others have argued support incompatibilism.

2.1 Support for compatibilism

The contemporary literature on free will and moral responsibility is littered with claims that incompatibilism is intuitively obvious. As such, much of the early experimental work in this area has focused on determining whether or not there is any empirical support for the following claim: *incompatibilism is intuitive*. Not surprisingly, the data conflicts.

On the compatibilist side of things the work of Eddy Nahmias, Stephen Morris, Thomas Nadelhoffer, and Jason Turner (hereafter referred to as NMNT) has had a significant impact.[74] NMNT set out to empirically test the claim, made widely by various incompatibilists[75], that ordinary people start out as "natural incompatibilists" and that incompatibilism best aligns with the intuitions of laypersons. In order to do so they begin with the following incompatibilist prediction, or hypothesis:

> (P) When presented with a deterministic scenario, most people will judge that agents in such a scenario do not act of their own free will and are not morally responsible for their actions. (NMNT 2006, 36)

They then presented subjects who had not previously studied the free will debate with three different deterministic scenarios. The first is a Laplacean conception of determinism, in which a supercomputer that

> …can look at everything about the way the world is and predict everything about how it will be with 100% accuracy…

predicts that an agent, Jeremy, will rob a bank at a particular time in the future and this prediction is correct (2006, 36). The second is a "simple" conception of determinism (echoing van Inwagen's "roll-back of the universe"[76]) in which:

> …there is a universe that is re-created over and over again, starting from the exact same initial conditions and with all the same laws of nature. In this universe the same conditions and the same laws of nature produce the exact same outcomes, so that every single time the universe is re-created, everything must happen the exact same way. (2006, 38)

[74] See Nahmias et al (2005, 2006) and Nahmias (2006).
[75] See, for example, Strawson (1986), Kane (1999), and Ekstrom (2002).
[76] See van Inwagen (1983).

In this universe Jill decides to steal a necklace at a particular time, and every time the universe is re-created she makes the same decision at that same time. The third and final deterministic scenario is intended to make salient the fact that agents' actions are "deterministically caused by factors outside their control," and subjects were asked to:

> Imagine there is a world in which the beliefs and values of every person are caused completely by the combination of one's genes and one's environment. (2006, 38)

In this world Fred and Barney are identical twins put up for adoption. Fred is raised by the selfish Jerkson family and Barney by the kindly Kinderson family. Correspondingly, when Fred and Barney each find a wallet containing $1000 Fred Jerkson keeps the money, while Barney Kinderson returns it to its owner.

The results of NMNT's study failed to support (P) in all three scenarios. In the Laplacean deterministic scenario 83% of participants judged that Jeremy was morally responsible for robbing the bank (and 88% judged him responsible for the alternative praiseworthy action presented – saving a child). In the roll-back deterministic scenario 77% of participants judged Jill morally responsible for stealing the necklace. Finally, in the third deterministic scenario focusing on the causal determination of one's genes and environment 60% of participants judged Fred morally responsible for stealing the wallet and 64% judged Barney morally responsible for returning it. NMNT conclude that this data raises an empirical challenge to the claim that ordinary people start off with incompatibilist intuitions or that incompatibilism best aligns with the intuitions of laypersons. Furthermore, they conclude that the data suggests there is actually a burden on *incompatibilists* to motivate their view in light of the fact that it is more metaphysically demanding than ordinary intuitions require (2006, 39).

Robert Woolfolk, John Doris, and John Darley (2006; hereafter WDD) also report empirical results that lend support to the intuitiveness of compatibilism, in particular

compatibilist views which take judgments of responsibility to depend on the degree to which an agent identifies with her action. For the purposes of their study WDD define "identification" as "the degree to which an actor wants or desires to perform a behavior and maintains a positive 'fundamental evaluative orientation' (Watson 1966) towards that behavior" (WDD 2006, 286). They set out to investigate whether identification influences our judgments of moral responsibility, even when the actor in question is strongly constrained (2006, 286). And they do so via a series of experiments.

In the first experiment, WDD set out to assess the impact of identification and constraint on responsibility attributions for a violent action – murder – and they hypothesize that both identification and constraint will affect judgments of responsibility (2006, 287). Subjects were presented with one of four vignettes, and all four shared the following initial feature:

> …two married couples, Susan and Bill and Elaine and Frank, are depicted on a Caribbean vacation, and subsequently on board an airliner returning home. It is revealed that Susan and Frank have been involved in a love affair and that Bill has discovered proof of the affair. (2006, 287)

Subjects were then presented with a *High Identification* or *Low Identification* vignette in either a *High Constraint* or *Moderate Constraint* condition. In both constraint conditions the couples' plane is hijacked. In the *High Constraint* condition, the hijackers give Bill a pistol with one bullet, point their machine guns at him, and order him to shoot Frank in the head. Upon doing so:

> Bill realized that there was no way to resist or overpower the hijackers, because he and the other passengers were no match for 8 heavily armed men; any attempted heroics on his part would result in more loss of life than obeying the hijackers' orders. (2006, 288)

In *High Identification/High Constraint* Bill has already previously decided to kill Frank (because it is the only way he can deal with the affair), and in *Low Identification/High Constraint* he has already previously decided to forgive Frank. In both identification conditions, Bill shoots Frank.

In the *Moderate Constraint* condition the plane is also hijacked, the hijackers give Bill a pistol with one bullet in it, and another hijacker points his gun at Bill. However, the rest of the vignette is quite different:

> Looking out the window, Bill saw that the plane was surrounded by heavily armed anti-terrorist forces. Bill quickly reviewed his options. He could try to persuade the hijackers that their situation was hopeless. He could stall until the anti-terrorist forces stormed the plane. The hijackers had been distracted by the arrival of the armed troops. Both the leader and the man holding a gun on Bill were nervous, frequently glancing out the windows of the plane. Perhaps, Bill thought, he could shoot the hijacker with the gun and the rest of the passengers could subdue the other two kidnappers. It was a risky move, but it could work. Bill though he just might be able to pull it off, but the hijackers were angrily ordering him to "get on with it." (2006, 288)

Again, in both the high and low identification conditions for *Moderate Constraint* Bill shoots Frank. But, in *High Identification/Moderate Constraint* Bill views the situation as an opportunity to kill his wife's lover and get away with it, and "feeling no reluctance, he placed the pistol at Frank's temple and proceeded to blow his friend's brains out" (2006, 288). And in *Low Identification/Moderate Constraint* Bill is "horrified," is certain that he does *not* want to shoot Frank despite his affair with his wife, and "although he was appalled by the situation and beside himself with distress, he reluctantly placed the pistol at Frank's temple and proceeded to blow his friend's brains out" (2006, 288). The results of this experiment were that, even in *High Constraint*, higher identification increased subjects' willingness to attribute moral responsibility (2006, 287).

In their second experiment WDD further increase the level of constraint by adding an *Absolute Constraint* condition in which they attempt to make salient that Bill acts intentionally in both *High Identification/Absolute Constraint* and *Low Identification/Absolute Constraint* but "was *unable* to do otherwise" (2006, 291). The results of this experiment replicated the findings of the first experiment, and WDD take them to suggest that identification elevates judgments of moral responsibility even when it is "highly plausible to suppose that the actor 'could not have done otherwise'" (2006, 291).

WDD conclude that the results of these experiments provide empirical support for the claim that ordinary people take non-causal elements such as identification into account when making judgments about moral responsibility. They do so even when the agent in question is coerced to perform an action, and this provides support for the claim that ordinary intuitions about moral responsibility are compatibilist.

2.2 Support for incompatibilism

On the incompatibilist side of things, experimental studies conducted by Shaun Nichols and Joshua Knobe have perhaps been most influential (Nichols 2004, Nichols & Knobe 2007). In one early paper Nichols (2004) argues that there is empirical evidence that young children possess the concepts of both a "Correlation Principle" (that if there is action, there is an agent), and a "Causal Principle" (that agents have causal powers to produce actions) (2004, 478). According to Nichols, this in turn supports the claim that young children have a notion of agent-causation according to which "(i) actions are caused by agents, and (ii) for a given action, an agent could have done otherwise" (2004, 474). If Nichols is correct and young children do in fact have and regularly apply the concept of agent-causation, then this would seem to provide at

least some support for the claim that libertarianism (and thus incompatibilism) aligns with ordinary intuitions about moral responsibility.

Nichols and Joshua Knobe (2007) also make one of the first calls in the literature to look beyond the question of *what* people's ordinary intuitions about free will and moral responsibility are to the question of *why* they have the intuitions that they do (663). In particular, they raise questions about the nature of the underlying psychological processes that generate these intuitions, and whether or not we have reason to think that these processes are generally reliable, biased, or distorted (2007, 664). Nichols and Knobe begin their study with the following hypothesis:

> Our hypothesis is that people have an incompatibilist theory of moral responsibility that is elicited in some contexts but that they also have psychological mechanisms that can lead them to arrive at compatibilist judgments in other contexts. (2007, 664)

They are especially interested in the role of *affect* and *concreteness* in generating intuitions about moral responsibility, and predict that people will make compatibilist judgments about responsibility when presented with concrete scenarios and incompatibilist judgments when presented with abstract scenarios.[77]

Nichols and Knobe report a series of experimental data that they take to support this hypothesis and confirm their prediction about affect and concreteness. They first presented subjects with descriptions of two different universes, Universe A and Universe B, in which an agent, John and Mary respectively, decides to eat French Fries. In Universe A:

> …everything that happens is completely caused by whatever happened before it….So, if everything in this universe was exactly the same up until John made his decision, then it *had to happen* that John would decide to have French Fries. (2007, 669; emphasis their own)

[77] I will discuss this asymmetry in much greater detail in Chapter 5.

Universe A is thus deterministic. On the other hand, in Universe B:

> ...*almost* everything that happens is completely caused by whatever happened before it. The one exception is human decision making….Since a person's decision in this universe is not completely caused by what happened before it, even if everything in the universe was exactly the same up until Mary made her decision, it *did* ***not*** *have to happen* that Mary would decide to have French Fries. She could have decided something different. (2007, 669; emphasis their own)

So, Universe B is indeterministic. Subjects were then asked which of these two universes is more like our own, and an overwhelming majority (over 90%) chose the indeterministic Universe B (2007, 669).

Subjects were then randomly assigned to either a *concrete* or *abstract* condition, and asked a question about Universe A (the deterministic universe). Those in the concrete condition were presented with a vignette in which an agent, Bill, is attracted to his secretary, decides that the only way to be with her is to kill his wife and three children, and correspondingly decides to do so. They were then asked whether or not Bill is "fully morally responsible" for killing his wife and children (2007, 670). Those in the abstract condition were simply asked whether or not it is possible (in Universe A) for a person to be "fully morally responsible" for their actions" (2007, 670). The results of the experiment were that 72% of participants in the concrete condition judged that Bill *is* morally responsible for killing his wife and children, while 86% of those in the abstract condition judged, to the contrary, that it *is not* possible for agents in Universe A to be fully morally responsible.[78]

Nichols and Knobe take this data to support their initial hypothesis, and to suggest that concrete vignettes trigger an underlying psychological mechanism, in particular an affective

[78] To avoid worries about the level of detail involved in the vignette presented to those in the concrete condition Nichols and Knobe also ran the experiment with a *simple concrete* condition, in which the case is simplified to, "Bill stabs his wife and children so that he can be with his secretary" (2007, 670). In this simplified concrete scenario 50% of participants still judged that Bill is fully morally responsibility for killing his wife and children.

response, which generates compatibilist intuitions while abstract vignettes generate overwhelmingly incompatibilist judgments. The final conclusion that Nichols and Knobe draw is that people have *both* compatibilist and incompatibilist intuitions, and that "these different kinds of intuitions are generated by different kinds of psychological processes" (2007, 681).[79] They remain agnostic about what these results show about our folk concept of moral responsibility more generally, and about whether or not we ought to take one of these psychological mechanisms and the corresponding intuitions that they generate to be more reliable than the other. However, they do suggest that some version of a *performance error model* might be the most plausible explanation for their data[80]. If some kind of performance error model is correct then concrete, affect-laden vignettes trigger an underlying psychological process that leads to biased or distorted judgments about moral responsibility. So, we should take our intuitions about abstract cases to be, in at least some sense, more reliable or informative than those generated by concrete cases. And since Nichols and Knobe's results suggest that ordinary intuitions about abstract cases are overwhelmingly incompatibilist, if some kind of performance error model is in fact the best explanation for the data then this lends empirical support to the claim that incompatibilism is intuitive.

Nichols and Knobe conducted an additional study in conjunction with Hagop Sarkissian, Amita Chatterjee, Felipe De Brigard, and Smita Sirker examining whether or not the results of the above study are stable across cultures (Sarkissian et al. 2010). Sarkissian et al. conducted a cross-cultural study of the intuitions of undergraduate students in the United States, India, Hong Kong, and Columbia. Subjects were presented with the same descriptions of two universes, Universe A and Universe B, from Nichols and Knobe (2007). They were then asked which of

[79] They also propose three possible psychological models that might explain this result. These proposals will be discussed in further detail in Chapter 5.
[80] I will discuss this model in much greater detail in Chapter 5.

these two universes is most like our own, and whether or not in Universe A (the deterministic universe) it is possible for a person to be fully morally responsible for their actions. The results suggested a significant level of cross-cultural convergence. The majority of subjects across all four cultures judged that Universe B, the indeterministic universe, was most like our own, and that it was not possible for a person to be fully morally responsible for their actions in Universe A (2010, 353). Sarkissian et al. conclude that these results provide prima facie empirical support for the claims that both indeterminism and incompatibilism best align with the intuitions of the folk.

3 Incompatibilist intuitions are merely apparent

In light of the studies discussed above it looks as though initial data generated by empirical work centered around the compatibility question conflicts. Some of these results suggest that compatibilism is intuitive, while others suggest that incompatibilism is. One strategy for resolving this conflict is to argue that it is merely apparent. The main proponents of this strategy have been Eddy Nahmias (2006, 2011), and colleagues Coates and Kvaran (2007; hereafter referred to as NCK) who argue that descriptions of determinism often generate the intuition that our agency is being *bypassed*, and that it is the incompatibility of responsibility with a mechanistic, reductionist, or fatalist picture that people find intuitive, not the incompatibility of responsibility and determinism per se.

Nahmias (2006) suggests that ordinary people do not take determinism, but rather certain reductionist or mechanistic descriptions of decision-making to be a threat to free will and moral responsibility. He first raises an objection to some of the data presented by Nichols (2004), which Nichols takes to indicate that people (in particular young children) have the libertarian

intuition that human decisions are uniquely indeterministic. Nahmias ran his own experiment testing for this intuition with three different scenarios. In Scenario L a lightning bolt hits a particular tree at a particular time, in Scenario I a woman decides to order vanilla ice cream over chocolate, and in Scenario S she decides to steal a necklace. Subjects were then asked to imagine the universe being re-created over and over, and judge whether or not these events would happen in the exact same way every time. The results were that subjects responded that the events would *not* happen the same way every time, and only 9 out of 99 said that the *physical* event (the lightning bolt) would happen the same way every time but that the human decisions would not (2006, 219-220). According to Nahmias, these results suggest that people are indeterminists about certain *complex* processes, not human-decision making in particular, and that this casts doubt on Nichols' claim that young children and ordinary adult human beings more generally have libertarian intuitions.

Nahmias then argues that previous data taken to suggest the intuitiveness of incompatibilism reveals only *apparent* intuitions in support of incompatibilism. Nahmias focuses in particular on the data presented by Nichols and Knobe (2007), and takes issue with their description of Universe A. He argues that using the locution "*has to happen*" in describing Universe A primes the intuition that in this universe human agency is bypassed, and that some version of fatalism, reductionism, or epiphenomenalism is true (2006, 227). Nahmias then ran an additional experiment to test the hypothesis that the intuitiveness of incompatibilism is merely apparent, and that it rests on a failure to distinguish between the threat of determinism and these alternative potential threats. Nahmias presented subjects with two 'twin earth' scenarios in which the actions of beings very similar to us, Ertans, are completely caused by prior events.

The only difference between these two deterministic scenarios is the type of events that cause the Ertans' decisions and actions. In the *Neuro-scenario*:

> …the Ertan neuroscientists have discovered exactly how Ertan's brains work. The neuroscientists have discovered that every single decision and action Ertans perform is *completely caused* by the particular chemical reactions and neurological processes occurring in their brain at the time, and that these chemical reactions and neurological processes in the brain are *completely caused* by earlier events involving their particular genetic makeup and physical environment. (2006, 231; emphasis author's own)

And in the *Psych-scenario*:

> …the Ertan psychologists have discovered exactly how Ertans' minds work. The psychologists have discovered that every single decision and action Ertans perform is *completely caused* by the particular thoughts, desires, and plans they have at the time, and that these thoughts, desires, and plans are *completely caused* by earlier events involving their particular genetic makeup and upbringing. (2006, 231; emphasis author's own)

The results generated by this experiment were that in the Neuro-scenario only 18% of participants judged that the Ertans act of their own free will and only 19% judged that they deserve credit or blame for their actions. However, in the Psych-scenario 72% judged that the Ertans *do* act of their own free will, and 77% judged that they *do* deserve credit or blame for their actions (2006, 231). Nahmias takes these results to support the hypothesis that it is reductionism, mechanism, and bypassing, *not determinism* that ordinary people take to be a threat to free will and responsibility. The conflicting data previously reported by Nichols and Knobe can be explained by the fact that they use the "has to happen" locution to describe determinism, and this description suggests a reductionist, mechanistic picture in which human agency is bypassed.

Nahmias, Coates, and Kvaran (2007; hereafter referred to as NCK) expand on this work. NCK refine Nahmias' earlier hypotheses and make a distinction between *Pure Incompatibilism* (PI), the view that determinism and free will and responsibility are incompatible, and *Mechanism Incompatibilism* (MI), the view that reductive mechanism is incompatible with free will and responsibility (2007, 215). NCK argue that, at the very least, MI is more central to people's intuitions about free will and responsibility than PI, and that PI merely appears to be intuitive because it trades on intuitions that support MI by conflating determinism with mechanism (2007, 216). They also more clearly articulate the concept of *bypassing* as the circumstance under which our deliberations and conscious purposes are causally irrelevant to our decision-making and actions (2007, 220).

Based on their hypotheses NCK make the following predictions:

1. When determinism is described mechanistically people are less likely to judge that agents act freely and are morally responsible.
2. When determinism is described in *non*-mechanistic psychological terms people are more likely to judge that agents act freely and are morally responsible.
3. The degree to which *indeterminism* is described mechanistically has the same influence on people's judgments about whether agents act freely and are morally responsible. (221)

NCK take the data reported in Nahmias (2006) to lend empirical support for (1) and (2). In addition, they ran a new experiment which makes several more distinctions than Nahmias' Neuro and Psych scenarios alone. In particular, they give these scenarios a moral valence and test subjects' intuitions about morally good and morally bad actions for both the Neuro and Psych

conditions, and also test their intuitions about what is possible given different deterministic descriptions of the *actual* world (as opposed to Erta).

Nahmias (2011) provides a helpful summary of the results of these additional experiments. NCK's subjects were asked to report a range of judgments about whether the agents in each scenario "act of their own free will," "are morally responsible," "deserve blame (praise) for their actions," and whether the "agents' decisions are up to them" (2011, 564). Here for ease of exposition I will focus only on the results regarding their judgments about moral responsibility.[81] In the *Neuro Abstract (Alt. World)* condition, where determinism is described on Erta in neuroscientific terms and subjects were asked to judge the actions of Ertans in general, only 52% judged that Ertan agents are morally responsible for their actions. In the *Psych Abstract (Alt. World)* condition, where determinism is described on Erta in psychological terms and subjects were asked to judge the actions of Ertans in general, over 70% judged that Ertan agents are morally responsible. In the *Neuro Concrete Bad (Alt. World)* condition, 79% of subjects judged that the Ertan in question was responsible for killing his wife and children, while 63% judged the Ertan in question responsible for donating money to an orphanage in the *Neuro Concrete Good (Alt. World)* condition. The difference between the results generated by the *Psych Concrete Bad (Alt. World)* and *Psych Concrete Good (Alt. World)* conditions was similar to the *Neuro Concrete Good/Bad (Alt. World)* conditions. Finally, in the *Psych Abstract (Real World)* condition 89% of subjects judged that we are morally responsible, while in the *Neuro Abstract (Real World)* condition only 41% judged that we are morally responsible (Nahmias 2011, 564; NCK 2007, 230).

[81] Interestingly, NCK did find some intuitional differences in these closely related concepts, but they are not relevant to my current project of providing a survey of the recent experimental work on moral responsibility in particular. For further discussion of these results see NCK (2007) and Nahmias (2011).

NCK and Nahmias in particular draw several conclusions based on this data. First, they take it to duplicate Nichols and Knobe's (2007) data and offer further empirical support for the claim that the concreteness or abstractness of a particular scenario influences ordinary people's intuitions about moral responsibility, and that people are more willing to judge that agents are responsible in concrete scenarios than they are in abstract scenarios. However, they offer an alternative explanation to the affective performance error model suggested by Nichols and Knobe. Rather, NCK suggest that the affective response triggered by concrete cases is better understood as an *enabling* factor for making competent judgments about responsibility. This alternative model will be discussed in further detail in the next two chapters, and so I will not discuss it further here.

The main conclusion drawn by NCK is that the intuitions people often express in support of PI (pure incompatibilism) are merely *apparent*, and that it is actually varieties of bypassing, mechanism, and reductionism that they intuitively take to be a threat to free will and moral responsibility. And none of these alternative threats are entailed by the truth of determinism. The expression of apparent pure incompatibilist intuitions can be explained by the fact that certain descriptions of determinism (in particular the "has to happen" locution used by Nichols and Knobe) increase judgments that the agent in question is being bypassed. So, apparent incompatibilist intuitions can be explained by the fact that ordinary people tend to conflate determinism with bypassing. NCK take these results to cast doubt on the claim that previous data provides genuine empirical support for the intuitiveness of incompatibilism. To the contrary, they take their data to provide empirical support for the claim that *compatibilism* in fact best aligns with the ordinary intuitions of laypersons.

4 Variantism, invariantism, contextualism, and multiple concepts – alternative interpretations of the data

By now it should be clear that there has thus far been a great deal of philosophical disagreement regarding how to interpret the conflicting results generated by initial empirical work on the compatibility question. Rather than attempting to explain away this conflict as merely apparent (as Nahmias and his colleagues do) several strategies have been employed for interpreting these results in terms of differing views about the nature of our folk concept of moral responsibility. There are at least three main strategies for such interpretations of the conflicting data taken at face value.[82] First, one might argue that it is best be explained by the fact that our concept of moral responsibility is *variant*, and that there is no single set of conditions or criteria for responsibility that hold fixed across all contexts. Second, one might attempt to explain how we could maintain commitment to the claim that our concept of moral responsibility is *invariant*, despite the conflicting data. Finally, one might argue that the conflicting data indicates that our folk concept of moral responsibility is in fact *contextual*, or that there are actually *multiple folk concepts* of moral responsibility.[83] Here it is important to stress that none of these options express *metaphysical* claims about moral responsibility, merely possible ways of interpreting and explaining existing empirical data about ordinary intuitions about moral responsibility. All of these options are therefore positions on might hold on the nature of our *folk concept* of responsibility. Below I briefly survey attempts to pursue each of these options.

[82] These are not intended to be exhaustive, merely to reflect the main dialectic of attempts to interpret the existing data thus far.

4.1 Variantism

First, John Doris, Joshua Knobe, and Robert Woolfolk (2007; hereafter DKW) make a case for the claim that some kind of variantism best aligns with our intuitions about moral responsibility. They take typical philosophical approaches to moral responsibility to be committed to both *invariantism* and *conservatism*. Invariantist theories, "posit exceptionlessly relevant criteria for moral responsibility attributions," or that the criteria for moral responsibility is the same in every context (2007, 184). On the other hand, conservatism is defined as a commitment to the methodological principle that "folk belief is a constraint on philosophical theorizing" (2007, 185). A commitment to conservatism need not entail that philosophical theorizing is held hostage to our folk theories, only that at minimum a theory that stands sharply at odds with our intuitions has a comparative disadvantage to other theories. DKW argue that these two commitments cannot be jointly maintained because there is strong empirical support for the claim that our folk thinking about moral responsibility is *not* invariantist, and that it is instead likely that the criteria for our attributions of moral responsibility vary with the circumstances in which these attributions are made (2007, 185).

DKW take both traditional philosophical disagreement about the compatibility of moral responsibility and determinism and existing empirical data to support this claim. Here I will set aside their discussion of the former and focus on the latter. In particular, they focus on four different asymmetries that suggest that our attributions of moral responsibility vary with either the evaluative valence or outcome of the action in question.

First, DKW cite data regarding *side-effect asymmetry* produced by a much discussed study designed by Joshua Knobe (2003).[84] In this study Knobe presented subjects with a case in which the chairman of a board makes a decision with the goal of increasing profits, but is aware

[84] What has since been coined the "Knobe effect" will be discussed in further detail in Section 5.

that this decision will also bring about another foreseen side-effect that is not desired. Knobe randomly assigned subjects to one of two conditions, the *Harm* condition or the *Help* condition. In the Harm condition the foreseen side-effect is that the decision will also harm the environment (and the chairman does not care at all about this). In the Help condition the foreseen side-effect is that the decision will also help the environment (and the chairman does not care at all about this). The results of this study were that subjects tended to assign blame to the chairman in the harm condition, but were less likely to assign praise in the help condition (DKW 2007, 194).

Second, DKW cite data that provides empirical support for *overwhelming emotion asymmetry* (2007, 194). Pizarro et al. (2003) constructed a series of vignettes in which an agent, Jack, performs either a good (impulsively giving a homeless man his jacket) or bad (impulsively smashing the window of a car) action because of a potent emotion, which is then contrasted with a case in which he performs the same action calmly and deliberately. The results of this study were that subjects attributed considerably less blame for the bad action when it was the result of an overwhelming emotion, but there was no corresponding effect for the good action.

Third, DKW discuss data supporting *unrealized intention asymmetry*. Malle and Bennett (2002) presented subjects with two pairs of sentences, one pair describing a morally good action and intention, and one pair describing a morally bad action and intention. The morally good pair was presented as follows:

> [*action*] helped a neighbor fix his roof.
>
> [*intention*] intends to help a neighbor fix his roof. (cited in DKW 2007, 195)

And the morally bad pair was presented as follows:

> [*action*] sold cocaine to his teenage cousin.
>
> [*intention*] intends to sell cocaine to his teenage cousin. (2007, 195)

Subjects were presented either with two sentences about the good and bad actions, or two sentences about the good and bad unfulfilled intentions, and were asked to judge how much praise or blame the agent deserved. The results were that subjects were considerably more willing to assign blame for the bad intention than praise for the good intention (2007, 195).

Finally, DKW cite a large body of data (beginning with Walster (1966)) regarding *negligence asymmetry*. This asymmetry is generated by cases in which harm is caused due to an accident, and is closely related to the legal standard of "reasonable care," according to which an agent should not be held responsible for accidental harm if they took all of the precautions they could reasonably be expected to (DKW 2007, 195). According to DKW, this body of data indicates that standards of reasonable care depend on the evaluative status of the outcome of the action in question. In particular, people often attribute moral responsibility even when negligence is minor for actions that result in *severe* harms, but are not willing to attribute responsibility for *slight* harms unless the agent was exceptionally negligent. For example, Walster (1966) presented subjects with a vignette in which a man parks his car at the top of a hill, remembers to put on his emergency brake, but has neglected to have the brake cables serviced. In this study participants were significantly more likely to judge the man responsible when the negligence resulted in serious injury to an innocent child than when it resulted in a damaged fender (DKW 2007, 196).

DKW take the body of data in support of these four asymmetries to provide significant empirical support for the claim that our attributions of moral responsibility are normative "all the way down" (2007, 197). Rather than thinking that we first identify whether the relevant causal relations and psychological states obtain in order to establish whether a particular attribution of moral responsibility is appropriate, and then looking to the normative status of the action in order

to determine whether praise or blame is appropriate (and to what degree), evidence of these asymmetries indicates that the normative status of the outcome of an action actually plays a role in determining whether or not we make the attributions themselves (2007, 193). And the above data suggests that the effect of these normative considerations on our attributions of responsibility depends on a variety of factors that differ across cases. According to DKW, this casts serious doubt on the claim that the criteria for moral responsibility are invariant, and that ordinary people do in fact apply the *same* criteria in *every* case.

According to DKW we are therefore left with two options. First, we might jettison our commitment to conservatism, disregard some features of our folk thinking suggested by the data discussed above, and adopt some kind of *revisionary invariantism* about moral responsibility. Alternatively, we might accept what this data suggests about the variant features of our folk thinking, abandon our current commitment to invariantism, and adopt *conservative variantism* about moral responsibility. DKW argue for the latter option, largely because they foresee the cost of revision to be too high (2007, 202-205).

4.2 Reconciling the data with invariantism

Dana Nelkin (2007) resists DKW's variantist conclusion and attempts to reconcile the data that they discuss with an invariantist or, to use her preferred terminology, *unificationist* view of responsibility. According to Nelkin, the empirical data which provides support for the asymmetries discussed above suggests the following conclusion:

> (*Empirical Conclusion*) There is no plausible set of criteria that fit with all (or even most) of our ordinary judgments. (2007, 247)

And this conclusion seems at odds with the following two assumptions:

> (*Unity Assumption*) There is a single set of conditions for moral responsibility that applies in all cases.
>
> (*Fit Assumption*) The criteria for moral responsibility attributions fit with all (or most) of our ordinary judgments. (2007, 246)

Given the Empirical Conclusion, it looks as though either the Unity Assumption or the Fit Assumption is false. However, Nelkin resists this challenge to the Unity and Fit assumptions for two reasons. First, she points out that even those philosophers who rely most heavily on intuitions in their theorizing about responsibility often do so within the larger framework of reflective equilibrium, and so are concerned not so much with our *initial* intuitions, but our *reflective* intuitions (2007, 251). Given this larger methodological background, the Empirical Conclusion undermines the Unity Assumption only if the picture of our intuitions it suggests is "sufficiently complex to resist *any* and *all* unifying principles" (2007, 251; emphasis my own). Second, Nelkin suggests that our judgments of moral responsibility depend not just on our intuitive concept of responsibility, but also on our *understanding* of the relevant cases. This understanding can include both case-specific features and more general empirical assumptions. And, people can and likely often do make mistakes about both of these things (2007, 251). According to Nelkin, the best way to understand the meaning of 'fit' in both the Fit Assumption and the Empirical Conclusion is to:

> …think of the principles as fitting a complete story of the intuitions, background empirical assumptions (whether right or wrong), and features of the case as understood by subjects. (2007, 251)

And once we understand 'fit' in the relevant sense, the proponent of a unifying theory of responsibility can adopt a "combination strategy" for explaining how the empirical data fails to

show that there is no single criteria that applies in all cases, and how this data can fit with a unified, invariant theory of responsibility (2007, 251).

Nelkin provides an example of how this combination strategy might be employed, and offers her own explanations for how the data at issue might be reconciled with a unified Reasons view of responsibility. As discussed in Chapter 1, the main contours of a reasons view are that one is morally responsible if and only if they "act with the ability to do the (or a) right think for the right reasons" (Nelkin 2007, 245). Furthermore, Nelkin's preferred version of this view is asymmetrical: moral responsibility requires the ability to do otherwise for blameworthiness, but not praiseworthiness.

Nelkin suggests a series of background assumptions that might underlie the judgments about responsibility reported in the various studies cited by DKW. In regards to the data on *overwhelming emotion asymmetry*, she suggests that the following assumption might underlie the judgments represented in the data reported by Pizarro et al.: negative emotions block one's ability to see reasons and the ability to act on them, while positive emotions (i.e. empathy) are often a vehicle or enabling factor for recognizing these reasons (Nelkin 2007, 252). In regards to the *unrealized intention asymmetry*, Nelkin suggests that the results of Malle and Bennett's study might be explained by the fact that we often generally assume that the level of commitment associated with positive intentions is lower than the level of commitment associated with negative ones (Nelkin 2007, 253). In regards to the data on *side-effect asymmetry*, Nelkin suggests that this asymmetry might be explained by corresponding asymmetries embedded in morality, in particular asymmetries regarding positive and negative duties. In Knobe's chairman cases, for example, one might think that the chairman has a (negative) duty to avoid harmful consequences to the environment, but no corresponding positive duty to help the environment

(Nelkin 2007, 253). In regards to the data on *negligence asymmetry* and severity Nelkin suggests that the following underlying assumption might be at play: people generally assume that the higher the severity of consequences, the higher the level of commitment to the action (2007, 254). However, this would only explain severity asymmetry in intentional cases, and so people's asymmetrical judgments about cases of negligence like those reported by Walster (1966) would still be in need of explanation. Finally, Nelkin argues that data supporting the *abstract/concrete asymmetry* can also be explained.[85] Here she appeals to the work of Eddy Nahmias and his colleagues (discussed above), and endorses their conclusion that subjects often conflate determinism with bypassing, mechanism, or reduction. And she makes the further suggestion that describing cases concretely and in terms that evoke high affect might be enough to dispel the widespread, mistaken assumption that determinism entails reductionism (Nelkin 2007, 255).

While Nelkin does not take any of her proposed explanations to amount to definitive arguments that a unified reasons view of responsibility is correct, they do suggest at least one way in which the Unity Assumption might be consistently maintained along with the Fit Assumption and the Empirical Conclusion. And this is enough to defend the claim, contra DKW, that unified or invariant theories of responsibility can be reconciled with the growing body of empirical evidence.

Brandon Warmke (2011) also defends invariantism in light of the data cited by DKW. Like Nelkin, he challenges the empirical conclusion, which he phrases as follows:

> *Empirical Conclusion*: Empirical studies of ordinary judgments of responsibility attribution reveal that there is no single set of conditions under which the folk attribute responsibility. (2011, 181)

[85] This particular asymmetry is not discussed by DKW, but is discussed widely elsewhere. For example, Nichols & Knobe (2007) first reported results suggesting that abstractness and concreteness influence our judgments about responsibility. This particular symmetry will be the main topic of discussion in Chapter 5.

Warmke argues that the existing empirical data does not in fact support the empirical conclusion. In particular, he points out that the studies cited in support of this conclusion are "fragmented" (they do not all test the same kind of judgment), and argues that in light of this the data they generate "simply does not provide unified evidential support for the Empirical Conclusion" (2011, 192). In order to argue the point Warmke divides the relevant studies into three separate groups: A, B, & C. The studies in Group A are those in which subjects are asked to assign a certain degree of blame. The problem with these studies is that while they may provide evidence for the claim that people use a variantist standard in determining *how much* blame or praise to attribute to an agent, this does not show that the same standards apply to attributing praise and blame in the first place (2011, 193-194).

The studies in Group B are those in which subjects are asked whether agents deserve blame full stop. Here Warmke cites two reasons for thinking that these studies do not provide support for the Empirical Conclusion either. First, he points out that judgments of blame and moral responsibility might be conceptually distinct.[86] And if they are, showing that people use variantist conditions to attribute blame need not entail that they also use variantist conditions to attribute moral responsibility (and the latter is what is at issue). However, one might argue that even if blame and responsibility are conceptually distinct, the fact that an agent is blameworthy entails that they are morally responsible. The latter is required for the former, though one might remain silent on whether the entailment holds in the opposite direction. At the very least, Warmke argues that such a move is a "substantive philosophical thesis" in need of further argument (2011, 195). Second, Warmke cites independent empirical support for the claim that ordinary people do not in fact apply the same conditions for responsibility and

[86] Warmke cites John Martin Fischer (2006) as a prominent proponent of the claim that they do in fact come apart conceptually.

blameworthiness.[87] In light of this we should, at the very least, be suspicious of studies that draw conclusions about responsibility based on the conditions under which people attribute blame.

Finally, the studies in Group C are those in which subjects are asked directly about moral responsibility. Here the problems raised for the first two groups do not apply. However, Warmke points out that even these results fail to support the Empirical Conclusion on any explanation which attributes a widespread *performance error* to subjects.[88] If this sort of explanation is correct then the various asymmetries discussed above can be explained by the fact that some subjects are making a *mistake* – the underlying process which usually generates their competent responsibility judgments is being distorted in some way (according to the most popular going story, by an affective reaction). However, if this is correct then the asymmetrical results "*do not* provide evidence for the claim that the folk have a variantist theory of moral responsibility" (2011, 197). If some version of a performance error model is correct, then the distorted judgments that contribute to the asymmetries in question don't reveal anything about people's underlying theory of responsibility, just that particular factors distort or bias our competency in making judgments that reflect the underlying theory.

Warmke claims that those arguing for a variantist theory of responsibility must reject the performance error model if they hope to garner support for the Empirical Conclusion. But rejecting the performance error model will itself require further empirical work which has not yet been done. Until then, Warmke concludes that we should be, at best, skeptical about the Empirical Conclusion (2011, 198). And, if variantist arguments depend on some version of the

[87] Warmke cites a study by Harvey and Rule (1978), in which subjects' responses indicated that their judgments across a number of responsibility-related dimensions revealed two distinct clusters, one associated with responsibility and another with blame or moral evaluation (Warmke 2011, 196).
[88] As does Nichols and Knobe's (2007) preferred explanation for their own results.

Empirical Conclusion – as it seems they do – we therefore have reason to be skeptical about variantism.

4.3 Contextualism & multiple concepts of moral responsibility

There are at least two alternative strategies to variantism and invariantism. First, Woolfolk, Doris, and Darley (2005) suggest a third strategy for making sense of the apparently conflicting data. Despite the fact that the results they report seem to provide support for the claim that ordinary intuitions about moral responsibility are compatibilist (and in particular identificationist), they actually propose the following explanation:

> In fact, we would suggest that folk theories of responsibility are most likely *contextualist*, meaning that differing considerations are salient to moral responsibility attributions in different contexts, and that patterns of responsibility attribution may also vary culturally and developmentally. (2005, 298)

WDD point out that our discussion of responsibility takes places across many different contexts, and that these contextual differences may even be salient for the same individual over time. Differences between Nichols' results suggesting that young children have a libertarian conception of agency and Nahmias' results suggesting that ordinary adults may have compatibilist intuitions, for example, highlight the fact that this kind of contextual difference might be salient. Perhaps the concept of responsibility plays different roles in these different contexts, and so different standards or criteria might be relevant in different contexts. Furthermore, the context in which the philosophical discussion of responsibility takes place is significantly different from many social, political, and legal discussions of responsibility. It may be the case that the same concept of responsibility plays very different roles in these different contexts, and that different standards for responsibility might be relevant in each of them.

WDD merely suggest this contextualist strategy, and they do not provide anything like an account of how the details might be cashed out. However, in the past two decades several other philosophers have gestured towards a contextualist account of moral responsibility in light of the apparent intractability of the traditional philosophical debate on the compatibility question.[89] I will not discuss the details of these views here, rather I intend only to gesture towards this kind of contextualist strategy as another possible alternative for interpreting the existing empirical data. Rather than arguing that the data shows there are no invariant criteria for moral responsibility that can and should be applied in every case, or that the data can in fact be reconciled with the claim that there are invariant criteria, the contextualist strategy represents a possible middle way. Such a strategy might allow one to argue that while there is a single set of criteria for moral responsibility the standards for meeting these criteria might shift in accordance with certain contextual features. The question of what these criteria are and what those contextual features might be is (as WDD acknowledge) an interesting subject for further research (2005, 299).

Finally, Feltz, Cokely, and Nadelhoffer (2009; hereafter referred to as FCN) set out to examine whether the best interpretation of the data might be that there are actually *multiple folk concepts* of moral responsibility. They ran an additional experiment with Nichols and Knobe's vignettes using a *within-subject* design rather than the between-subject design of Nichols and Knobe's own experiment. FCN hypothesized that if the same individual is given *both* the high and low affect conditions they will give matched responses to each, suggesting that discrete groups of people actually have stable intuitions about moral responsibility. Their results aligned with this hypothesis: 25% of participants gave compatibilist responses in both conditions, 67% gave incompatibilist responses in both conditions, and only 8% gave mixed responses. FCN then

[89] For example, see Heller (1996).

ran a second study in which participants were asked about 'free will' rather than 'moral responsibility. The results of this study were similar to the first. Finally, they ran a third, larger study (110 rather than 65 participants), which generated results nearly identical to the first. FCN conclude that these results cast doubt on Nichols and Knobe's interpretation of the data, as only a small percentage of participants reported mixed intuitions about the high and low affect cases. Furthermore, they argue that their results provide prima facie evidence that there are actually multiple folk concepts of moral responsibility. While the majority of people seem to have consistently incompatibilist intuitions a significant minority also seem to have consistently compatibilist intuitions.

Here one might object that rather than providing prima facie evidence for multiple concepts of responsibility these results actually provide support for the claim that incompatibilism best aligns with the ordinary intuitions of laypersons (the view FCN refer to as "natural incompatibilism"). However, FCN reject this possible conclusion due to what they take to be three serious flaws in all of the studies based on Nichols and Knobe's experiment conducted thus far (including their own). First, all of these studies use the locution "has to happen" to make the determinism in Universe A salient, and as has been much discussed elsewhere (NMNT 2006) this locution may lead many subjects to conflate determinism with fatalism. Second, the "has to happen" locution is also "too ambiguous with respect to what it means to say that an agent could have done otherwise" (FCN 2009, 15). In particular, it is not clear whether this locution implies a conditional or unconditional analysis of whether agents in Universe A could not have done otherwise, and subjects might be relying on one or the other of these interpretations in making their judgments about responsibility. Finally, none of these studies are designed to control for the possibility that, for at least some subjects, the belief in free

will and responsibility might be "so deeply entrenched" that these subjects will judge an agent to be free and responsible "*no matter what*" (2009, 16). While manipulation checks might filter out some of these individuals, if a sizeable portion of the population has such beliefs then it is not clear that their reported judgments should be discarded. But failing to distinguish between individuals with such beliefs and genuine natural compatibilists might unduly raise the rate of those who report seemingly compatibilist judgments. So, at the very least future studies should find some way to control for this possibility. FCN take these three considerations to be serious methodological flaws, and thus conclude that their results provide prima facie evidence for stable intra-personal intuitional differences and in turn a multiplicity of folk concepts of moral responsibility, rather than natural incompatibilism.

This concludes my discussion of the main options for interpreting the apparently conflicting data generated by empirical work on the compatibility question. This survey is in no way intended to be exhaustive, merely representative of some of the primary strategies for interpreting the data which have been pursued thus far.

5 Individual factors that influence our responsibility judgments

The above discussion of various apparent asymmetries in our responsibility judgments suggests an alternative empirical research program to the one discussed thus far. Rather than taking the compatibility question as one's central focus in an empirical approach to moral responsibility, many experimental philosophers have recently opted for a more fine-grained approach. In particular, they have focused their work on identifying a variety of individual factors that seem to influence our judgments of moral responsibility, and on providing an explanation of the various asymmetries that these individual factors seem to generate.

Regardless of which general strategy one opts for in making sense of the initially conflicting data generated by the compatibility question, the results of this more fine-grained approach are likely to be relevant. For example, if one hopes to argue that the conflict is merely apparent and that our intuitions about moral responsibility really do fit best with compatibilism or incompatibilism, then an explanation of why we sometimes judge otherwise in terms of an underlying bias or performance error will be helpful (if not necessary). Likewise, if one hopes to argue that the apparent conflict and asymmetries can be reconciled with an invariantist view of responsibility. On the other hand, if one hopes to argue for a variantist or contextualist view of responsibility then filling out the details of such an account will require explaining what different factors lead us to apply different standards, concepts, or criteria in different circumstances.

So, it looks as though on any general strategy for making sense of the data generated by work on the compatibility question it will be helpful (again, if not necessary) to be able to tell a story about at least some of the individual factors that influence our judgments, and about why they do so. In this section I survey some of the individual factors that have thus far garnered a great deal of attention: the moral valence of the *action* being judged, the moral character of the *agen*t being judged, and stable features of *the subject doing the judging*.

This is by no means an exhaustive account of the individual factors that look to influence our responsibility judgments. In particular, I leave out any mention of the influence of the way in which particular cases or vignettes are *described*, which has perhaps received the most attention in the literature thus far. Asymmetries generated by this particular feature will be the primary subject of Chapter 5.

5.1 Knobe's side effect asymmetry

Joshua Knobe's (2003) paper, "Intentional Action and Side Effects in Ordinary Language," has spurred an explosion of discussion and further research in action theory. Knobe reports empirical results suggesting that ordinary people are considerably more willing to blame an agent for the foreseen side effects of an action that are bad than to praise them for good foreseen side effects. In his first experiment Knobe presented subjects with a vignette in which the chairman of a board of directors decides to implement a new program that will make a great deal of money for his company and also have some side effect. The chairman foresees this side effect, but "doesn't care at all" about it (2003, 190). Subjects were randomly assigned to either the *Harm* or *Help* condition. In the Harm condition the foreseen side effect harms the environment, and in the Help condition the foreseen side effect helps the environment. After being presented with the case, those in the Harm condition were then asked how much blame the chairman deserved for what he did (on a scale from 0-6), and whether or not they thought he had intentionally harmed the environment. Those in the Help condition were asked how much praise the chairman deserved for helping the environment, and whether or not they thought that he had acted intentionally (2003, 192-192). The results of this experiment were that 82% of participants judged that the chairman *did* act intentionally in the Harm condition, while 77% judged that he *did not* act intentionally in the Help condition (2003, 191).

In Knobe's second experiment the vignette is changed to a military setting in an attempt to control for people's potential emotional responses towards corporations and environmental harm. In this experiment a sergeant is ordered to send his squad to a particular location, again with a foreseen side effect. In the Harm condition the side effect is that the soldiers are killed in the line of fire, and in the Help condition the side effect is that the soldiers escape the battle.

This experiment replicated the asymmetrical results reported in the first experiment: 77% of subjects judged that the sergeant acted intentionally in the Harm condition while 70% judged that he did not act intentionally in the Help condition.

Finally, in both of these experiments subjects attributed a higher degree of blame in the Harm condition than praise in the Help condition. On average, the degree of blame assigned in the Harm condition was 4.8 (out of 6), but the degree of praise assigned in the Help condition was only 1.4 (out of 6). And these results suggest that ordinary judgments about responsibility are affected by the moral character of the action in question. In particular, ordinary people are more likely to attribute blame for bad actions (or at least those with bad side effects) than praise for good actions (or at least those with good side effects).

Many others have replicated and expanded Knobe's results in support of this side effect asymmetry for our ordinary judgments about praise and blame and intentional action. I will now turn to one particular expansion of these results.

5.2 Moral character (of the agent or action judged)

Thomas Nadelhoffer (2004, 2005, 2008) is one of many to have replicated and extended Knobe's (2003) results. Nadelhoffer interprets the growing body of empirical data initiated by Knobe to suggest that people are more likely to judge that a morally negative action or side effect was brought about intentionally than a structurally similar good or non-moral action (2007, 150). Additionally, in light of new explanatory models in moral psychology that suggest that *emotional, non-rational* processes (rather than deliberate, rational ones) are primarily what

generate our moral judgments[90], Nadelhoffer proposes that our judgments of praise and blame *themselves* can have a similar influence on our judgments of intentionality (2007, 151-152).

In his own experiment Nadelhoffer (2007) presented subjects with one of two cases. Both cases are structurally similar in that an agent approaches the window of a car brandishing a gun and the driver speeds off with the approaching agent holding on to the side of the car. In each case the driver zig zags away, knowing (but not caring) that their erratic driving will put the agent holding onto the car in grave danger. Finally, in each case the agent holding onto the car rolls into oncoming traffic and is killed. While each of the two cases share these structural similarities they differ in regards to the moral character of the driver's actions. In *Case 1* the driver's actions have a negative moral character. In this case the driver of the car is a thief, the car is full of recently stolen goods, and the agent who approaches the car brandishing a gun is a police officer. The thief's attempts to escape the police officer are successful, and the officer dies. On the other hand, *Case 2* describes an "innocent" driver whose actions bring about the death of an attempted carjacker (2007, 156-157).

Subjects were then asked three questions regarding whether or not the driver described *knowingly* brought about the other agent's death, whether they did so *intentionally*, and how much *blame* (on a 6-point scale) the driver deserved. Replicating Knobe's (2003) data, the results of this experiment were that subjects were more likely to say that the thief in *Case 1* knowingly brought about the police officer's death than that the innocent driver knowingly brought about the carjacker's death in *Case 2*. Specifically, 75% said that the thief knowingly brought about the death of the police officer while only 51% said the driver knowingly brought about the death of the attempted carjacker (2007, 156-157). Furthermore, 37% said that the thief *intentionally* brought about the police officer's death while only 10% said that the driver

[90] For example, see Haidt (2001).

intentionally brought about the attempted carjacker's death. Finally, the average degree of *blame* attributed in *Case 1* was 5.11, while the average degree of blame attributed in *Case 2* was only 2.01 (156-157).

Nadelhoffer concludes that these results are noteworthy for at least two reasons. First, they replicate Knobe's (2003) results in support of side-effect asymmetry. Second, they suggest that moral considerations (regarding either the moral character of the agent or action in question) affect ordinary judgments about whether an agent *knowingly* brings about a particular result and the degree to which an agent is *blameworthy*. According to Nadelhoffer, it looks as though we are putting the "moral cart before the intentional horse," and these results might best be explained in terms of some kind of performance error model for the underlying psychological processes that generate our intuitions in these cases (2007, 157). I will discuss this model in further detail in the next chapter. I now turn to one additional factor that empirical results suggest affects our judgments of praise and blame – the deep personality traits of the subject making the judgment.

5.3 Personality (of the judge)

In addition to Knobe, Nadelhoffer, and many others, Feltz and Cokely (2009) also attempt to identify sources of diversity in our folk intuitions about responsibility. However, they diverge from Knobe, Nadelhoffer, and others working on the influence of the moral character of the action or agent being assessed, and focus instead on the potential for features of *the agent who is making the judgment* to influence responsibility judgments. In particular, Feltz and Cokely report new data suggesting that general personality traits can be used to predict variations in intuitions about free will and moral responsibility (2009, 342). They examine specifically the

influence of extraversion on these judgments, and hypothesize that extraverts are more likely to be influenced by certain factors surrounding an action, especially factors that have a socially important and potentially affective dimension (2009, 345).

In their experiment Feltz and Cokely used scenarios identical to NCK's (2007) "psychologically non-reductionistic real world abstract" scenarios. However, the second paragraph (originally concerned with agents in general in the actual world) is replaced with the following:

> So, once specific earlier events have occurred in a person's life, these events will definitely cause specific later events to occur. For example, one day a person named John decides to kill his wife so that he can marry his lover, and he does it. Once the specific thoughts, desires, and plans occur in John's mind, they will definitely cause his decision to kill his wife. (Feltz & Cokely 2009, 345)

Feltz and Cokely cite three particular reasons for altering the scenario in this way:

> First, it involves a socially important action of a man killing his wife. Second, it describes determinism in terms of complete causation while at the same time avoiding using terminology implying that events had to happen as they did. Third, we know that in the psychologically non-reductionist, concrete scenarios the majority of people have compatibilist intuitions (NCK 2007, 230). (345)

Subjects were given this scenario and asked to rate (on a 7-point scale; 1 = strongly agree, 4 = neutral, and 7 = strongly disagree) how much they agreed with the statements that John's action was "up to him," that he decided to kill his wife "of his own free will," and that he is "morally responsible" for killing his wife (2009, 345). Subjects were also given a brief version of the Big 5 personality inventory (Gosling, Rentfrow, & Swann, 2003), and asked to rate to what degree they thought each prompt described them. In particular, they were asked to rate how much the

following to pairs of adjectives describe them: (a) extraverted, enthusiastic, and (b) reserved, quiet.

There are several noteworthy features of Feltz and Cokely's results. First, they replicated NCK's (2007) compatibilist-friendly results.[91] Second, their data suggests that extraversion is "positively, linearly related to compatibilist judgments," and thus explains a moderate amount of the variance reported in subjects' judgments (Feltz & Cokely 2009, 346). Third, there was a "clear quantitative shift between those who are high and low in extraversion in relation to the person being judged morally responsible" (2009, 346). Subjects who were moderately extraverted strongly agreed with the statement that John was morally responsible for killing his wife, while subjects who were low in extraversion only weakly agreed (2009, 346).

Feltz and Cokely conclude that stable psychological traits, in particular extraversion, predict whether a person has compatibilist or incompatibilist intuitions. And they take this to be important for at least two reasons. First, these results expand on other results regarding responsibility, intentionality, ethics, and epistemology that suggest that there is not necessarily any single, stable, widespread set of folk intuitions regarding these concepts (2009, 347).[92] Either a "single, homogeneous" set of folk intuitions is being *obscured,* or there is no such set (2009, 347). Either way, Feltz and Cokely suggest that future work should focus not just on the data regarding folk intuitions, but on the underlying causes that generate them (2009, 347). In regards to their own data, they propose the following explanation for the variation it reveals: perhaps extraverts are more highly influenced by the affective and social dimensions of actions when making judgments of responsibility than deterministic features. And perhaps this is

[91] The means were as follows: "up to him": $M = 2.72$, $SD = 2.03$; "free will": $M = 2.95$, $SD = 2.06$, "responsible": $M = 2.35$, $SD = 2.13$ (Feltz and Cokely 2009, 346).
[92] Feltz and Cokely cite, for example, Cushman & Mele (2008), Feltz (2007), Feltz & Cokely (2007, 2008), and Nichols & Ulatowski (2007).

because *holding* people responsible appears to serve an important regulatory function in social dynamics, a function that extraverts may be more sensitive to and concerned with preserving than introverts (2009, 345).

5.4 Interpretational diversity

Nichols and Ulatowski (2007) draw conclusions similar to Feltz and Cokely's when it comes to intuitions about intentionality. While Nichols and Ulatowski are primarily concerned with explaining Knobe's side-effect asymmetry they imply that their proposed explanation may generalize to a variety of other intuitions, including intuitions about responsibility. They hypothesize that Knobe's results are susceptible to ordering effects, and that people who received the Help condition first would be more likely to deny that the CEO acted intentionally in the Harm condition. They then conducted a new experiment to confirm this hypothesis, but much to their surprise found no ordering effects. However, upon further examination of the data they found that the results generated did fit a pattern:

> There was a strong correlation between responses. That is, people tended to answer the same way on the harming and helping questions ($r = .516$, $N = 44$, $p < 001$, two tailed). Responses split roughly into thirds. One third said neither [Harm nor Help] was intentional; another third said both were, and another third responded asymmetrically. The asymmetric responses were all of the same variety. No one said that the helping was intentional but not the harming. (2007, 355)

Given the correlations found in the minority responses (those who gave the *same* responses in Harm and Help) Nichols and Ulatowski conclude that it is a mistake to attribute these responses to randomness. Rather, they suggest that these minority responses reflect systematic individual differences in the way that people interpret 'intentional'. On one interpretation of 'intentional' (one that appeals to foreknowledge) the CEO acts intentionally in both Harm and Help, and on

another (one that appeals to motive) the CEO does not act intentionally in either condition (2007, 356).[93]

Nichols and Ulatowski conclude that because the term 'intentional' looks to exhibit at least some degree of semantic under-specification, neither matched group is making an *error*. And, those who report asymmetrical responses in Help and Harm might be "*flexible*" in the way they interpret 'intentional' in different contexts, though Nichols and Ulatowski leave open the details of what factors might actually push the subjects to "flex" one way rather than the other (2007, 360). They conclude that their hypothesis about intra-personal differences in interpretation might have wider ranging implications for our intuitions more generally.

Conclusion

The goal of this chapter has been to provide some background on the field of experimental philosophy (its primary goals and methodology), as well as a survey of the main dialectic in this field in regards to the topic of moral responsibility. At present, the conflicting results of empirical work on the compatibility question fall short of providing clear support for the claim that our folk concept of responsibility is either compatibilist or incompatibilist. Several different strategies have been implemented to interpret these results, including attempts to argue that the conflict is merely apparent, that our folk concept of responsibility is variant, contextual, or admits of multiple concepts, or that these results can in fact be reconciled with an invariant, unified concept of responsibility.

But which of these strategies for interpreting the data is most promising? Answering this question is notoriously difficult, in part because new results are being added to the relevant body

[93] The explanations provided by subjects for why they made the judgments that they did also seem to fit with this interpretation of the data (Nichols & Ulatowski 2007, 348-349).

of data all the time. For my purposes, I will set this question aside. Rather than examining *what* these results tell us about our ordinary thinking about responsibility (is our folk concept compatibilist, incompatibilist, variant, invariant, contextualist, or something else entirely?) I will examine further the question of *why* these results turn out the way that they do first place. I take this latter question to be far more relevant to the overall revisionist project. Taking up this question will be the topic of the next chapter, where I will focus on one asymmetry in our responsibility judgments in particular: the abstract/concrete asymmetry. Evidence for this particular asymmetry has been well established, and I will argue that providing a tenable explanation for it can lay the groundwork for a revisionist response to the normativity-anchoring problem.

Chapter 5
The Abstract/Concrete Asymmetry

Introduction

In this chapter I focus on one particular factor which influences our judgments about moral responsibility: the way that a particular case, question, or vignette is described or framed. More specifically, there is a growing body of empirical evidence indicating that abstractness and concreteness have a significant and very different influence on our responsibility judgments.

The goals of this chapter are threefold. First, I attempt to characterize the distinction between abstractness and concreteness. This is no easy task given that the distinction can be made along several different dimensions.[94] While making the distinction precise is beyond my current purposes I provide what I hope to be some helpful clarificatory remarks on the dimensions across which it is relevant to our judgments about moral responsibility. Second, I discuss the growing body of empirical data suggesting that abstractness and concreteness influence our responsibility judgments in very different ways. Third, I discuss various attempts to explain this asymmetry. In the next chapter I will provide some philosophical and empirical support for one family of explanation in particular, *concreteness enabling models*. I will argue that this kind of explanation can and should be utilized by revisionists as part of the positive project of saving revisionism from the normativity-anchoring problem.

1 The abstract/concrete distinction

I begin with an exercise for the reader: think of your favorite thought experiment intended to elicit intuitions about moral responsibility. Perhaps you are thinking of a certain Frankfurt case, or Pereboom's Four Case Manipulation Argument. Maybe a version of one of

[94] Nichols & Knobe (2007) and Sinnott-Armstrong (2008) have also attempted to make this distinction more precise.

the many vignettes discussed in Chapter 4 comes to mind. Chances are, whatever case you are thinking of involves a particular agent performing a particular action. And, if the data discussed in this chapter is to be taken seriously, this feature of the case affects your intuitions in a way that more abstract musing (say, for example, on the theoretical relationship between responsibility and determinism) does not.

The task of *explaining* this particular asymmetry will be the topic of Section 3 and, to some extent, the following chapter. In Section 2 I present a sampling of the wealth of data which supports the descriptive claim that our responsibility judgments are in fact subject to this asymmetry, and that concreteness makes a difference when it comes to our judgments about moral responsibility. However, before jumping into discussion of the data (*more* discussion, of *even more* data – the end is coming soon!) it is helpful to get clear on what the asymmetry in questions amounts to. Attempting to make this distinction more precise is the purpose of this section.

What I am here interested in is the influence of several closely related factors on our ordinary judgments about moral responsibility. One such factor is the level of detail with which particular cases, vignettes, or thought experiments are described (hereafter I will simply talk of "cases"). Generally, the more specificity involved in the presentation of a particular case, the more concrete it will be. And the sorts of things that increase the specificity of the presentation of the case are things like the following: the case makes reference to a particular action, performed by a particular agent (named or otherwise described with some level of detail), against the background of some particular set of circumstances. Already it should be clear that many of the cases used by philosophers to elicit intuitions about moral responsibility are likely to be

found somewhere on the concrete end of the spectrum when considering this dimension of concreteness independently.

A second factor is the degree to which the case described seems *possible*, *likely*, or *actual*. The degree to which a case seems possible, likely, or actual to the subject presented with it also increases the concreteness of the case. On the other hand, thought experiments about possible worlds significantly different from ours (or at least the way that we view ours), or agents significantly different from us (or at least the way that we view ourselves) will be more abstract, even if they describe the specific actions of specific agents.

Finally, the degree to which a case described elicits an *emotional*, or *affective*, response also increases the concreteness of a case. Cases that describe actions with a clear moral valence will likely be more concrete. This is particularly true of cases that describe *bad* acts. For whatever reason – and this will be discussed further in Section 3 – it seems to be a straightforward psychological fact that bad actions tend to elicit stronger emotional reactions than good acts. Here an example may be instructive. Consider a both a good and bad action with similar moral valences. Say, *donating* $1000 to a charity versus *stealing* $1000 from a charity. While impressive, most of us would likely have a rather ho-hum positive emotional response to the former, perhaps something along the lines of, "Well isn't that a nice thing to do." On the other hand, we are likely to have a much more pronounced *negative* emotional response to the latter. Perhaps, for example, we experience a certain level of moral disgust that might be expressed along the lines of, "What a reprehensible thing to do!" Again, for whatever reason, it looks to be an uncontroversial, descriptive fact that bad acts get us going emotionally in a way that good acts generally do not.[95]

[95] Of course, there will be some exceptions to this generalization. In particular, contextual factors might generate counterexamples. If, for example, you have a child with terminal cancer you might be moved to tears by a small act

So, there are at least three dimensions to the abstract/concrete distinction. The level of detail, the degree to which a case seems possible or actual, and the degree to which a case elicits an affective response all contribute to the concreteness of that case. This is by no means intended to be an exhaustive account of the abstract/concrete distinction, merely to capture some intuitive ways that the distinction might be made more precise. It is perhaps most helpful to think of this distinction as a spectrum, with particular cases lying somewhere closer to the abstract or concrete end of things. Different considerations may also conflict in pushing particular cases closer to one end or the other. For example, while a particular case may describe a specific agent's bad action in a high degree of detail and elicit a strong affective response, it may also be described as taking place in a possible world very different from the actual world or the way that we view the actual world. While the first two considerations might push the case in question towards the concrete end of the spectrum, radical divergences between the way that this possible world is described and the way that we view the actual world may in the end neutralize these considerations in terms of concreteness. So, the degree to which a particular case is concrete will depend on a variety of factors, including the way in which abstract and concrete considerations weigh against one another.

The main point suggested by the data discussed in the sections to follow is that where a particular case lies on the abstract/concrete spectrum can have a significant influence on our ordinary judgments about responsibility. I turn now to discussion of this data.

---

of kindness performed by a stranger. Though clearly not universal, the asymmetry in our emotional responses to good and bad actions does seem to hold true in general.

2 The data

Empirical evidence suggesting that concreteness has a significant influence on many of our philosophical intuitions is well-documented and widespread. While support for this claim need not be restricted to responsibility, here I will focus on several studies which indicate that concreteness has an influence on our responsibility judgments in particular. Some of these studies will be familiar from Chapter 4. For example, in Section 2.1 I will discuss further the work of Nichols and Knobe and Nahmias and colleagues (though here I will focus on what these studies suggest about the abstract/concrete asymmetry in particular). In the remainder of Section 2 I discuss some additional studies that provide evidence for the abstract/concrete asymmetry not discussed in Chapter 4. These include Small and Loewenstein (2005), Nichols and Roskies (2008), and Sinnott-Armstrong (2008).

2.1 Nichols, Knobe, and Nahmias – a review

In this section I return to the data reported by Nichols, Knobe, and Nahmias et al. already discussed at length in Chapter 4. Nichols and Knobe were the first to suggest that concreteness might significantly influence our responsibility judgments. And Nahmias and his colleagues have in several places replicated some of Nichols and Knobe's results, though they have interpreted them quite differently.

Nichols & Knobe focus on the role of *affect* in particular in generating intuitions about moral responsibility. Recall from the discussion in Chapter 4 that they intentionally place their subjects in either an abstract or concrete condition and ask them a question about what is possible in a deterministic universe (Universe A). In the abstract condition subjects were asked the general question of whether or not agents in Universe A can be fully responsible. Subjects in

the concrete condition were asked whether, in Universe A, Bill is fully responsibility for murdering his wife and children to be with his secretary. Here the abstract/concrete distinction between the two conditions is clear, and holds over several different dimensions. Not only does the question asked of subjects in the abstract condition fail to mention any particular agent or particular action, it also requires subjects to assess what would be the case in a universe very unlike their view of the actual universe (recall that 90% of subjects reported that Universe B, the indeterministic universe, was most like our own). Furthermore, there are no features of the question asked of subjects in the abstract condition that elicit an affective response. On the other hand, those in the concrete condition were presented with a specific agent (Bill), and a specific action (murdering his wife and children), that is quite clearly *wrong* and likely to elicit a strong affective response (moral anger or indignation). And, as hypothesized, subjects responded quite differently in these two conditions. 86% judged that agents in Universe A could *not* be fully morally responsible in the abstract condition while 72% judged that Bill *is* fully responsible for murdering his wife and children in the concrete condition.

Nahmias et al. (2007) and Nahmias (2011) reported similar results. They presented subjects with one of the following scenarios:

(1) **Ertan Abstract Psych**: 71.9% agreed that Ertans are morally responsible

(2) **Ertan Abstract Neuro**: 52.4% agreed that Ertans are morally responsible

(3) **Ertan Concrete Psych Good**: 68.5% agreed that Smit is morally responsible

(4) **Ertan Concrete Psych Bad**: 81.1% agreed that Smit is morally responsible

(5) **Ertan Concrete Neuro Good**: 63.0% agreed that Smit is morally responsible

(6) **Ertan Concrete Neuro Bad**: 79.2% agreed that Smit is morally responsible

(7) **Psych Real**: 88% agreed that Smit is morally responsible

(8) **Neuro Real**: 40.7% agreed that Smit is morally responsible

All of the Ertan scenarios (1-6) describe creatures much like us, on a planet much like ours, whose actions are completely determined. In the Ertan Psych scenarios (1, 3, and 4) psychologists have discovered that the Ertans' actions are completely caused by things like thoughts, desires, and plans. In the Ertan Neuro scenarios (2, 5, and 6) neuroscientists have discovered that the Ertans' actions are completely caused by things like specific chemical reactions and neural processes. The Abstract scenarios (1 and 2) describe the actions of Ertans in general, while the Concrete scenarios (3-6) describe a particular good or bad action performed by a particular Ertan named Smit. The good action is donating a large sum of money to an orphanage and the bad action is murdering his wife so that he can marry his lover. Finally, the Real scenarios (7 and 8) ask subjects to imagine that the best neuroscientists or psychologists have discovered that our actions are determined in the actual world.

Here the different results generated by several of these scenarios lend support to the abstract/concrete asymmetry. First, holding fixed the Neuro component, judgments of responsibility jumped from 52.4% in Ertan Abstract Neuro to 63% in Ertan Concrete Neuro Good. Even more significantly, judgments of responsibility jumped even higher to 79.2% in the Ertan Concrete Neuro *Bad* scenario. In fact, what I take to be most relevant to the asymmetry in question is that the three scenarios located farthest on the concrete end of the spectrum correspond with the three highest percentages for attributions of responsibility. Scenarios 4, 6, and 7 all describe a particular agent (Smit) performing a particular act that is clearly wrong (murdering his wife) and are likely to elicit an emotional response from subjects. Furthermore, the Psych Real (7) scenario is made even *more* concrete in that it no longer describes Ertans in a possible world, but a hypothetical discovery made in the *actual* world. Interestingly, this

scenario generated by far the highest percentage of agreement (88%) that Smit is morally responsible (NCK 2007).

How should we interpret this data? This question will be the subject of Section 3. For the moment I will set it aside and discuss some further results which provide further empirical support for the mere descriptive claim that our responsibility judgments are significantly influenced by concreteness.

2.2 Small & Loewenstein

Small and Loewenstein (2005) examine whether or not people are more punitive towards identifiable wrongdoers, even when they are not given any further information about them. They conducted an experiment in which participants drew a number between 1 and 10 and were given the following instructions for two rounds of the experiment:

> Round 1: At the beginning of Round 1 you, and every other participant, receive $5. You and each of the other nine group members must decide whether to contribute your $5 to the group or to keep it for yourself. If you contribute the money, then everyone in the group will receive $1.25 from you. If you do not, then everyone in the group will receive nothing from you. Therefore, your income from the experiment depends on what you do and what everyone else does. (2005, 314)

Participants were then asked to choose whether to keep their $5 or contribute it to the group. The best possible individual outcome would be to keep the $5 while everyone else contributed, in which case you would receive $16.25. The worst possible individual outcome would be to be the only one to contribute, in which case you would receive nothing. The best possible outcome for everyone in the group would be for everyone to contribute, in which case everyone would receive $11.25.

In Round 2 all of the participants who contributed in Round 1 were then given the opportunity to *punish* those who did not contribute. They were given a range of choices from not penalizing at all to penalizing up to $5 (in $1 increments). Penalizing also came at a cost of $.20 per $1 penalized. Participants who were able to punish were randomly assigned to either the *unidentified* or *identified* condition. In the unidentified condition participants were asked to decide on a punishment *before* randomly drawing the number of a group member who did not contribute and who would then receive the punishment. Participants in the identified condition randomly drew the number of the non-contributor they would be punishing before making a decision about how much, if any, to penalize them. (2005, 314-315)

Small and Loewenstein made the following three predictions: (1) people would punish an identified non-contributor more severely than an unidentified one, (2) there would be an increase in reported anger towards an identified non-contributor than an unidentified one, and (3) the effect of identifiability on punishment would be mediated by feelings of anger (2005, 313). All of these predictions were confirmed by their results. The effect of identifiability on severity of punishment was significant, but was mediated almost completely when anger was controlled for. Small and Loewenstein conclude that identification makes a situation more *concrete*, because it reduces the social distance between judges and targets. Perhaps most relevant to the topic at hand, participants also reported that they *blamed* identified non-contributors more (2005, 315-317).[96]

The results reported by Small and Loewenstein therefore provide further support for the claim that concreteness influences our responsibility judgments. Those in the identifiable condition were asked to penalize a *specific*, rather than abstract non-contributor (even though

[96] On a scale of 1-5, blame reported in the identifiable condition was $M = 2.92$, $SD = 1.38$, while in the unidentifiable condition $M = 2.29$, $SD = .98$ (2005, 316).

they were only given a *number* for that non-contributor, rather than a name), and also had a stronger emotional, affective response.

2.3 Nichols & Roskies

Shaun Nichols and Adina Roskies (2008) attempt to isolate another dimension of the abstract/concrete distinction and its influence on our responsibility judgments. Given that appeal to thought experiments usually involve asking people to report their intuitions about hypothetical scenarios in alternative or possible worlds, to what extent does this feature affect our judgments? In regards to moral responsibility we would expect our judgments to depend solely on features of the situation described, and *not* on our relationship to that situation. However, Nichols and Roskies found that, quite to the contrary, our judgments of moral responsibility "diverge across worlds even when facts about those worlds are the same" (2008, 371).

Nichols and Roskies randomly assigned participants to either the *Actual* or *Alternate* condition. Participants were given the same presentation of a deterministic universe. However, in the Actual condition that universe was implied to be our own, while in the Alternate condition they were first asked to "Imagine an alternate universe, Universe A, that is much like earth" (2008, 373). Participants in both conditions were then asked to rate their level of agreement (on a scale of 1 to 7, where 1 was "disagree completely," and 7 "agree completely") with each of the following statements respectively:

> *Alternate*: If these scientists are right, then it is impossible for a person in Universe *A* to be fully morally responsible for their actions.
>
> *Actual*: If these scientists are right, then it is impossible for a person to be fully morally responsible for their actions. (2008, 374)

The mean response in *Alternate* was 5.06, while the mean response in *Actual* was 3.58 (2008, 374). Nichols and Roskies then asked participants in both conditions to rate their agreement with the statement that, if these scientists are right, people "should still be morally blamed for committing crimes" (2008, 374). The mean response in *Alternate* was 3.67, while the mean response in *Actual* was 5.35 (2008, 375).

Nichols and Roskies conclude that whether or not a particular vignette or case is set in the *actual* world versus an alternate or *possible* world influences our responsibility judgments. And this is true *even if we hold fixed all other facts about the cases*. In particular, people are more inclined to judge that agents are responsible and that blaming them is appropriate in a deterministic world described as actual than they are for a deterministic world described as merely possible. And this holds true even if it is made explicit that the possible world is "much like earth" (2008, 373). In isolating this particular dimension of concreteness Nichols and Roskies' results therefore lend further empirical support to the abstract/concrete asymmetry.[97]

2.4 Sinnott-Armstrong

Whereas the studies discussed above focus on the influence of concreteness on our responsibility judgments in particular, Sinnott-Armstrong (2008) provides a helpful survey of evidence which suggests this phenomenon is in fact more philosophically widespread. Sinnott-Armstrong canvasses data on our judgments about responsibility (which he categorizes in the area of metaphysics), skepticism (epistemology), ethics, action theory (the Knobe effect on judgments of intentional action), and other traditional philosophical paradoxes in metaphysics in order to offer support for the claim that the abstract/concrete asymmetry persists across diverse

[97] Nichols and Roskies also provide a helpful discussion of the implications of these results, and of several different models for understanding and explaining the processes that generate them. I will discuss the explanation they propose in further detail in Section 3.2 of this chapter.

areas of philosophy. Here I will briefly outline the work cited by Sinnott-Armstrong on skepticism and deterrence in particular. I will return to Sinnott-Armstrong and his discussion of what he takes to be the best explanation for this asymmetry in Section 3.

After canvassing the abstract/concrete asymmetry in regards to our judgments of responsibility discussed above Sinnott-Armstrong presents a study by Nichols, Stitch, and Weinberg (2003) on people's ordinary intuitions about skepticism. Participants were given a lengthy vignette in which two roommates (George and Omar) have a late-night philosophical discussion (Sinnott-Armstrong 2008, 217-218). They discuss the future possibility of scientists being able to grow a disembodied virtual-reality brain, and debate whether or not they can know that they themselves are not in such a situation. George points out the fact that he has legs, that a disembodied virtual-reality brain would not have legs, and concludes that in light of this he will continue to believe that he is not a virtual-reality brain. Participants were told that George and Omar are in fact real humans in the actual world, and so George's belief is true. They were then asked whether or not George "knows," or "only believes" that his belief is true (2008, 217-218). The results of this study were that among the subjects that had taken three or more philosophy courses only 20% reported that George really knows, while 55% of subjects who had taken two or fewer philosophy courses reported that he did know (2008, 218).

Sinnott-Armstrong suggests that this difference reflects a distinction in abstract and concrete intuitions. The vignette presented by Nichols, Stitch, and Weinberg has a mixture of concrete and abstract elements. It is concrete "insofar as it refers to a particular incident, particular people, and a particular claim: that George has legs" (Sinnott-Armstrong 2008, 219). However, it is abstract in that the final question asked of subjects is whether or not George really knows that he is not a virtual-reality brain, which is "abstract insofar as it affects almost all of

George's beliefs indiscriminately," and the actual circumstances of being a virtual-reality brain are left rather vague (Sinnott-Armstrong 2008, 219). Sinnott-Armstrong hypothesizes that this difference in intuitions might be explained by the fact that subjects who had previously taken three or more philosophy courses were more inclined toward abstract thinking, while those who had not were less inclined. In order to test this Sinnott-Armstrong ran his own study with the hope of demonstrating the influence of concreteness alone on ordinary intuitions about knowledge. He presented subjects with the following *abstract* question:

> People sometimes believe things for no good reason. For example, people sometimes believe what a politician says about the economy when they have no good reason to trust what the politician says. Our question is about knowledge: If a person cannot give any good reason to believe a claim, is it possible that the person *knows* that the claim is correct? (2008, 220)

A second group was presented with a *concrete* version of the same question:

> If you cannot give any good reason to believe that the person whom you believe to be your mother really is your mother, is it possible that you *know* that she is your mother? (2008, 221)

The results were that only 52% of subjects who received the abstract question answered "Yes," while 88% of subjects who received the concrete version answered "Yes" (2008, 221). These two questions are structurally similar, and the only difference between them seems to be that the abstract question is phrased as a general, universal question about when it is possible for a person to know something, while the concrete question presents subjects with a specific, affect-laden question about what it is possible for the subjects themselves to know.

In addition to the results of this experiment Sinnott-Armstrong also gestures towards additional examples of what appear to be conflicting intuitions in ethics generated by abstractness and concreteness. For example, he points out that while people are ordinarily moved by Singer's (1972) theoretical arguments about the obligations of the affluent to the

needy, they have quite a different response to the concrete consequences of these principles when it comes to giving up their own goods and luxuries (Sinnott-Armstrong 2008, 222).

This concludes my survey of the wealth of data in support of the abstract/concrete asymmetry. There is a great deal of empirical evidence in support of the descriptive claim that concreteness has a significant influence on our intuitions in a wide range of different philosophical areas, and especially in regards to our judgments about moral responsibility. I now turn to discussion of the variety of attempts to explain this asymmetry that have been offered thus far.

3 Explaining the abstract/concrete asymmetry

If we accept the evidence that concreteness significantly influences our responsibility judgments, what should we make of it? Here I will discuss four main competing explanations first presented and assessed by Nichols and Knobe (2007) in response to their own results. Then I will briefly discuss some alternatives to these models including Nichols and Roskie's (2008) appeal to two dimensional semantics, and Mandelbaum and Ripley's (2012) NBAR hypothesis.

3.1 Four models

In response to the data generated by their 2007 study Nichols and Knobe propose four possible models that might provide the best explanation for the abstract/concrete asymmetry: the *performance error model*, the *affective competence model*, the *concrete competence model*, and the *hybrid model*.

Before discussing these models in any detail it is helpful to make a few brief remarks about them more generally. First, they are all intended to provide a picture of the underlying

psychological processes that generate our ordinary judgments about moral responsibility. Determining what those processes actually amount to is a job for the psychologists. As such, each model discussed below will be little more than that: a *model*. However, there is a significant amount of work in psychology which might be appealed to in support of each of these models. Here I will try to at least gesture in the direction of some psychological support for each model as I proceed. Finally, each of these models is intended to provide an explanation for the way that our ordinary responsibility judgments are generated. They aim to capture what is going on under the surface when we make *competent* judgments of this kind. It is important to point out that these models are silent on whether or not these judgments are *correct*. As should by now be quite clear, the fact that our ordinary judgments *are* generated by a particular picture of our underlying psychological processes captured by one of these models does not entail that those judgments are *true* or correct.

Nichols and Knobe take some version of a *performance error model* to be the best explanation of the data generated by their own study, and so I begin with this model. According to the *performance error model* strong affective reactions somehow distort or bias our responsibility judgments. Because concrete cases describe particular agents performing particular actions that are often clearly morally wrong, concrete cases often elicit affective reactions. According to this model the abstract/concrete asymmetry can be explained as follows: judgments about responsibility generated by abstract cases are generated by an underlying psychological process that is not subject to any bias or distortion, but judgments generated by concrete cases are not. More specifically, when we are asked an abstract question about responsibility such as, "Is it possible for people in a deterministic universe to be fully morally responsible?" our answer usually depends on some underlying *theory*, one that we may of course

be committed to only tacitly. But, when we are asked a concrete question about responsibility such as, "Is Bob fully morally responsible for murdering your best friend to avoid paying his taxes?" something *gets in the way* of the process that utilizes the underlying theory and normally generates our responsibility judgments. What is "getting in the way" here is an emotional response like moral anger or indignation. This affective reaction interferes with the normal application of our tacit underlying theory of the criteria for responsibility. Whereas our response to the abstract question will be generated at least in part by the theory, the affective response generated by the concrete question interferes with the usual underlying psychological process. So, judgments generated by concrete cases are subject to a *bias* or *performance error* (Nichols & Knobe 2007, 671-672). Responses to abstract cases really do reveal features of our ordinary conceptual commitments regarding responsibility, whereas responses to concrete cases are merely the product of a degenerate version of these psychological processes distorted or biased by affective reactions.[98]

In sharp contrast to the performance error model one might instead argue that affect is actually an essential component of the underlying psychological processes which ordinarily generate our responsibility judgments. This is the main idea behind the *affective competency model*. If this kind of model is correct then our abstract theories of responsibility don't actually play any role in generating ordinary responsibility judgments. Instead, an affective reaction is what is required (likely in combination with some other psychological processes) to enable us to competently[99] judge whether an agent is responsible or not. The idea here is that our theoretical beliefs have little or nothing to do with how we actually make ordinary judgments about moral responsibility. So, we might explain the abstract/concrete asymmetry according to this model in

[98] Nichols and Knobe cite a variety of psychological work in support of the plausibility of this kind of model, including Kunda (1990) and Lerner et al. (1998).
[99] Here, again, "competency" should be interpreted as something akin to "normal," "ordinary," or "unbiased."

the following way: responses generated by abstract cases reflect only theoretical commitments that have little to do with the way that our ordinary judgments about responsibility are normally generated, whereas concrete cases which elicit an affective reaction enable the psychological processes that do generate these judgments.

Nichols and Knobe cite an array of psychological support for the affective competency model. In particular, they cite studies indicating that people with deficits in emotional processing sometimes offer bizarre patterns of responses to questions that require moral judgments (Nichols & Knobe 2007, 673).[100] These studies suggest that if no affective reaction occurs then these people have difficulty applying consistent moral criteria more generally. Furthermore, this model fits nicely with currently popular views in moral psychology, such as Haidt's (2001), which take our moral judgments to primarily be a product of intuition or affective reactions and not some form of moral reasoning.

The *concrete competence model* is similar to the affective competence model, though according to this model it is dimensions of concreteness other than affect which play an essential role in generating our ordinary responsibility judgments. According to this model, our responses to concrete cases are generated by a completely different, though still purely cognitive psychological process from the one that generates our responses to abstract cases. Nichols and Knobe discuss one version of this model that they find particularly appealing: that our intuitions in concrete conditions are generated by an innate "moral responsibility module" (2007, 673). According to this kind of view our ordinary responsibility judgments are the result of a "swift, automatic, and entirely unconscious" process that takes as input "information about an agent and his or her behavior and then produce[s] as output an intuition as to whether or not that agent is

[100] For further discussion of these patterns, especially as they have been evidenced in psychopaths, see Blair (1995), Blair et al. (1997), and Hauser et al. (2006).

morally responsible" (2008, 673). If this picture is correct then our judgments about abstract and concrete cases are generated by two separate, independent cognitive processes. While some combination of our theoretical beliefs about responsibility and moral reasoning generate judgments about abstract cases, a separate cognitive process – perhaps a moral responsibility module – generates our judgments about concrete cases. Because these two separate processes do not interact with one another, they might each yield very different judgments about responsibility, which would explain the asymmetry at issue.

Haidt's (2001) social intuitionist model of moral judgment perhaps fits even better with the concrete competence model than the affective competency model, though it might be reconciled with either. And support for the particular view that the independent cognitive process which generates judgments about concrete cases might be something like the module discussed by Nichols and Knobe can be found in the work of Fodor (1983), Leslie (1994), Dwyer (1999), Harman (1999), and Hauser (2006).

Finally, the three models discussed above are not all mutually exclusive. One might opt for some version of a *hybrid model* which combines various elements of all three. I will not attempt to discuss all of the different possibilities for a hybrid model here, but Sinnott-Armstrong's (2008) *dual systems hypothesis* provides an instructive example. Sinnott-Armstrong suggests that there might be a parallel between what look to be different, independent memory systems and the way that we make abstract and concrete judgments. He draws on Klein et al.'s (1996) distinction between *episodic* memory (which represents particular events) and *semantic* memory (which represents abstract properties and general traits) (Sinnott-Armstrong 2008, 222). Sinnott-Armstrong suggests that, like memory, our intuitions might be generated by different, independent systems. He argues that this would explain why philosophical intuitions

are so persistent – because those on both sides of the debate *have* both systems and so can feel the force of the intuitions appealed to by the opposing side. It would also provide an evolutionary explanation for *why* we have the conflicting intuitions that we do – in conjunction separate abstract and concrete systems have a great deal of utility. In regards to moral responsibility, while abstractly generated incompatibilist intuitions allow us to identify excuses quickly (such as being caused to act by factors beyond one's control) intuitions generated by the concrete system prevent us from overextending these principles (2008, 223). And part of the latter system might also involve storing representations of particular instances or paradigms where agents are fully morally responsible despite having been caused to act by external factors (2008, 223).

Sinnott-Armstrong's dual-systems hypothesis draws from several of the models above. In particular, the "concrete system" would likely appeal to features of both the affective and concrete competency models in that it might posit that some combination of affect and concreteness generate *some* of our competent judgments about responsibility. On the other hand, this model also allows that cognitive, abstract processes also generate some of our ordinary judgments. This picture is of course wildly speculative, as Sinnott-Armstrong himself admits (2008, 222), but presents a helpful example of what a particular hybrid model might look like.

This concludes discussion of the four main hypotheses currently on offer to explain the underlying causes of the abstract/concrete asymmetry. While these look to be the main options, I will conclude this chapter by canvassing a few alternatives.

3.2 Alternative explanations

It is no great surprise that not all explanatory models fit neatly into the four categories outlined by Nichols and Knobe. This is in part because Nichols and Knobe are primarily concerned with explanations for the abstract/concrete asymmetry in terms of the underlying psychological processes that generate our responsibility judgments. However, one might look elsewhere for a suitably deep explanation for this asymmetry. Here I will briefly discuss two such explanations proposed by Shaun Nichols and Adina Roskies (2008), and Eric Mandelbaum and David Ripley (2012).

Nichols and Roskies first offer a variety of possible psychological explanations for the data supporting the abstract/concrete asymmetry generated by their own experiment. First, the asymmetry might be explained by the fact that people process problems more "fully and accurately" when the questions they are presented with are personally relevant to them (2008, 378). Questions posed about the actual world are clearly more personally relevant to subjects than those about abstract possible worlds, even those described as much like our own. Second, the asymmetry might be explained by appeal to differences in motivation. In particular, it may be the case that people are more likely to believe things that they want to be true.[101] And it seems quite plausible that while we want to hold people responsible in the actual world, we care less about whether possible agents in alternative universes are responsible. Finally, Nichols and Roskies appeal to the role of affect. In particular, they gesture towards a version of the affective competence model and cite psychological research supporting the claim that emotional processes are recruited when we make at least some moral judgments.[102] Because possible worlds or alternative universes are more removed from our own, and as noted above we may be less

[101] For arguments and data in support of this claim see Kunda (1999).
[102] For example, they refer the reader to Greene et al. (2001), Moll et al. (2002), and Schaich Borg et al. (2006).

personally invested in what holds true in them, it seems plausible that responding to questions about such universes involves different emotional areas of our brains than questions about the actual world.

However, Nichols and Roskies also offer a *non-psychological* explanation for their results. This explanation appeals to two-dimensional semantics and in particular what they call *analytical conditional analysis*. They suggest that applying two-dimensional semantics to 'moral responsibility'[103] allows us to capture the fact that ordinary people seem to have a *non-negotiable* intuition that agents in the actual world are at least sometimes responsible. They then offer the following conditional analysis of the concept of moral responsibility:

> If the actual world is indeterministic, then moral responsibility is compatible with determinism; else in the actual world compatibilism is true and is true of moral responsibility in other worlds. (2008, 384)

So, whether or not our concept of moral responsibility is compatibilist or incompatibilist depends on whether or not determinism is true in the actual world. This analysis seems to fit with the results generated by Nichols and Roskies' study, and it explains why our intuitions vary with whether or not the world we are considering is taken to be the actual world. Furthermore, it allows us to understand the concept of responsibility despite the fact that we do not know whether the actual world is deterministic or indeterministic, while the concept itself still depends to some degree on whether or not determinism is true.

Mandelbaum and Ripley (2012), offer a further alternative explanation for the abstract/concrete asymmetry. Taking Sinnott-Armstrong's dual systems hypothesis as a starting point they acknowledge that this explanation has the appeal of allowing that apparently

[103] Whereas sentences containing the term 'moral responsibility' have both a C-intension (the set of truth values of the sentence evaluated at every possible world, taking one particular world as actual) and an A-intension (a function which evaluates the truth value of the sentence at every possible world, taking the world at which it is being evaluated as actual) (Nichols & Roskies 2008, 383).

inconsistent judgments of responsibility can arise even under conditions that are not affectively loaded (2012, 356). However, they point out that the dual systems hypothesis has two serious shortcomings. First, it should predict that concrete cases will *all* (even those that concretely describe mundane actions such as buttering a slice of bread) produce an increase in judgments of moral responsibility[104], but this does not seem plausible (2012, 357). Second, the dual systems hypothesis seems to confuse the *stimuli* that cause representations with the *vehicle* of representation, and fails to provide a satisfying explanation for why abstract stimuli get encoded by abstract representations (and likewise for concrete stimuli and representations) (2012, 357-358). Mandelbaum and Ripley conclude that proponents of a dual systems hypothesis like Sinnott-Armstrong's face a dilemma. They can either explain the connection between the representational system and abstract/concrete judgment in question by appeal to affect, in which case the view just boils down to either an affective competence or affective performance error model, or they can appeal to differences in belief states, in which case appeal to distinct representational *systems* is no longer necessary (2012, 358).

Mandelbaum and Ripley opt for the latter and propose a new explanation for the abstract/concrete asymmetry. Their explanation combines general principles of cognitive dissonance theory (the hypothesis that having inconsistent beliefs puts the believer in a negative emotional state or creates a desire to relieve that dissonance) and a single hypothesized tacit belief shared by the majority of participants in the studies which support the asymmetry (2012, 359). They refer to the tacit belief doing much of the explanatory work here as NBAR, or "Norm Broken, Agent Responsible" (2012, 359). If NBAR is correct then most people share a well-entrenched belief that whenever a norm is broken there is some agent responsible for

[104] You might, of course, take issue with this claim if you assume that the concept of responsibility only applies to actions with a moral valence and not morally neutral actions like buttering a slice of bread.

breaking it (2012, 359). And Mandelbaum and Ripley interpret what counts as a norm quite widely. They allow that as far as NBAR is concerned a norm is any world-to-mind state or mind-to-world expectation (2012, 365-366). And they take great care to explicate that they are not claiming that NBAR is true, just that most people tacitly endorse it (2012, 360).

This view has an appealing degree of simplicity (compared to the dual systems hypothesis) in that it allows that both concrete and abstract stimuli are processed in similar ways, or by the same psychological system or module. So, it avoids speculative commitments about our psychological architecture. And it looks to be capable of explaining much of the data in support of the abstract/concrete asymmetry. For example, most of the concrete conditions discusses above describe a bad action and it is reasonable to suppose, as Mandelbaum and Ripley do, that bad happenings are always situations in which we think that a norm has been broken (2012, 360). And in other less obvious cases, such as Nichols and Roskie's study, Mandelbaum and Ripley are able to offer further explanations in terms of NBAR. When it comes to Nichols and Roskies's results, for example, they suggest that while NBAR is a widespread and well-entrenched tacit belief about what holds true in the actual world, people may be less committed to it in worlds described as possible or alternative to our own.

Finally, Mandelbaum and Ripley argue that the NBAR hypothesis has some further explanatory power in addition to its relevance to the abstract/concrete asymmetry. For example, it seems capable of explaining the phenomenon of anthropomorphizing inanimate objects when they don't work as well as we expect them to (2012, 362). Mandelbaum and Ripley also identify some predictions that should hold if the NBAR hypothesis is correct: empirical results should confirm that even people who consciously deny NBAR do in fact tacitly hold the belief, and it should take people less time to responds to abstract cases because they do not put us in a state of

cognitive dissonance (2012, 364). More empirical work needs to be done to either confirm or disconfirm these predictions, but Mandelbaum and Ripley take them to be plausible.

Conclusion

Here I hope to have met the three main goals of this chapter: to clarify what the abstract/concrete distinction amounts to, to survey some of the empirical data which provides evidence for the descriptive claim that abstractness and concreteness have very different influences on our responsibility judgments, and to present some of the most promising explanations for this asymmetry currently on offer. In the next chapter I will argue that making use of some form of *competence* model can provide revisionists with the necessary tools to respond to the normativity-anchoring problem, and that there are good philosophical reasons for thinking that this particular model is plausible in addition to the empirical considerations discussed in this chapter.

Chapter Six

Defending a New Methodological Assumption: MAP

Introduction

The main goal of this chapter is to propose and defend a qualified methodological assumption that I argue revisionists can and should accept, one that can allow them to avoid the normativity-anchoring problem while preserving much of the skeptical spirit used to motivate revisionism. This methodological assumption respects the influence of abstractness and concreteness on our responsibility judgments discussed in the previous chapter, and identifies some judgments generated by concrete cases as having a privileged epistemic status. I will say much more about this assumption later in the chapter, but will attempt a first pass at characterizing it here:

> **MAP:** We lack good epistemic reasons for thinking that, in general, our intuitions inform us about what moral responsibility is. But *some* of these intuitions have a privileged epistemic status that allows them to play a (defeasible[105]) evidentiary role in our moral responsibility theorizing. Judgments which are **convergent** and generated by **concrete** cases that elicit a strong **affective** response have this status.

Here I will argue that we are justified in accepting MAP as a basic assumption about the kind of responsibility judgments a theory of moral responsibility should respect. And, if these arguments are correct, acceptance of MAP allows revisionists to preserve much of the motivation for revisionism while avoiding the normativity-anchoring problem.

---

[105] I will discuss some possible defeaters in further detail in Section 3.

The arguments offered in this chapter proceed in several stages. First, in Section 1 I offer some philosophical support for the initial plausibility of MAP. These considerations are intended to lend plausibility the contingent claim that concreteness and affectivity play an important role in generating competent responsibility judgments. So, they are intended to make plausible some version of a concreteness or affective competency model discussed in the previous chapter. As such further empirical support for the plausibility of MAP can also be found in the psychological work which lends support to these two models discussed in Chapter 5. It is important to note that the arguments in Section 1 are in no way intended to show that concreteness and affectivity generate *true* or *correct* responsibility judgments. Whether or not we are justified in assuming that they do will be the subject of Section 2. Rather, the arguments in Section 1 are intended to show only that we have good reason to think that concreteness and affectivity play a necessary, enabling role in the way in which we do actually make normal (unbiased or undistorted) responsibility judgments.

In Section 2 I argue that we are justified in accepting MAP. Here I present both a conservative and ambitious strategy available to the proponent of MAP. However, the primary focus of this section is the conservative strategy, which is to offer a series of companions in guilt style arguments for MAP. I take this strategy to be sufficient to the task of justifying our acceptance of MAP. In Section 2.1 I discuss the contours of this general argumentative strategy, make explicit the distinction between different versions of this strategy, and identify some of its general strengths and weaknesses. In Section 2.2 I identify an analogy between the features of a particular class of moral judgments and the class of responsibility judgments identified by MAP. I argue that we already take these moral judgments to have a privileged (or at least uncontroversial) epistemic status in our moral theorizing, and that the features they share with

the responsibility judgments identified by MAP provide the best explanation for this status. I then use this analogy as the basis for a series of companions in guilt style arguments for our acceptance of MAP in Section 2.3. Taken together, I conclude that these arguments are sufficient to justify our acceptance of MAP. However, for those who remain skeptical I suggest an additional ambitious strategy that the proponent of MAP might also pursue. This strategy appeals to the fact that responsibility is a moral concept, and so there is good reason to take MAP to be an appropriate methodological assumption for our theorizing about responsibility in light of the analogy identified in Section 2.2. I conclude that the arguments in Section 2 establish that we are in fact justified in accepting MAP, and should take widely convergent judgments generated by concrete cases which elicit a strong affective response to have an adequate epistemic status to constrain our theorizing about moral responsibility.

In Section 3 I present and assess some potential objections to the arguments in Section 2. First, one might object that none of our responsibility judgments are in fact convergent to the degree necessary to make them sufficiently similar to the moral judgments identified in Section 2.2. Second, one might argue that the moral judgments identified in Section 2.2 do not have the epistemic status I attribute to them. Third, one might argue that even if the analogy identified in Section 2.2 holds, it is not epistemically relevant. Or, one might argue that there are disanalogies between the two kinds of judgment in questions which act as defeaters to the claim that they share the same epistemic status. I argue that the proponent of MAP can successfully respond to each of these objections. Finally, in Section 4 I argue that acceptance of MAP allows revisionists to preserve the primary motivation for their view while avoiding the normativity-anchoring problem. I conclude that if the arguments in this chapter are successful then

revisionism remains a live, interesting, and potentially fruitful option in the contemporary philosophical debate on moral responsibility.

1 Why think that the judgments picked out by MAP are *competent*?

The first step towards meeting the goals outlined above is to provide support for the claim that concreteness plays a necessary, enabling role in the way that we make competent judgments about moral responsibility. Providing such support will be the subject of this section. Before doing so it may be helpful to make a few clarificatory remarks on what this claim amounts to. First the claim in question is contingent. It is a claim about how creatures like us, which evolved in the same sorts of ways that we did actually make normal judgments about responsibility. The claim itself is therefore silent on whether or not we could have come to make normal responsibility judgments in different ways. Second, this claim is also silent on whether or not these judgments are correct, truth-tracking, or refer to any metaphysically robust properties in the world. Here 'competent' is intended to convey only that the judgments in question are produced *in the normal sort of way*, and that the normal processes that generate these judgments are not being biased or distorted. Finally, while this claim is distinct from MAP, it is important to laying the groundwork for this assumption. That is because MAP is intended to play a primary role in our theorizing about moral responsibility. If it turned out that the kind of judgments picked out by MAP as having a privileged epistemic status are generated in ways radically different from the way that we do normally form judgments about responsibility, then this would be a serious problem for those who wish to appeal to MAP. It

would be of little use to us in our responsibility theorizing to appeal to an assumption about the epistemic status of judgments that we rarely, if ever, actually make or could not make.

So what kind of support might be offered for the contingent claim that concreteness plays a necessary, enabling role in generating competent responsibility judgments for creatures like us? First, one might offer empirical support for this claim. I will not go into lengthy discussion of this kind of support here, but some of the empirical evidence relevant to MAP has already been mentioned in the discussion of the affectivity and concreteness enabling models in Chapter 5, Section 3. Two of the psychological considerations mentioned in the previous chapter are particularly relevant to the plausibility of MAP. First, data generated by studies on psychopaths and others with emotional processing deficits lends support to MAP.[106] This data indicates that people with such deficits have difficulty applying moral criteria, which suggests that affectivity in particular plays a necessary enabling role in making moral judgments more generally. Second, rationalist views in moral psychology like those of Kohlberg (1969) and Piaget (1932/1965) which take our moral judgments to primarily be the product of some form of moral reasoning have at present fallen largely out of favor in light of this kind of empirical evidence. Views which take our moral judgments to primarily be the product of intuition or affective reactions such as Haidt's (2001) have largely replaced them. Both the data and the current plausibility of this kind of comprehensive view in moral psychology lend empirical support to MAP (at the very least in regards to its affective component). If we have good empirical reason to think that affectivity plays an essential role in generating ordinary *moral* judgments, then affectivity likely plays an essential role in generating a particular subset of these judgments, judgments about moral responsibility.

[106] For example, see Blair (1995), Blair et al. (1997), and Hauser et al. (2006).

Second, one might offer philosophical support for MAP. The most prominent source of this kind of support can be found in P.F. Strawson's (1962) distinction between the participant and objective stances. While we can temporarily adopt the latter towards our fellow human beings as a resource (for example, in forming social policy or in our interactions with small children or those with serious psychological deficiencies), according to Strawson the participant stance is essential to our inter-personal relationships. Strawson characterizes the participant stance in terms of the attitudes involved in occupying this stance:

> What I have called the participant reactive attitudes are essentially natural human reactions to the good or ill will or indifference of others towards us, as displayed in *their* attitudes and actions. (1962, 80)

And these reactive attitudes are themselves characterized as

> …the non-detached attitudes and reactions of people directly involved in transactions with each other…the attitudes and reactions of offended parties and beneficiaries…such things as gratitude, resentment, forgiveness, love, and hurt feelings. (1962, 74-75)

According to Strawson, occupying the participant stance towards an agent and viewing them as the appropriate subject of the reactive attitudes is a necessary condition for that agent to be a candidate for our attributions of moral responsibility.

What does this view have to do with concreteness? First, because occupying the participant stance towards an agent is a necessary condition for that agent to be an appropriate candidate for *any* attributions of moral responsibility, if this view is correct then occupying the participant stance is also a necessary condition for making competent responsibility judgments. Without occupying this stance we could never make such judgments. If this is correct then

making competent responsibility judgments will require viewing an agent as the appropriate subject of our reactive attitudes, for example gratitude, resentment, forgiveness, love, and hurt feelings. And there are at least some background conditions that must be satisfied in order for us to hold these attitudes in the first place. For example, there must be a particular *agent* under consideration as the appropriate target of these attitudes, and a particular *action* that these attitudes are respondent to. A minimal degree of reflection reveals that we do not hold attitudes like gratitude and resentment in a vacuum, nor can they be generated by abstract musings. I cannot, for example, genuinely summon up a feeling of resentment towards an agent, *X*, whom I am told to imagine has committed some wrong action, *Y*. This is not to say that it is impossible to occupy the participant stance towards hypothetical agents described under a set of possible circumstances. Rather, the point is that it looks as though a sufficient degree of detail – or *concreteness* – is a necessary background condition for an agent to be the appropriate subject of our reactive attitudes. We simply cannot generate these attitudes without it. In order to feel resentment, or gratitude, or love towards an agent, I must know something *about* that agent, and the particular act in question.

Here the point might be put in terms of the following argument:

1. Occupying the participant stance is a necessary condition for making attributions of moral responsibility. (AP; call this "the Strawsonian view")
2. Occupying the participant stance requires that we take agents to be the appropriate target of our reactive attitudes.
3. We can only take agents described in concrete scenarios to be the appropriate targets of our reactive attitudes.

4. Therefore, concreteness is a necessary condition for making *any* attributions of moral responsibility.
5. So, concreteness is a necessary condition for making *competent* attributions of moral responsibility.

So, the Strawsonian view looks to provide some philosophical support for the claim that competent judgments of responsibility require concreteness. On this view, *all* judgments of responsibility require concreteness. We simply cannot summon the reactive attitudes which constituted the practice under purely abstract circumstances. Furthermore, according to Strawson we are not even *capable* of abandoning the participant stance, and our commitment to these reactive attitudes is a "necessary feature of the general framework of human life" (1962, 84).

Of course, the above considerations in no way *prove* that our responsibility judgments require concreteness. They merely suggest that if one finds these basic features of Strawson's view plausible then there is good reason for thinking that our responsibility judgments require concreteness. While countless persuasive objections have been raised against other features of Strawson's account, the central role it attributes to the reactive attitudes is perhaps one of its least controversial features. So, I take the Strawsonian view to provide at least some *prima facie* philosophical support for the claim that competent responsibility judgments require concreteness.[107] And these considerations, in addition to the empirical support discussed above, support the contingent claim that concreteness and affectivity play a role in generating at least some of our ordinary responsibility judgments. So, MAP is at least initially plausible – we have

[107] Nahmias (2011) makes a similar suggestion regarding the role of the participant stance in enabling competent attributions of responsibility. He takes the importance of the participant stance to our attributions of responsibility to provide a possible explanation for why bypassing makes subjects less willing to attribute responsibility. Nahmias suggests that cases that elicit the intuition that someone's agency is being bypassed interfere with our ability to occupy the participant stance towards the agent in question, thus interfering with an important component of how we normally make attributions of responsibility. Dana Nelkin (2007) also suggests that interference with our ability to occupy the participant stance might interfere with our ability to make competent responsibility judgments.

good reason to think that at least some of our normal responsibility judgments will fall under the class of judgments picked out by this assumption.

2 Defending MAP

As discussed at the outset of the previous section the contingent fact that concreteness and affectivity do play a necessary, enabling role in generating competent judgments about moral responsibility does not entail that they should, or that we have any reason for thinking that these judgments actually track the truth in some way. The goal of this section is to provide some independent reasons for thinking that we are justified in taking these judgments to have an epistemic status adequate to constrain our theorizing about moral responsibility, and in doing so provide support for the qualified methodological assumption laid out at the start of this chapter, MAP:

> **MAP:** We lack good epistemic reasons for thinking that, in general, our intuitions inform us about what moral responsibility is. But *some* of these intuitions have a privileged epistemic status that allows them to play a (defeasible) evidentiary role in our moral responsibility theorizing. Judgments which are **convergent** and generated by **concrete** cases that elicit a strong **affective** response have this status.

How might one go about offering support for the latter part of this assumption, that at least some of our judgments about moral responsibility, namely those that are widely convergent and generated by concrete cases which elicit a strong affective response, have a privileged epistemic status? This will largely depend on what is meant by 'privileged epistemic status' and, in particular, what features of this class of intuitions confer this status. For the revisionist purposes at issue here, the privileged status in question cannot be grounded in the mere fact that this class

of judgments are competent, or that they are generated by the usual sorts of underlying psychological processes. If it did, then MAP would equate to something like the contingent claim discussed in the previous section. In order for MAP to do any work for revisionists in regards to the normativity-anchoring problem the privileged status of this class of judgments must be normatively grounded in some way.

In this section I will argue that we are justified in *assuming* that the responsibility judgments picked out by MAP are in some sense correct, reliable, or truth-truth tracking, because they share a number of epistemically relevant features with the kind of paradigmatic *moral* judgments that we already take to have this status and constrain our ethical theorizing more generally. While it would of course be difficult (if not impossible!) to *prove* that the responsibility judgments picked out by MAP are in fact correct, reliable, or truth-tracking, the fact that they share these epistemically relevant features with other judgments that we already take to have a privileged (or at the very least uncontroversial) epistemic status provides sufficient support for our acceptance of the assumption that they are. Or so I will argue in this section.

My arguments in support of MAP will make use of what is often referred to as a *companions in guilt* style argumentative strategy. In Section 2.1 I outline the general structure of this kind of argumentative strategy and identify what I take to be its general strengths and weaknesses. In Section 2.2 I establish an analogy between the responsibility judgments picked out by MAP and the kind of paradigmatic moral judgments that constrain our ethical theorizing. In Section 2.3 I argue that this analogy can be used to ground a series of companions in guilt arguments in support of MAP. While I take these arguments to be sufficient to justify our acceptance of MAP, in Section 2.4 I propose an additional strategy for offering support for this assumption. Taken together, I conclude that the arguments in this section establish that we can

and should accept MAP as a basic methodological assumption appropriate to our theorizing about moral responsibility.

2.1 Companions in guilt & companions in innocence

What I will here refer to as a *companion in guilt* argumentative strategy has been appealed to in a variety of different philosophical areas in a variety of different forms, often with insufficient attention paid to differences in structure and the implications of these differences. Here I hope to diffuse any initial skepticism about this general argumentative strategy by clearly identifying the different versions of it that one might offer, make explicit which versions I will appeal to in support of MAP, and attempt to identify what I take to be its general strengths and weaknesses.[108]

Hallvard Lillehammer provides an instructive description of the general form of companions in guilt style arguments:

> Companions in guilt arguments as I shall understand them here are arguments designed to defend the metaphysical or epistemological credentials of one set of claims by comparing them to a different set of claims with which they have some apparently problematic features in common. (2007, 4)

Call the first set of claims in question X-claims, and the second set Y-claims. The motivation for offering companions in guilt style arguments is often that the "metaphysical or epistemological credentials" of one set of claims, say the X-claims, have come under fire.[109] On the other hand, the other set of claims, the Y-claims, have thus far enjoyed an uncontroversial metaphysical or

[108] This discussion draws heavily on Hallvard Lillehammer's (2007) helpful analysis of the scope of this kind of argumentative strategy as it has traditionally been employed in defense of ethical objectivity.
[109] Though, as Lillehammer (2007, 4) points out, this is not the only possible motivation for employing a companions in guilt argument. I will discuss the different variations of the overall argumentative strategy shortly.

epistemological status. The proponent of a companions in guilt style argument must first establish that the apparently problematic features of the X-claims are in fact shared by the comparatively uncontroversial Y-claims. Once the relevant similarity has been established, the proponent of this kind of argument then concludes that, by parity of reasons, the metaphysical or epistemological credentials of these two sets of claims should stand or fall together.

There are at least two versions of a companions in guilt style argument. Which version on offer will depend on how we weigh the problematic nature of the X-claims against the apparently uncontroversial status of the Y-claims. If we take the problematic nature of the X-claims to outweigh our reasons for previously accepting the status of the Y-claims, then the argument in question is a genuine *companions in guilt* argument. The fact that the previously uncontroversial Y-claims share the same problematic features of the X-claims yields the conclusion that worries about the status of the X-claims also apply to the previously uncontroversial Y-claims. In other words, we ought to view the metaphysical or epistemological status of the Y-claims with the same degree of skepticism with which we view the X-claims. Hereafter I will refer to this version of the argumentative strategy as a CIG. Alternatively, if we take our reasons for previously accepting the uncontroversial status of the Y-claims to outweigh the problematic nature of the X-claims then the argument yields a very different conclusion. The fact that the features of the X-claims under fire are shared by the uncontroversial Y-claims yields the conclusion that we ought to confer the same uncontroversial status to the X-claims that we do the Y-claims. Hereafter I will refer to this version of the argument as a *companions in innocence* argument, or CII.

The difference between these two versions of the general companions in guilt argumentative strategy can be formalized as follows:[110]

CIG:

1. Feature (*f*) of a particular set of claims (*X*) appears to be epistemologically problematic.
2. A distinct set of claims (*Y*) are at present epistemologically uncontroversial.
3. *Y*-claims share feature *f*.
4. Our reasons for taking feature *f* to be epistemologically problematic outweigh our reasons for accepting the epistemological status of the *Y*-claims.
5. Therefore, *Y*-claims share the same **problematic** epistemic status as *X*-claims.

CII:

1. Feature (*f*) of a particular set of claims (*X*) appears to be epistemologically problematic.
2. A distinct set of claims (*Y*) are at present epistemologically uncontroversial.
3. The *Y*-claims share feature *f*.
4. Our reasons for taking the epistemological status of the *Y*-claims to be uncontroversial outweigh weigh our reasons for taking *f* to be epistemologically problematic.
5. Therefore, *X*-claims share the same **uncontroversial** epistemic status as *Y*-claims.

While these two versions of the overall argumentative strategy are structurally identical they yield drastically different conclusions. Whether or not a companions in guilt style argument ends up being a CIG or CII depends entirely on the fourth premise and our weighing of the

[110] For simplicity I here restrict the status of the claims at issue to their epistemological status, as this is the kind of status that will be relevant to my arguments in support of MAP. It is important to note that this general argumentative strategy might, in principle, be employed in regards to a variety of different philosophically relevant statuses of a particular set of claims.

problematic features of the X-claims against the previously uncontroversial status of the Y-claims. If we take our reasons for concern about the relevant features of the X-claims to be stronger, then we ought to extend these worries to the Y-claims (CIG). But, if we take our reasons for accepting the status of the Y-claims to be stronger than our reasons for concern about the relevant features of the X-claims, then the fact that we accept the uncontroversial status of the Y-claims serves to diffuse our concerns about the X-claims (CII).

This distinction between a CIG and CII is not the only distinction that might be made regarding versions of this general argumentative strategy. One thing that CIGs and CIIs have in common is that they are both *arguments by analogy*. For each, the third premise – that Y-claims are relevantly similar to X-claims in that they share feature *f* – is essential to the validity of the argument. However, it is also possible to provide a companions in guilt style argument *by entailment* (Lillehammer 2007, 11). This version of the argument is structurally different from a CIG or CII. Call this version of the argument a CBE (*companions by entailment*). The structure of a CBE can be formalized as follows:

CBE:

1. A particular set of claims (*Y*) are epistemically uncontroversial.
2. We have strong, independent reasons not to deny (1). (AP)
3. *Y*-claims entail another particular set of claims (*X*).
4. *X*-claims are epistemically uncontroversial. (On pain of denying (1)).

As Lillehammer (2007, 11) points out, CBEs are more powerful than CIGs or CIIs, but face some serious limitations. The first limitation is that it is often difficult to establish an entailment relation. In particular, it is far more difficult to establish this kind of relation than the contingent similarity relation that CIGs and CIIs depend on. As Lillehammer notes:

> Thus, arguments by entailment fail if the less problematic companion [in the above argument, Y-claims] can, in principle, be completely understood without any commitment to the claims associated with the more problematic companion [in the above argument, X-claims]. (Lillehammer 2007, 11)

So, it is far more difficult to offer a successful CBE than a CIG or CII. A second limitation for CBEs is that even if the relevant entailment can be established this may fail to generate the intended conclusion. We might, under these circumstances, decide that the strong, independent reasons to accept (1) above are not strong enough. In this case, rather than establishing the intended conclusion the CBE in question turns out to be a *reductio*. Rather than showing that X-claims are epistemically uncontroversial, the fact that Y-claims entail such *problematic* X-claims instead establishes that Y-claims should themselves inherit this problematic status. The point here is that even if the first limitation for CBEs can be overcome and the relevant entailment relation established, the success of the argument will still depend on the degree of support for the assumption made in the second premise. If our independent reasons for accepting (1) are not strong enough (and, in particular, outweighed by the degree to which the problematic features of the X-claims are such), then the argument backfires. In other words, CBEs face a serious worry that "companionship in guilt can *cut both ways*" (Lillehammer 2007, 11; emphasis my own). So much the worse for CBEs, and I will not attempt to offer this version of a companions in guilt style argument in support of MAP. Rather, in what follows I will pursue a combination strategy and offer a series of CIIs and a CIG.

What are the strengths and weakness for this kind of argument by analogy? The most obvious strength is that it is far easier to provide support for the crucial premise this kind of argument depends on than the entailment relation that must be established for a successful CBE. Again, the crucial premise for CIGs and CIIs is the third premise – that the apparently

uncontroversial Y-claims share the apparently problematic feature of X-claims. Both CIGs and CIIs depend on this similarity claim. However, the fact that these arguments depend on a similarity claim also gives rise to perhaps the most serious weakness for this argumentative strategy. First, similarity is notoriously tricky. As Lillehammer and countless others have pointed out, "everything is similar to everything else in some respect of other."[111] In offering a CIG or CII one must take care to establish that the similarity appealed to not only holds, but is in fact *relevant* to the epistemic or metaphysical status of the particular class of claims in question.

What is required in order to establish the epistemic or metaphysical relevance of the similarity relation in question? Here Lillehammer's remarks are also instructive. He proposes two desiderata for establishing relevance: *centrality* and *lack of defeaters*.[112] First, if the two sets of claims in question share a feature that has little to do with their epistemic or metaphysical status then this similarity will not yield the conclusion that the uncontroversial status of the Y-claims is shared by the X-claims (or vice versa). For example, for a successful CII the shared feature appealed to must be *central* to the apparently uncontroversial status of the Y-claims. Another way to put the point is that the shared feature appealed to should play some role in our explanation of *why* the Y-claims have the uncontroversial epistemic or metaphysical status that they do.

Here an example may be instructive. CIIs are often employed in defense of the objectivity of moral claims. For example, Hilary Putnam argues that evaluative claims are relevantly similar to scientific claims, and as such the objective status that we already confer on scientific claims should be extended to evaluative claims. Putnam appeals to at least two ways in

---

[111] Lillehammer (2007, 13). See also Lewis (1973).

[112] Lillehammer (2007, 13) puts the point as follows:

The importance of the fact that two sets of claims share a certain feature therefore depends on at least two further issues. The first is the centrality of the relevant feature to the claims in question. The second is the presence of other relevant features that may otherwise distinguish the two sets of claims with respect to the question at hand.

which scientific and evaluative claims are similar: the inquiry which generates each kind of claim is both holistic and essentially perspectival.[113] However, even if we grant these similarities one might object that *these* features of scientific inquiry do not play an explanatory role in understanding why we confer the objective status to scientific claims that we do. So, even if we grant the similarity that Putnam appeals to, it might still fail to do the work it is intended to and establish the objectivity of evaluative claims. Further arguments are needed to show that the similarity identified plays some role in explaining why we accept the objectivity of scientific claims before it can justify conferring that same status to evaluative claims.

Second, even if a particular CIG or CII succeeds in identifying a similarity relation that is centrally relevant, this similarity might be outweighed or *defeated* by epistemically or metaphysically relevant disanalogies. Here I will continue to use Putnam's arguments for the objectivity of evaluative claims as an instructive example. Say we grant Putnam's claim that evaluative and scientific inquiry are similar in that they are both holistic and essentially perspectival, and also that these shared features are central to explaining why we take scientific claims to be objective. This might still fail to show that we ought to grant that evaluative claims have the same objective status, because the similarity identified (though relevant and central) is outweighed by some other relevant disanalogy. In this case, there are a variety of seemingly relevant disanalogies one might appeal to. Take, for example, convergence. While there is a great deal of convergence in scientific inquiry, lack of convergence in ethical inquiry provides ammunition for a wealth of arguments against ethical objectivity. And it seems reasonable to claim that convergence plays a central role in explaining the objective status of scientific inquiry. One might therefore object that even if Putnam is correct to point out that scientific and evaluative claims share some relevant similarities, there are relevant disanalogies between the

[113] See Putnam (1981, 1993, and 2002).

two sets of claims which outweigh these considerations. Namely, if we take convergence to be *more* central to explaining the objective status of scientific claims than the holistic and essentially perspectival features Putnam identifies, then the fact that scientific and evaluative claims actually differ in this respect counts as a defeater against their relevant similarities. A CIG or CII is therefore only successful in the absence of such defeaters.

I take worries about centrality and defeaters to be two of the most serious potential problems for this general argumentative strategy. The proponent of a CIG or CII must therefore take great care to avoid them, and I will address potential objections that my own arguments fall prey to these worries in Section 3. I turn now to the first step of my own use of this argumentative strategy in defense of MAP: identifying a relevant similarity between the responsibility judgments picked out by MAP and the paradigmatic moral judgments that we take to have an uncontroversial epistemic status in our ethical theorizing.

2.2 Establishing the similarity relation

The first step in my arguments in support of MAP is to establish an epistemically relevant similarity between the responsibility judgments picked out by this assumption and another class of judgments that we already take to have an privileged (or at least uncontroversial) epistemic status. Establishing that this similarity relation holds and is epistemically relevant is the purpose of this section.

Though it is rarely made explicit it looks as though something very much like MAP best explains why we take certain moral judgments to play a defeasible evidentiary role in our moral theorizing, but not others. Here I attempt to make this explicit using a series of examples. First, I introduce a particular case, Timmy the Toddler, which generates judgments paradigmatic of the

kind of judgment we take to be epistemically privileged and thus constrain our ethical theorizing. In practice we assume that moral theories should respect this kind of judgment, and we take it to be methodologically appropriate to use this kind of judgment in support of more general moral principles. I then contrast this case with a second case, Charlie the Chicken, which generates the kind of judgment that we think *does not* have this status nor can appropriately be used to constrain our ethical theorizing. Timmy and Charlie's cases are structurally identical, and I argue that the particular features which best explain the difference in the epistemic status of the kind of judgment generated by each are that judgments like those generated by Timmy's case are widely convergent, generated by concrete cases, and elicit a strong affective response, while judgments like those generated by Charlie's case are not.

Consider first the following case:

> **Timmy the Toddler**: Timmy is a toddler. One day his older brother, Jimmy, is left in charge of Timmy. Jimmy really enjoys torturing Timmy, and not in the colloquial sense that usually applies to siblings. For example, Jimmy enjoys tying Timmy up and burning him with lit cigarettes.

I assume that this case generates the same judgment in an overwhelming majority of adult human beings with normally functioning cognitive capacities. When asked whether or not they think what Jimmy does to Timmy is *wrong*, *bad,* or *harmful* I take it that most (if not all) normally functioning adults judge that it is. Furthermore, I take it that most (if not all) moral philosophers take this widespread judgment about Timmy's case to be uncontroversially correct and philosophically relevant. It is just the sort of case that we, in practice, expect our moral theories to respect. Here what I mean by "respect" can be cashed out in a number of different ways. First, all else being equal we would be very suspicious of any moral theory that yields the wrong result in Timmy's case, the result that what Jimmy does is *not* wrong, bad, or harmful. To the extent

that we value counterexamples in ethics a theory which entails that Jimmy's action has no negative moral valence looks to be about as clear a counterexample as one can find. So, as our moral methodology currently stands, failure to respect the judgment that Jimmy's act of burning Timmy with lit cigarettes for fun is wrong, bad, or harmful counts against a moral theory. Second, all else being equal we would consider it a merit of a theory if it not only yields the right results in Timmy's case, but also goes some way towards explaining *why* what Jimmy does is wrong, bad, or harmful. Again, as our moral methodology currently stands, it counts in favor of a theory if has the tools to explain the judgment that what Jimmy does to Timmy is wrong. So again, judgments generated by Timmy's case look to be paradigmatic of the kind of judgment that we take to have a privileged epistemic status in our moral theorizing as it stands. These judgments do in fact play a (defeasible) evidentiary role. Furthermore, Timmy's case is in no way unique, and there are countless other cases like it we might identify as generating judgments which share this status.

However, not *all* cases generate moral judgments which have this status. Take, for example, the following case:

> **Charlie the Chicken***:* Charlie is a chicken. One day the farmer who is his caretaker, Old McDonald, gathers some of the eggs that Charlie has recently laid. Old McDonald really enjoys eating omelets, and one day he prepares and eats one made with Charlie's freshly laid eggs.

This case and Timmy's case are *structurally identical*. They each describe a particular agent performing a particular action. However, we do *not* take judgments generated by Charlie's cases to have the same epistemic status that those generated by Timmy's do. To help elucidate this point, consider the relation between judgments generated by Timmy's case, judgments generated by Charlie's case, and the following two general principles:

Torture: torturing people for fun is wrong.

Veganism: consuming animal products is wrong.

In addition to the way in which we expect a moral theory to respect the widespread judgment generated by Timmy's case discussed above, we might also cite the judgment that Jimmy's act is wrong, bad, or harmful as providing some direct support for Torture. Or, we might use this judgment as a premise, or support for a premise, in an argument intended to establish Torture. Either of these is, in practice, a respectable methodological move to make. However, the same cannot be said for judgments generated by Charlie's case. Presenting someone who does not already assent to Veganism with Charlie's case will not help motivate the principle further. Nor would it provide any non-question begging support in an argument for Veganism or for premises in an argument for Veganism. At the very least, such an appeal would not be persuasive. It is highly unlikely that anyone who does not already assent to the principle that consuming animal products is wrong will be moved by judgments about Charlie's case to accept Veganism. So, though structurally similar, judgments generated by Charlie's case do not play the same methodological role that judgments generated by Timmy's do. At least as things currently stand, they do not have the same privileged epistemic status that those generated by Timmy's do. And this difference cannot be explained by appeal to structural differences between the two cases that elicit these judgments.

So, what features of judgments generated by Timmy's case and those generated by Charlie's case might account for the apparent difference in epistemic status? Why, *ceteris paribus* should a theory respect the former, but not the latter judgments?

It looks as though the best explanation for this difference depends on two features in particular. The first is the degree to which judgments generated by each of these cases are convergent. Judgments about Charlie's case do not converge in the same way that those

generated by Timmy's do. There is likely to be a great deal of disagreement among normal adults about whether or not what Old McDonald does is wrong, bad, harmful, or even has a moral valence of any kind. And this disagreement is likely to depend on a variety of independent considerations, not features of the case itself. Given the importance that we often attribute (implicitly or explicitly) to convergence in other domains it is reasonable to conclude that the high level of convergence in judgments about Timmy's case explains, at least in part, the privileged epistemic status that these judgments have in our moral theorizing. Judgments about Charlie's case do not enjoy this level of convergence, and so this difference can provide a plausible explanation, again at least in part, for the comparatively controversial epistemic status of these judgments.

However, appeal to this difference in convergence is inadequate to the task of explaining the epistemic difference between these two kinds of judgments fully. This can be seen by restricting the class of individuals whose judgments we are interested in to those whose judgments about Charlie's case *do* converge. In particular, we are interested in those whose judgments about Charlie's case converge in that they agree that what Old McDonald does is *not* wrong, bad, or harmful. Assume also that this restricted class of individuals is made up entirely of normal adult human beings (those who are not suffering from any psychological deficiencies), and so the judgments of the individuals in question also converge in respect to Timmy's case. They also agree that Jimmy's act is wrong, bad, or harmful. Consider a world in which only individuals in this restricted class exist. If the different epistemic status of judgments like those generated by Timmy's case and judgments like those generated by Charlie's case can be explained fully by convergence alone then, in the world we are considering, these judgments should share the same epistemic status. But surely they do not. Even in such a world it is

difficult, if not impossible, to imagine that one's moral theory ought to respect judgments generated by Charlie's case in the same way as those generated by Timmy's. Nor does it seem likely that, in such a world, if one were to come into contact with someone who had never considered Veganism as a general moral principle that presenting them with Charlie's case would move them to accept it. But it does seem likely that, in such a world, judgments generated by Timmy's case would continue to have the same epistemic status that they do in the actual world. So, we are in need of something in addition to convergence to adequately explain the different epistemological statuses of these two kinds of judgments.

I think that the following, in addition to convergence, provides the best explanation for this difference: there is a *qualitative* difference between these two kinds of judgments that is epistemically relevant. Even if we restrict our considerations to those whose judgments about Charlie's case converge, they are likely to have qualitatively different responses to Timmy's case and Charlie's. In particular, they will be *horrified* by Timmy's case and feel *certain* that Jimmy's act is wrong, bad, or harmful. Timmy's case is, in an important way, *more concrete* than Charlie's in that it elicits a strong affective response on the part of individuals presented with it. We react to Timmy's case with some degree of moral anger, indignation, or resentment. Furthermore, this affective reaction to Timmy's case best explains *why*, at least introspectively, we feel certain that our judgment about Timmy's case is correct. Judgments about Charlie's case do not share this feature. Even amongst those who judge that what Old McDonald does is wrong it is doubtful that their affective response (if there is one) to his eating the omelet is anywhere near on par with their response to what happens to Timmy. And the fact that Charlie's case fails to elicit the same kind of affective response that Timmy's does best explains *why* we are less certain that our judgment about this case is correct. It also helps to explain why attempts to

support a principle like Veganism with Charlie's case are likely to be met with something like the incredulous stare. Perhaps a case describing, in graphic detail, the suffering imposed on a particular animal as a result of the human consumption of animal products would provide such support. Or, perhaps watching a documentary on the conditions animals living in factory farms must endure would do it. However, one thing is certain: Charlie's case will not.

So, it looks as though when we go about the business of moral theorizing we let some of our judgments about cases (like Timmy's) but not others (like Charlie's) move us towards certain conclusions. And we assume that our moral theories should, *ceteris paribus*, respect some of our judgments (like those generated by Timmy's case) but not others (like those generated by Charlie's case). Here I hope to have established that a combination of convergence, concreteness, and affectivity best explain why we take the former to have a privileged epistemic status in our moral theorizing.

If these arguments are correct then they suggest that something akin to MAP is already at play in our moral theorizing: we take convergent judgments generated by concrete cases that elicit a strong affective response to have a privileged epistemic status and constrain our moral theories.

2.3 Justifying MAP

In this section I argue that we are justified in accepting MAP as a methodological assumption about the epistemic status of a particular class of responsibility judgments that our theories of responsibility ought to respect. Here I offer both a conservative and ambitious strategy. The conservative strategy is to provide a series of companions in guilt style arguments based on the epistemically relevant similarity between these responsibility judgments and the

class of moral judgments discussed in Section 2.2. In Section 2.3.1 I provide this series of arguments. However, for those who remain skeptical in Section 2.3.2 I propose an additional, ambitious strategy for defending MAP. This strategy is to argue that there are in fact independent considerations for thinking that MAP is an *appropriate* methodological assumption for theorizing about moral responsibility in particular. I begin with the conservative strategy.

2.3.1 The conservative strategy: companions in guilt & innocence

If we grant that the similarity relation laid out in Section 2.2 holds then it can be used to establish a series of companions in guilt arguments in support of MAP. These arguments take the form of a strong CII, a weak CII, and finally a CIG argument. If one accepts the first, the strong CII argument, then this is all that is needed to establish the claim that we are justified in accepting MAP. However, for those not persuaded by the strong CII argument, a weaker version can also do the work required. Finally, I argue that those unwilling to accept the conclusion of either CII argument are left in an uncomfortable position in regards to our moral methodology, and I establish this by way of a CIG argument.

I begin with the strong CII argument. This argument has the same structure as the general form discussed in Section 2.1, though it also employs the initial revisionist assumption that *all* of our judgments about moral responsibility are epistemically problematic or controversial (or, in other words, appeal to the skeptical claim discussed in Chapter 1). The argument goes as follows:

The Strong CII:

1. Judgments about moral responsibility which are convergent and generated by concrete cases which elicit a strong affective response appear to be epistemically problematic. (From the revisionist AP)
2. Judgments like those generated by cases like Timmy's which are convergent and generated by concrete cases which elicit a strong affective response have, at present, a privileged epistemic status in our moral theorizing.
3. The judgments referred to in (1) and (2) share the following features: convergence, concreteness, and affectivity.[114]
4. Our reasons for accepting (2) outweighed our reasons for accepting (1).
5. Therefore, the judgments identified in (1) share the same *privileged* epistemic status as those identified in (2).

The main line of reasoning appealed to here is that if we grant that judgments like those generated by cases like Timmy's have a privileged epistemic status in our moral theorizing then, by parity of reasons, we should extend the same epistemic status to the particular class of responsibility judgments picked out by MAP. And this because there is an epistemically relevant similarity between the judgments identified in (2) and those picked out by MAP: they share the features of convergence, concreteness, and affectivity.

There is, however, an obvious objection to this strong CII argument, namely that it is not clear that the arguments in Section 2.2 offer adequate support for (2). In particular, I have not provided sufficient arguments for taking judgments generated by cases like Timmy's to in fact

[114] I am here assuming that these features are relevant to the epistemic status of both kinds of judgments. I will return to the question of whether or not this assumption is plausible in Section 3, where I discuss a number of potential objections to both the CIG and CII arguments offered in this section.

have a *privileged* epistemic status, nor have I made clear what this status amounts to. However, identifying precisely how best to characterize the status of these judgments is a daunting task, and one that I will not attempt to undertake here as justifying our acceptance of MAP does not depend on the strong CII argument (though, for those willing to grant (2) the strong CII argument will be sufficient). Instead, we might offer a weaker, less controversial version of the argument. In particular, we might weaken (2) to something like the following:

> (2)': Judgments generated by cases like Timmy's have an *uncontroversial* epistemic status.

Here by 'uncontroversial' I mean something like the following: barring defeaters we generally assume that such judgments are correct. Rather than claiming that these judgments have a *privileged* epistemic status, we might instead opt for the less controversial claim that we do not (again, barring defeaters) take there to be anything epistemically problematic about these judgments. The conclusion generated by this version of the CII argument is also weaker. This version establishes only that, by parity of reasons, we ought to extend the same uncontroversial epistemic status of the judgments identified in (2)' to the responsibility judgments picked out by MAP. So, the formulation of MAP supported by this version of the CII argument will also be weaker, and will identify the class of responsibility judgments picked out as having, again, merely an uncontroversial epistemic status. But this is all that is required for the purposes at hand. Recall that the motivation for defending MAP in the first place is to provide revisionists with a methodological assumption capable of preserving a general skepticism about most of our responsibility judgments (thus preserving the motivation for the descriptive/prescriptive distinction) while allowing that *some* have sufficient epistemic standing to constrain our responsibility theorizing (thus avoiding the normativity-anchoring problem). A weakened

version of MAP which picks out a set of our responsibility judgments that, barring defeaters, we can generally assume are correct will therefore do the job for revisionists. And so the weak CII argument is sufficient to justify our acceptance of a suitable version of MAP.

Finally, for those who remain skeptical about even the weak CII argument the proponent of MAP might appeal to the following argument: if we *do not* accept MAP then, by parity of reasons, we ought to reject its analogue in moral theorizing. But this would be problematic for two reasons. First, it is not at all clear that we *can* reject this underlying assumption in our moral theorizing. Such a rejection would entail reducing judgments about cases like Timmy's to the same epistemic standing as cases like Charlie's. In fact, depending on one's reasons for rejecting MAP it may even entail that judgments generated by cases like Timmy's have a particularly *problematic* epistemic status that judgments generated by cases like Charlie's do not. Not only would this result be radically counterintuitive, it would largely undermine our methods of moral theorizing as they currently stand. If nothing else, judgments about cases like Timmy's serve to constrain our moral theories. Of course, they do not constitute *indefeasible* evidence. But they are the closest thing to evidence that we have in this domain, and we do in fact treat them as such. And, it is not at all clear how we might proceed without them.

Those who reject MAP must therefore respond to the following CIG argument:

1. Judgments about moral responsibility which are convergent and generated by concrete cases which elicit a strong affective response are epistemically problematic. (From rejection of MAP)
2. Judgments generated by cases like Timmy's which are convergent and generated by concrete cases which elicit a strong affective response have, at present, an unproblematic epistemic status in our moral theorizing.

3. The judgments referred to in (1) and (2) share the following features: convergence, concreteness, and affectivity.
4. Our reasons for accepting (1) outweighed our reasons for accepting (2).
5. Therefore, the judgments identified in (2) share the same *problematic* epistemic status as those identified in (1).

But again, this conclusion would have serious negative consequences for our moral methodology as it currently stands. So, the above argument provides strong reason to accept MAP – rejecting it would, by parity of reasons, leave us in the uncomfortable position of (at best) taking judgments about cases like Timmy's and Charlie's to be on par epistemically or (at worst) taking judgments about cases like Timmy's to be especially problematic. I take this to be a cost to our moral theorizing high enough such that our reasons to avoid it will outweigh any independent worries about the epistemic status of the responsibility judgments identified by MAP.

To sum up, I take this series of companions in guilt style arguments to provide sufficient justification for our acceptance of MAP. If one is willing to grant that judgments generated by cases like Timmy's have a privileged epistemic status in our moral theorizing then the strong CII establishes that, by parity of reasons, we should extend this status to the responsibility judgments identified by MAP. However, for those unwilling to grant the second premise of the strong CII argument it is possible to offer a weaker, less controversial version of the argument. So long as one is willing to grant that judgments generated by cases like Timmy's are epistemically *uncontroversial*, the weak CII argument justifies an acceptable version of MAP. Finally, those who wish to *reject* MAP are faced with the above CIG argument. If we reject MAP then, by parity of reasons, our moral methodology is in need of radical overhaul. So, rejecting MAP looks to come at too high a cost.

I take this series of arguments to show that we can and should accept MAP as a methodological assumption in our theorizing about moral responsibility.

2.3.2 The ambitious strategy

While the above arguments are sufficient to show that we are justified in accepting MAP, it is also possible to opt for a more ambitious strategy and offer some positive support for this assumption. There are a number of ways that I think one might pursue this ambitious strategy and here I will sketch one that I take to be particularly fruitful. That is to emphasize, like Vargas, that responsibility is a *moral* concept. As discussed at length in earlier chapters, the kind of responsibility at issue in the philosophical debate is whatever it is that licenses our attributions of *moral praise and blame*. So, those who accept MAP can say more than merely that the judgments picked out by this assumption should stand and fall with a relevantly similar class of moral judgments. They might argue further that we should accept MAP because the same kind of assumption plays a central methodological role in our moral theorizing.

This ambitious strategy depends on the claim that, conceptually speaking, RESPONSIBILITY is nearer in kind to concepts like JUSTICE, FAIRNESS, and GOODNESS. So, the methods we employ for providing an account of responsibility should align with the methods we employ for theorizing about concepts like these. And, even if one takes RESPONSIBILITY to be less straightforwardly a moral concept than those just mentioned, it is surely more akin to these concepts than paradigmatic metaphysical concepts such as TIME or PERSISTENCE. So, at the very least, we should err on the side of taking the methodological constraints on theorizing about the former, not the latter, to be appropriate to theorizing about moral responsibility. If this is correct, then those who depart drastically from the general methods of ethical theorizing in

their theorizing about moral responsibility are in need of some explanation for why this departure is appropriate. Why treat MORAL RESPONSIBILITY like the straightforwardly metaphysical concepts mentioned above?

The general form of the ambitious strategy might be summed up as follows: the proponent of MAP should emphasize the *moral* part of moral responsibility. Once we attend to this important feature of the concept, there may in fact be a burden of proof on those who reject this assumption to explain why alternative methodological constraints on our responsibility theorizing (which depart from the constraints of our ethical theorizing) are appropriate in this particular domain.

3 Objections

Here I survey what I take to be three of the most salient potential objections to the above arguments in support of MAP. First, one might be skeptical about whether or not there are any responsibility judgments that have the features identified by MAP. In particular, one might claim that none of our judgments about responsibility are convergent to the same extent that the moral judgments identified in Section 2.2 are. Second, one might deny the claim that judgments generated by cases like Timmy's do in fact have the privileged (or at least uncontroversial) epistemic status that I attribute to them. Third, one might argue that the CIG and CII arguments presented above fall prey to traditional objections to companions in guilt style arguments. In particular, they might argue either that the shared features of the moral and responsibility judgments identified in both of these arguments are not epistemically relevant, or that there are relevant disanalogies between these two kinds of judgments which constitute defeaters to the claim that the shared features identified make them companions in guilt or innocence.

3.1 Objection #1: none of our responsibility judgments have the features identified by MAP

The first objection concerns whether or not it is reasonable to suppose that any of our actual responsibility judgments do in fact have the features identified by MAP. If not, then appeal to this assumption will not help revisionists avoid the normativity-anchoring problem. This assumption will fail to provide revisionists with a class of judgments which can in fact be used to constrain their prescriptive account of responsibility.

This kind of objection might be offered in a number of different ways, depending on which of the features identified by MAP one takes to be problematic. Because many of our judgments about responsibility are in fact generated by concrete cases involving particular agents and actions, concreteness is not a likely candidate. I take it that one of the most plausible attempts to raise this version of the objection will instead target the features of affectivity and convergence. First, in regards to affectivity, one might claim that none of our responsibility judgments elicit an affective response on par with the kind of response elicited by Timmy's case. We simply do not get as worked up about praise and blame as we do about torturing for fun. So, none of our actual responsibility judgments have all of the epistemically relevant features identified by MAP.

Whether or not this claim is true is a contingent matter to be confirmed or disconfirmed empirically. So, I will not discuss it at length here. However, in light of the wealth of empirical data discussed in Chapters 4 and 5 I think it is reasonable for the proponent of MAP to dismiss this version of the objection. Much of this data suggests that many cases and questions about moral responsibility *do* elicit an affective response. This is especially true of cases describing a *wrong* action, which tend to elicit moral anger and resentment. Whether or not the strength of this response is in fact equivalent to that of our response to cases like Timmy's is, again, an

empirical question. But given the data as it stands it seems plausible to claim that at least some of the cases that generate judgments about moral responsibility do elicit a strong affective response. Take, for example, Nichols and Knobe's (2007) concrete condition in which Bill is attracted to his secretary, decides that the only way to be with her is to murder his wife and children, and correspondingly decides to do so. I take it that most readers do in fact have a strong affective response to this case. And, there are countless other cases like it in the literature. While future empirical work may show that this affective response is not on par with the response elicited by cases like Timmy's I take assuming that many of our responsibility judgments are on par to be reasonable in light of the current evidence.

However, there is a much stronger version of the objection at hand. Rather than focusing on affectivity, one might argue that none of our responsibility judgments are in fact convergent to the extent that judgments generated by Timmy's case are. Given the apparent intractability of many philosophical debates about moral responsibility this looks initially plausible. Many of these debates boil down to bedrock disagreements about the correct judgments about a particular case or cases. Take, for example, disagreement about Frankfurt cases and Pereboom's Four Case Manipulation Argument. In light of such disagreement, why think that any of our judgments about moral responsibility actually converge to the same degree that judgments generated by cases like Timmy's do?

Given the extent of disagreement in the philosophical literature, the proponent of MAP must take this objection seriously. If it were in fact true that none of our responsibility judgments are adequately convergent to have the suite of epistemically relevant features identified by MAP, this would again render MAP irrelevant to revisionists' purposes. However, it is not at all clear that *none* of our responsibility judgments are adequately convergent. While it is true that there is

a great deal of disagreement regarding the judgments generated by many of the most prominent cases offered in the philosophical literature, this does not extend to *all* cases intended to generate judgments about moral responsibility. And all the proponent of MAP requires is that *some* of our responsibility judgments are adequately convergent. They can even allow that the class of judgments that have this feature is rather small. So long as there are some such judgments, these judgments can serve to constrain our theorizing about moral responsibility in a non-trivial way.

However, even if we grant that it is possible that some of our responsibility judgments are adequately convergent this objection still places a burden on the proponent of MAP to provide some evidence for thinking that they are. Again, this will require appeal to empirical work. It is unlikely that the proponent of MAP will be able to convince an interlocutor offering this kind of objection that some of our responsibility judgments are adequately convergent with *a priori* considerations alone. But again, the existent empirical data suggests there is good reason for the proponent of MAP to be optimistic. For example, many of the cases discussed in Chapters 4 and 5 generate judgments that are convergent to a high degree. Nichols and Knobe's (2007) data is again relevant here. In the concrete condition described above 72% of subjects judged that Bill was fully morally responsible for his action (Nichols & Knobe 2007, 670). Furthermore, many of the cases presented by Nahmias et al. (2006, 2007) generated agreement in more than 80% of subjects. And these examples do not exhaust the data in support of the claim that there is in fact a high degree of convergence for many of our responsibility judgments. So, as it stands, the existent empirical evidence looks sufficient to assuage worries about lack of convergence based on the apparent intractability of the contemporary philosophical debate alone.

The proponent of this version of the objection might still be unsatisfied with the above appeal to the existent empirical data. More specifically, they might argue that even the high

percentages of agreement cited above fall short of the level of convergence required for at least some of our responsibility judgments to be sufficiently similar to the moral judgments generated by cases like Timmy's. Here the proponent of MAP and those pursuing this objection will reach a kind of impasse. It is difficult to see what kind of non-question begging arguments might be offered in support of the claim that the level of agreement cited above is or is not adequate to establish the relevant similarity at issue. However, there looks to be a clear burden on the proponent of the objection to show that the level of convergence cited above (70-90%) is *not* adequate. Requiring a higher degree of convergence would seem, on the face of things, unreasonable. How much convergence would be adequate? 100% agreement is clearly too stringent a requirement. Even cases like Timmy's are unlikely to enjoy *that* degree of convergence. So, where should we draw line?

I am skeptical that the proponent of this version of the objection will be able to identify a level of convergence more stringent than the degree already suggested by the existent data that is not completely arbitrary. And so again I take the burden to be on the proponent of the objection to provide arguments for why this level of convergence is not adequate to establish the relevant similarity. Until then, the degree of convergence suggested by the existing data is enough to forestall this kind of objection to MAP.

3.2 Objection #2: judgments generated by cases like Timmy's are epistemically problematic

The remaining two objections focus on the epistemic status of the moral judgments discussed in Section 2.2. The first is to deny that these judgments have the epistemic status I attribute to them. This argument might take one of two forms. First, one might deny that these judgments have an epistemically *privileged* status. Establishing that these judgments are

genuinely privileged requires something further than appeal to the fact that we do generally assume that judgments like this are true in practice. This assumption might be mistaken, and in order to establish that these judgments are genuinely epistemically privileged further arguments are required. For example, one might attempt to provide independent considerations that count in favor of thinking that these judgments are correct or reliable. However, this version of the objection fails to undermine support for MAP, as only the strong CII depends on the claim that the judgments picked out by MAP have a genuinely *privileged* epistemic status. So, the proponent of MAP can grant the objection and go on to offer the weak CII argument, which requires appeal only to the claim that these judgments are uncontroversial.

The second form of the objection raises a more serious worry. One might deny that the moral judgments discussed in Section 2.2 are epistemically uncontroversial, and argue that even the weak CII argument fails. This version of the objection is based on one of the general weaknesses for companions in guilt argumentative strategies discussed in Section 2.1 – that these arguments run the risk of *cutting both ways*. Once we have established that a relevant similarity holds between the features of some of our moral judgments and some of our responsibility judgments, what this similarity establishes regarding the epistemic status of each depends on the weighing of our reasons for thinking that one is epistemically uncontroversial against our reasons for thinking that the other is epistemically problematic. One might therefore deny the fourth premise of the weak CII argument and claim that our reasons for taking the responsibility judgments identified by MAP to be epistemically problematic outweigh our reasons for taking the moral judgments identified in Section 2.2 to be epistemically uncontroversial. In other words, the proponent of this kind of objection might bite the bullet and accept the conclusion of the CIG argument – that moral judgments generated by cases like Timmy's are in fact

epistemologically problematic. Rather than showing that the class of moral judgments and responsibility judgments in questions are companions in innocence the similarity relation I have established instead shows that these judgments are companions in *guilt*. The moral judgments in question should inherit the problematic epistemic status of the responsibility judgments.

This objection is also one that the proponent of MAP must take seriously, though they have a response available. It is not at all clear what independent reasons one might appeal to for thinking that the responsibility judgments in question are epistemically problematic which would outweigh our reasons for taking the moral judgments in question to be uncontroversial. The best candidate for providing such reasons is likely the affective feature of these judgments. For example, one might argue that we have good reason to think that the affective feature of the responsibility judgments identified by MAP actually *biases* or *distorts* these judgments. So, if there is little reason to accept the uncontroversial status of the analogous moral judgments short of the fact that we do in practice generally assume that they are true, the fact that we have good independent reasons for thinking the responsibility judgments picked out by MAP are biased or distorted (and thus epistemically problematic) is enough to outweigh our reasons for accepting the uncontroversial status of the moral judgments.

This kind of objection will have to be dealt with by the proponent of MAP on a case by case basis, but it again looks to fall short in light of the empirical data discussed in Chapters 4 and 5, and the Strawsonian considerations appealed to in Section 1 of this chapter. At the very least there is no definitive evidence that an affective response does in fact bias our responsibility judgments. Furthermore, if we find the Strawsonian view that the reactive attitudes are a necessary condition for making attributions of responsibility in the first place appealing, then there is good reason for thinking that, when it comes to responsibility judgments in particular,

affectivity actually *enables* these judgments. While these considerations do not constitute a definitive response to the objection at hand they do significantly diffuse it, at least when formulated in terms of appeal to the claim that it is the affective feature of the responsibility judgments picked out by MAP that are problematic. As it stands, there is empirical *and* philosophical support for the opposing claim that the affective component of these judgments is an essential feature of how our responsibility judgments are normally generated. While other considerations might provide reasons weighty enough to move us to bite the bullet and accept the conclusion of the CIG argument, it is not at all clear what they might be and defenders of MAP will have to assess them on a case by case basis.

3.3 Objection #3: relevance and defeaters

Finally, one might object to the above arguments by targeting one of the two weaknesses for companions in guilt argumentative strategies in general. That is to argue either that the similarity relation at the heart of the CII and CIG arguments above is *not epistemically relevant*, or to argue that there are disanalogies between the moral judgments and responsibility judgments in question which act as *defeaters* to the claim that these two classes of judgments share the same epistemic status.

The response to the latter version of this objection borrows from the arguments above. If there are such defeaters, it is not at all clear what they are. The proponent of MAP will again have to respond to this kind of objection on a case by case basis. They will have to weigh our reasons for taking the disanalogy in question to be epistemically relevant against our reasons for taking the features of convergence, concreteness, and affectivity to be. I will not attempt to hypothesize about what specific disanalogies the proponent of this kind of objection might hope

to offer. However a successful response to objections regarding relevance may also go some way towards diffusing worries about defeaters. And if the arguments in Section 2.2 are correct then the features of convergence, concreteness, and affectivity provide the *best explanation* for why judgments generated by cases like Timmy's have the epistemic status that they do. If this is right, then it is difficult to see how one might possibly argue that the similarity between these judgments and the responsibility judgments picked out by MAP is not epistemically relevant. The fact that these features provide the best explanation for the epistemic status of the moral judgments in question again means that these particular shared features ground precisely the kind of similarity that *should* be used in a successful CIG or CII argument.

So, I do not take objections appealing to defeaters or relevance to pose a serious threat to my arguments in support of MAP either. I turn now to discussion of how acceptance of MAP bears on the normativity-anchoring problem.

4 Avoiding the normativity-anchoring problem

Recall that the heart of the normativity-anchoring problem is as follows: revisionists do not have a way to bridge the gap between the descriptive, psychological claim that the responsibility system promotes something that we value (in Vargas' case, moral considerations-responsive agency) and the normative claim that our participation in the responsibility system should, all things considered, continue. And this is a unique problem for revisionism. Conventional responsibility theorists *assume* that the best descriptive account – the one that best aligns with our refined intuitions about cases and principles – is the one that is most likely to be correct, and so the account that we should ultimately endorse. Revisionists deny this, and are committed to the skeptical claim which does much of the work in motivating the distinction

between descriptive and prescriptive accounts of responsibility. So, abandoning the skeptical claim entirely does not look like a tenable option for revisionists. In order to successfully respond to the normativity-anchoring problem revisionists must find some way to bridge the gap between the descriptive and normative claims above without undermining one of the most basic motivating features of their view.

Identifying a particular class of responsibility judgments that have a privileged or uncontroversial epistemic status provides revisionists with a way to preserve their general skepticism about the proper role of intuition in our responsibility theorizing while avoiding the normativity-anchoring problem. They can appeal to this class of judgments as having an epistemic status adequate to ground prescriptive claims about what we *should* think about responsibility, while allowing that many of our responsibility judgments do *not* have this status. Acceptance of MAP allows revisionists to make such a distinction. According to MAP, responsibility judgments that are convergent and generated by concrete cases that elicit a strong affective response have a privileged (or at the very least uncontroversial) epistemic status in our responsibility theorizing. And, if the arguments above are correct, there are good reasons to accept MAP as a methodological constraint on our theorizing about moral responsibility.

Here one might raise an obvious objection: even if revisionists can justify our acceptance of MAP, the claim that the responsibility judgments it picks out have the epistemic status they do is itself a *descriptive* claim. If Vargas' own descriptive claim that we value moral considerations responsive agency is not enough to normatively anchor his prescriptive account of responsibility how is MAP supposed to do this work either?

There is an important difference between Vargas' claim that we value moral considerations-responsive agency and acceptance of MAP. The former expresses a descriptive

claim about our psychology, and so requires the intermediate steps discussed in Chapter 3 to do the anchoring work needed:

Psychological claim → Axiological claim → Normative claim

But, MAP does *not* express a claim about our psychology. It is a basic methodological *assumption* about the epistemic status of a particular class of responsibility judgments. If revisionists accept this assumption then, like the conventional theorists, they get the normative claim needed to do the required anchoring work for free. They need not rely solely on appeal to the promotion of some independent value in order to meet the external requirement of (C1) and justify our continued participation in the practice of moral praising and blaming. Because their prescriptive account is constrained by MAP, revisionists who accept this assumption can also appeal to the fact that we are continuing to participate in a practice sustained by a theory of responsibility that we have most reason to think tracks facts about when agents genuinely *deserve* moral praise and blame.

Like the methodological assumptions underlying conventional responsibility theorizing, accepting MAP allows revisionists to sidestep worries that appeal to facts about independent value (not facts about *responsibility*) fail to justify our continued participation in our responsibility-related practices. If revisionists are justified in assuming that some of our responsibility judgments are tracking the truth in some way then this diffuse the normativity-anchoring problem. The problem itself is generated by the fact that commitment to the skeptical claim blocks appeal to the basic assumption that the methodology used to generate a particular prescriptive account is, in some sense, tracking the truth about when agents genuinely deserve praise and blame. And accepting MAP allows revisionists to appeal to a qualified version of this assumption. Like conventional theorists, revisionists who accept MAP can respond to the

question, "But how do you know whether your account tells us anything about when agents deserve praise and blame, and can justify our continued participation in the responsibility system?" with some version of the incredulous stare. More charitably, they can point to the fact that, in addition to being naturalistically plausible, normatively adequate, and providing an answer to the internal question, their account does the best job of respecting the judgments about moral responsibility that we take to be, *ceteris paribus*, true. Like conventional theorists they too can claim that, based on the best methodology for responsibility theorizing that we have, there is good reason to think that their particular prescriptive account *gets things right* and provides us with the best account of when agents genuinely deserve moral praise and blame.

Furthermore, revisionists who accept MAP have an additional methodological advantage over conventional theorists. They might argue that the prescriptive account generated by their methodology does a *better* job and is *more likely* to get things right than the conventional theorist. While the conventional theorist is silent on *which* judgments about moral responsibility our theory should respect, revisionists can clearly identify a particular class of these judgments: those that are convergent and generated by concrete cases that elicit a strong affective response. And the general contours of revisionism also provide further tools for identifying which judgments ought to constrain our theory. For example, if a particular responsibility judgment that satisfies MAP presupposes an account of responsibility that is naturalistically implausible, for revisionists this would be a clear violation of the *ceteris paribus* clause of MAP. Likewise for judgments that presuppose accounts of responsibility that are clearly normatively inadequate, or fail to answer the internal question and track what we take to be normatively relevant features of agents in the world. So, revisionists can appeal to principled reasons for taking some of our responsibility judgments, but not others, seriously. At present, it is not clear that conventional

theorists can do the same.[115] Therefore, not only can revisionists who endorse MAP avoid the normativity-anchoring problem, there are additional reasons for thinking that revisionism has significant methodological advantages to conventional responsibility theorizing.

Conclusion

The primary goal of this chapter has been to argue that revisionists are justified in accepting MAP. First, Section 1 provides some philosophical and empirical support for the initial plausibility of MAP. Section 2 is concerned primarily with my arguments that we are justified in accepting MAP. These arguments focus primarily on the conservative strategy, which is to offer a series of companions in guilt style arguments in support of MAP. Section 2.1 outlines the overall structure, strengths, and weaknesses of this kind of argumentative strategy. Section 2.2 establishes an epistemically relevant similarity between the features of the responsibility judgments identified by MAP and the features of a class of paradigmatic moral judgments that we take to be epistemically privileged in our moral theorizing. I argue that these features – convergence, concreteness, and affectivity – also provide the best explanation for the fact that the moral judgments in question have the epistemic status that they do. Section 2.3.1 presents the conservative strategy for justifying our acceptance of MAP by way of a series of companions in guilt style arguments. Taken together, these arguments are sufficient to justify our acceptance of this methodological assumption. Section 2.3.2 suggests an additional, ambitious strategy for supporting MAP.

[115] I take this difference to also have the potential to explain some of the apparent intractability in the contemporary philosophical debate on moral responsibility. Often these debates boil down to bedrock disagreements about particular intuitions. As mentioned at the outset of this project, revisionism (and in particular MAP-style revisionism) provides an alternative to conventional theorizing that might allow some significant progress on these issues. In particular, MAP-style revisionism identifies a principled way in which we might distinguish intuitions our theory must respect from those it does not.

Section 3 canvasses some potential objections to the arguments in Section 2. First, one might object that none of our responsibility judgments are in fact convergent to the degree necessary to make them sufficiently similar to the moral judgments identified in Section 2.2. Second, one might argue that the moral judgments identified in Section 2.2 do not have the epistemic status I attribute to them. Third, one might argue that the similarity relation identified in Section 2.2 is not epistemically relevant, or is undermined by defeaters. I argue that none of these objections are persuasive, and conclude that we are justified in accepting MAP.

Finally, in Section 4 I discuss how acceptance of MAP allows revisionists to avoid the normativity-anchoring problem. Appeal to MAP allows revisionists to sidestep worries that their prescriptive account fails to justify our continued participation in the responsibility system because they have good reasons for taking a prescriptive account constrained by this assumption to be the best account of when agents genuinely deserve moral praise and blame.

If the arguments in this chapter are successful then MAP is a methodological assumption that revisionists can and should accept. It allows them to preserve much of the motivation for revisionism by preserving a healthy amount of skepticism about the epistemic status of many of our responsibility judgments, while still allowing that some of these judgments – those that are convergent, concrete, and affective – can and sometimes do inform us about moral responsibility and can appropriately constrain our best theory. Thus, I conclude that revisionism remains a live, interesting, and potentially fruitful option in the contemporary philosophical debate.

Bibliography

Balagaur, M. (1999). Libertarianism as a Scientifically Reputable View. *Philosophical Studies*, 93, 189-211.

Balagaur, M. (2004). A Coherent, Naturalistic, and Plausible Formulation of Free Will. *Noûs*, 38, 379-406.

Bernstein, M. (2007). Experimental Philosophy Meets Experimental Design: 23 Questions. Presented at the MidSouth Philosophy Conference, February 2007, accessed at:https://resources.oncourse.iu.edu/access/content/user/jmweinbe/Filemanager_Public_Files/BernsteinXPhi-meets-experimenatal-design.pdf.

Blair, R. (1995). A Cognitive Developmental Approach to Morality: Investigating the Psychopath. *Cognition*, 57, 1-29.

Blair, R., Jones, L., Clark, E., Smith, M., and Jones, L. (1997). The Psychopathic Individual: A Lack of Responsiveness to Distress Cues?" *Psychophysiology*, 34(2), 194-198.

Carey, S. (2009). The *Origin of Concepts*. New York: Oxford University Press.

Carlsmith, K. M., Darley, J. M., and Robinson, P. H. (2002). Why Do We Punish? Deterrence and Just Deserts as Motives for Punishment. *Journal of Personality and Social Psychology*, 83, 284-299.

Clark, R. (1993). Toward a Credible Agent-Causal Account of Free Will. *Noûs*, 27, 191-203.

Clark, R. (1996). Agent Causation and Event Causation in the Production of Free Action. *Philosophical Topics*, 24(2), 19-48.

Clark, R. (2003). *Libertarian Accounts of Free Will*. New York: Oxford University Press.

Chisholm, R. (1966). Freedom and Action. In K. Lehrer (ed.) *Freedom and Determinism*, 11–44, New York: Random House.

Cushman, F., Mele, A. (2008). Intentinoal Action: Two-and-a-Half Folk Concepts? In J. Knobe & S. Nichols (eds.), *Experimental Philosophy*, 171-188, New York: Oxford University Press.

Doris, J., Knobe, J., and Woolfolk, R. (2007). Variantism about Responsibility. *Philosophical Perspecitives*, 21, 183-214.

Double, R. (1991). *The Non-Reality of Free Will*. New York: Oxford University Press.

Double, R. (1996). *Metaphilosophy and Free Will*. New York: Oxford University Press.

Dwyer, S. (1999). Moral Competence. In K. Murasugi and R. Stainton (eds.) *Philosophy and Linguistics*, 169-190, Boulder, CO: Westview Press.

Ekstrom, L. (2002). Libertarianism and Frankfurt-style Cases. In R. Kane (ed.) *The Oxford Handbook of Free Will 1st edition*, 309-322, New York: Oxford University Press.

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

Feltz, A. (2007). The Knobe Effect: A Brief Overview. *Journal of Mind & Behavior*, 28, 265-277.

Feltz, A., and Cokely, E.T. (2007). An Anomaly in Intentional Action Ascriptions: More Evidence of Folk Diversity. In D.S. McNamara & J.G. Trafton (eds.) *Proceedings of the 29th Annual Cognitive Science Society*, 1748, Austin, TX: Cognitive Science Society.

Feltz, A., and Cokely, E.T. (2008). The Fragmented Folk: More Evidence of Stable Individual Differences in Moral Judgments and Folk Intuitions. In B.C. Love, K. McRae, and V.M. Sloutsky (eds.) *Proceedings of the 30th Annual Conference of the Cognitive Science Society*, 1771-1776, Austin, TX: Cognitive Science Society.

Feltz, A., and Cokely, E.T. (2009). Do Judgments about Freedom and Responsibility Depend on Who You Are? *Consciousness and Cognition*, 18, 342-350.

Feltz, A., Cokely, E., Nadelhoffer, T. (2009). Natural Compatibilism versus Natural Incompatibilism: Back to the Drawing Board. *Mind & Language*, 24(1), 1-23.

Fischer, J.M. & Ravizza, M. (1998). *Responsibility and Control: A Theory of Moral Responsibility*. Cambridge: University Press.

Fischer, J.M., Kane, R., Pereboom, D., and Vargas, M., (2007). *Four Views on Free Will*. Malden, MA: Blackwell.

Fodor, J. (1983). *Modularity of Mind*. Cambridge, MA: MIT Press.

Fodor, J. (1990). *A Theory of Content and Other Essays*. Cambridge, MA: MIT Press.

Fodor, J. (1998). *Concepts: Where Cognitive Science Went Wrong*. New York: Oxford University Press.

Ginet, C. (1990). *On Action*. Cambridge: University Press.

Ginet, C. (2002). Reasons Explanations of Action: Causalist versus Noncausalist Accounts. In R. Kane (ed.) *The Oxford Handbook of Free Will 1st Edition*, 386-405, New York: Oxford University Press.

Ginet, C. (2007). An Action Can be Both Uncaused and Up to the Agent. In C. Lumer and S. Nannini (eds.) *Intentionality, Deliberation, and Autonomy*, 243-256, Burlington: Ashgate.

Gosling, S. D., Rentfrow, P. J., and Swann, W. B. (2003). A Very Brief Measure of the Big-Five Personality Domains. *Journal of Research in Personality*, 37, 504-528.

Greene, J.D., Sommerville, R.B., Nystrom, L.E., Darley, J.M., and Cohen, J.D. (2001). An fMRI Investigation of Emotional Engagement in Moral Judgment. *Science*, 292, 2105-2108.

Haidt, J. (2001). The Emotional Dog and Its Rational Tail: A Social Intuitionist Approach to Moral Judgement. *Psychological Review*, 108(4), 814-834.

Haji, I. (1998). *Moral Appraisability*. New York: Oxford University Press.

Haji, I. (1999). Moral Anchors and Control. *Canadian Journal of Philosophy*, 29(2), 175-204.

Haji, I. (2002). *Deontic Morality and Control*. Cambridge: University Press.

Harman, G. (1999). Moral Philosophy and Linguistics. In K. Brinkmann (ed.) *Proceedings of the 20th World Congress of Philosophy: Volume I: Ethics*, 107-115, Philosophy Documentation Center.

Hauser, M. (2006). *Moral Minds: The Unconscious Voice of Right and Wrong*. New York: Harper Collins.

Heller, M. (1996). The Mad Scientist Meets the Robot Cats: Compatibilism, Kinds, and Counterexamples. *Philosophy and Phenomenological Research*, 56, 333-337.

Hurley, S. (2000). Is Responsibility Essentially Impossible? *Philosophical Studies*, 99, 229-268.

Jackson, F. (1998). *From Metaphysics to Ethics*. Oxford: University Press.

Kane, R. (1996). *The Significance of Free Will*. New York: Oxford University Press.

Kane, R. (1999). Responsibility, Luck, and Chance: Reflections on Free Will and Indeterminism. *Journal of Philosophy*, 96, 217-240.

Kane, R. (2007). Libertarianism. In J.M. Fischer, R. Kane, D. Pereboom, and M. Vargas (eds.) *Four Views on Free Will*, 5-43, Malden, MA: Blackwell.

Klein, S. B., Loftus, J., and Kihlstrom, J. F. (1996). Self-Knowledge of an Amnesic Patient: Toward a Neuropsychology of Personality and Social Psychology. *Journal of Experimental Psychology*, 125, 250-260.

Knobe, J. (2003). Intentional Action and Side Effects in Ordinary Language. *Analysis*, 64, 181-187.

Knobe, J., and Nichols, S. (2008). An Experimental Philosophy Manifesto. In J. Knobe and S. Nichols (eds.) *Experimental Philosophy*, 3-16, Oxford: University Press.

Kohlberg, L. (1969). Stage and Sequence: The Cognitive-Developmental Approach to Socialization. In D.A. Goslin (ed.) *Handbook of Socialization Theory and Research*, 347-480, Chicago: Rand McNally.

Kripke, S. (1980). *Naming and Necessity*. Cambridge, MA: Harvard University Press.

Kunda, Z. (1990). The Case for Motivated Reasoning. *Psychological Bulletin*, 108(3), 480-498.

Kunda, Z. (1999). *Social Cognition*. Cambridge: MIT Press.

Laurence, S., and Margolis, E. (1999). Concepts and Cognitive Science. In E. Margolis and S. Laurence (eds.) *Concepts: Core Readings*, 3-82, Cambridge, MA: MIT Press.

Lerner, J., Goldberg, J., and Tetlock, P. (1998). Sober Second Thought: The Effects of Accountability, Anger, and Authoritarianism on Attributions of Responsibility. *Personality and Social Psychology Bulletin*, 24, 563-574.

Leslie, A. (1994). ToMM, ToBY, and Agency: Core Architecture and Domain Specificity. In L. Hirschfeld and S. Gelman (eds.) *Mapping the Mind*, 119-147, Cambridge: University Press.

Lewis, D. (1973). *Counterfactuals*. Malden, MA: Blackwell.

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow. *Noûs*, 13, 455–476.

Lillehammer, H. (2007). *Companions in Guilt: Arguments for Ethical Objectivity*. London: Palgrave Macmillan.

Malle, B. F., and Bennett, R. E. (2002). People's Praise and Blame for Intentions and Actions: Implications of the Folk Concept of Intentionality. *Technical Reports of the Institute of Cognitive and Decision Sciences, No. 02-2, Eugene, Oregon*.

Mandelbaum, E., and Ripley, D. (2012). Explaining the Abstract/Concrete Paradoxes in Moral Psychology: The NBAR Hypothesis. *Review of Philosophy and Psychology*, 3 (3), 351-368.

Margolis, E. (1999). How to Acquire a Concept. In E. Margolis & S. Laurence (eds.) *Concepts: Core Readings*, 549-567, Cambridge, MA: MIT Press.

McCann, H. (1998). *The Works of Agency: On Human Action, Will, and Freedom*. Ithaca, NY: Cornell University Press.

McKenna, M. (2009a). Compatibilism and Desert: Critical Comments on *Four Views on Free Will*. *Philosophical Studies*, 144, 3-13.

McKenna, M. (2009b). *Compatibilism*. The Stanford Encyclopedia of Philosophy, accessed at: http://plato.stanford.edu/entries/compatibilism/#3 .

Mele, A. (2006). *Free Will and Luck*. New York: Oxford University Press.

Moll, J., de Oliveriera-Souza, R., Eslinger, P.J., Bramati, I.E., Mourao-Miranda, J., Andreiuolo, P.A., etl al. (2002). The Neural Correlates of Moral Sensitivity: A Functional Magnetic Resonance Imaging Investigation of Basic and Moral Emotions. *Journal of Neuroscience*, 22, 2730-2736.

Moore, G.E. (1898). Freedom. *Mind*, 7(26), 179-204.

Murphy, G. (2002). *The Big Book of Concepts*. Cambridge, MA: MIT Press.

Nadelhoffer, T. (2004). Praise, Side Effects, and Intentional Action. *Journal of Theoretical and Philosophical Psychology*, 24, 259-269.

Nadelhoffer, T. (2005). Skill, Luck, and Intentional Action. *Philosophical Psychology*, 18, 343-354.

Nadelhoffer, T. (2008). Bad Acts, Blameworthy Agents, and Intentional Actions: Some Problems for Juror Impartiality. In J. Knobe & S. Nichols (eds.) *Experimental Philosophy*, 149-167, Oxford: University Press.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2005). Surveying Freedom: Folk Intuitions about Free will and Moral Responsibility. *Philosophical Psychology*, 18(5), 561-584.

Nahmias (2006). Folk Fears about Freedom and Responsibility: Determinism vs. Reductionism. *Journal of Cognition and Culture*, 6(1-2), 215-237.

Nahmias, E., Morris, S., Nadelhoffer, T., and Turner, J. (2006). Is Incompatibilism Intuitive? *Philosophy and Phenomenological Research*, 73(1), 28-53.

Nahmias, E. Coates, D.J., and Kvaran, T. (2007). Free Will, Moral Responsibility, and Mechanism: Experiments on Folk Intuitions. *Midwest Studies in Philosophy*, 31, 214-242.

Nahmias, E. (2011). Intuitions about Free Will, Determinism, and Bypassing. In R. Kane (ed.) *The Oxford Handbook of Free Will, 2nd Edition*, 554-577, New York: Oxford University Press.

Nelkin, D. (2005). Freedom, Responsibility, and the Challenge of Situationism. *Midwest Studies in Philosophy*, 29, 181-206.

Nelkin, D. (2007). Do We Have a Coherent Set of Intuitions About Moral Responsibility? *Midwest Studies in Philosophy*, 31, 243-259.

Nelkin, D. (2008). Responsibility and Rational Abilities: Defending an Asymmetrical View. *Pacific Philosophical Quarterly*, 89, 497-515.

Nichols, S. (2006). Folk Intuitions on Free Will. *Journal of Cognition and Culture*, 6(1&2), 57-86.

Nichols, S. (2004). The Folk Psychology of Free Will: Fits and Starts. *Mind & Language*, 19(5), 473-502.

Nichols, S. (2007). After Incompatiblism: A Naturalistic Defense of the Reactive Attitudes. *Philosophical Perspectives*, 21, 405-428.

Nichols, S., and Knobe, J. (2007). Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions. *Noûs*, 41(4), 663-685.

Nichols, S., and Roskies, A. (2008). Bringing Moral Responsibility Down to Earth. *The Journal of Philosophy*, 105, 371-388.

Nichols, S., Stich, S., and Weinberg, J. (2003). Meta-Skepticism: Meditations in Ethno-Methodology. In S. Luper (ed.) *The Skeptics*, 227-247, Aldershot, England: Ashgate.

Nichols, S., and Ulatowski, J. (2007). Intuitions and Individual Differences: The Knobe Effect Revisited. *Mind & Language*, 22(4), 346-365.

Nolan, D. (1997). Impossible Worlds: A Modest Approach. *Notre Dame Journal of Formal Logic*, 38(4), 535-572.

Nozick, R. (1981). *Philosophical Explanations*. Cambridge, MA: Belknap Press.

O'Connor, T. (1995). Agent Causation. In T. O'Connor (ed.) *Agents, Causes, and Events: Essays on Indeterminism and Free Will*, 173–200, New York: Oxford University Press.

O'Connor, T. (2000). *Persons and Causes: The Metaphysics of Free Will*. New York: Oxford University Press.

O'Connor, T. (2005). Freedom with a Human Face. *Midwest Studies in Philosophy*, 29, 207-227.

O'Connor, T. (2009). Agent-Causal Power. In T. Handfield (ed.) *Dispositions and Causes*, Oxford: University Press.

O'Conner, T. (2011). Agent-Causal Theories of Freedom. In R. Kane (ed.) *The Oxford Handbook of Free Will 2nd Edition*, 309-328, Oxford: University Press.

Pereboom, D. (2001). *Living Without Free Will*. Cambridge: University Press.

Pereboom, D. (2009a). Hard Incompatibilism and Its Rivals. *Philosophical Studies*, 144, 21-33.

Pereboom, D. (2009b). Free Will, Love, and Anger. *Ideas Y Valores*, 141, 169-189.

Piaget, J. (1965). *The Moral Judgment of the Child*. M. Gabain (Trans.). New York: Free Press. (Original work published 1932).

Pink, T. (2004). *Free Will: A Very Short Introduction*. Oxford: University Press.

Pizarro, D., Uhlmann, E., and Salovey, P. (2003). Asymmetry in Judgments of Moral Blame and Praise: The Role of Perceived Metadesires. *Psychological Science*, 14(3), 267-272.

Prinz, J. (2008). Empirical philosophy and experimental philosophy. In.J. Knobe & S. Nichols (eds.) *Experimental Philosophy*, 189-208, Oxford, University Press.

Putnam, H. (1981). *Reason, Truth and History*. Cambridge: University Press.

Putnam, H. (1993). Objectivity and the Science-Ethics Distinction. In M. Nussbaum and A. Sen (eds.) *The Quality of Life*, 143-64, Oxford: University Press.

Putnam, H. (2002). *The Collapse of the Fact/Value Distinction*. Cambridge: Harvard University Press.

Sarkissian, H., Chatterjee, A., De Brigard, F., Knobe, J., Nichols, S., and Sirker, S. (2010). Is Belief in Free Will a Cultural Universal? *Mind & Language*, 25(3), 346-358.

Schaich Borg, J., Hynes, C., Van Horn, J., Grafton, J.S.T., and Sinnott-Armstrong, W. (2006). Consequences, Action and Intention as Factors in Moral Judgments: An fMRI Investigation. *Journal of Cognitive Neuroscience*, 18, 803-817.

Sinnott-Armstrong, W. (2008). Abstract + Concrete = Paradox. In J. Knobe and S. Nichols (eds.) *Experimental Philosophy*, 209-230, Oxford: University Press.

Small, D., and Loewenstein, G. (2005). The Devil You Know: The Effects of Identifiability on Punishment. *Journal of Behavioral Decision Making*, 18, 311-318.

Smart, J.J.C. (1961). Free Will, Praise, and Blame. *Mind*, 70, 291-306.

Smilansky, S. (2000). *Free Will and Illusion*. Oxford: University Press.

Smith, M. (2003). Rational Capacities, or: How to Distinguish Recklessness, Weakness, and Compulsion. In S. Stroud and C. Tappolet (eds.) *Weakness of Will and Practical Irrationality*, 17-33, Oxford: University Press.

Sommers, T. (2010). Experimental Philosophy and Free Will. *Philosophy Compass*, 5(2), 199-212.

Strawson, G. (1986). Freedom and Belief. Oxford: University Press.

Strawson, G. (1993a). On Freedom and Resentment. In J.M. Fischer and M. Ravizza (eds.) *Perspectives on Moral Responsibility*, 67-100, Ithaca: Cornell University Press.

Strawson, G. (1993b). The Impossibility of Moral Responsibility. *Philosophical Studies*, 75, 5-24.

Strawson, P.F. (1962). Freedom and Resentment. In P.F. Strawson (ed.), *Studies in the Philosophy of Thought and Action*, 71-96, Oxford: University Press.

Van Inwagen, P. (1983). *An Essay on Free Will*. Oxford: Clarendon.

Van Inwagen, P. (2008). How to Think About the Problem of Free Will. *Journal of Ethics*, 12, 327-341.

Vargas, M. (2005). The Revisionist's Guide to Responsibility. *Philosophical Studies*, 125(3), 399-429.

Vargas, M. (2009). Revisionism About Free Will: A Statement and Defense. *Philosophical Studies*, 144(1), 45-62.

Vargas, M. (2011). Revisionist Accounts of Free Will: Origins, Varieties, and Challenges. In R. Kane (ed.) *The Oxford Handbook of Free Will 2nd Edition*, 457-474, Oxford: University Press.

Vargas, M. (2013). *Building Better Beings: A Theory of Moral Responsibility*. Oxford: University Press.

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

Wallace, R.J. (1994). *Responsibility and the Moral Sentiments*. Cambridge, MA: Harvard University Press.

Walster, E. (1966). Assignment of Responsibility for an Accident. *Journal of Personality and Social Psychology*, 3, 73-79.

Warmke, B. (2011). Moral Responsibility Invariantism. *Philosophia*, 39, 179-200.

Watson, G. (1987). Free Action and Free Will. *Mind*, 96, 154-172.

Wegner, D. (2002). *The Illusion of Conscious Will*. Cambridge, MA: MIT Press.

Williamson, T. (2007). *The Philosophy of Philosophy*. Oxford: Blackwell.

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

Woolfolk, R., Doris, J., & Darley, J. (2006). Identification, Situational Constraint, and Social Cognition: Studies in the Attribution of Moral Responsibility. *Cognition*, 100, 283-301.

KELLY ANNE MCCORMICK

Ph.D. Candidate, Syracuse University

CONTACT INFORMATION:

Address: 126 Jamesville Ave Apt L3, Syracuse, NY 13210
Phone: 518.859.4747
Email: kamcco02@syr.edu
Website: https://sites.google.com/site/kellyannemccormick47/

AREA OF SPECIALIZATION:

Ethics

AREAS OF COMPETENCE:

Metaphysics, Epistemology, Logic, Moral Psychology, Applied Ethics (Environmental, Media)

EDUCATION:

June 28, 2013	PhD	Syracuse University
2006	BA	Colgate University

DISSERTATION:

Title: *Towards a Revisionist Account of Moral Responsibility*

In my dissertation I raise a new problem for revisionism, the *normativity-anchoring problem*. The heart of this problem is that the methodological commitments used to motivate revisionism and distinguish the view from conventional theorizing about moral responsibility make it uniquely difficult for revisionists to justify our continued participation in the practice of moral praising and blaming. Solving this problem is necessary in laying the groundwork for any successful revisionist account. I argue that revisionists can ultimately avoid it by appeal to a principled difference between certain kinds of judgments about responsibility, and that we are justified in taking a particular class of these judgments to have a privileged epistemic status in our responsibility theorizing.

Committee: André Gallois (advisor), Ben Bradley, Mark Heller, Derk Pereboom

PAPERS UNDER REVIEW:

"Anchoring a Revisionist Account of Moral Responsibility"
"Responsibility and Justification: Are Our Theories Held Hostage to the Compatibility Question?"

PRESENTATIONS:

2011	"Anchoring a Revisionist Account of Moral Responsibility," Creighton Club: New York Philosophical Association Meetings, November 2011
2011	"Anchoring a Revisionist Account of Moral Responsibility," ABD Workshop, Syracuse University Philosophy Department, October 2011

2009 "A Causal Integrationist Response to Pereboom," Rocky Mountain Philosophy Conference, March 2009

TEACHING EXPERIENCE:
(*large lecture with TA mentoring responsibilities)

PHI 383: Free Will (Summer 2013 - online)

PHI 192: Introduction to Moral Theory (Spring 2013)

*PHI 251: Introduction to Logic (Fall 2011, Fall 2012)

PHI 109: Introduction to Philosophy, Honors (Spring 2012)

PHI 107: Theories of Knowledge and Reality (Fall 2009, Summer 2010, Spring 2011, Summer 2011, Summer 2012)

PHI 251: Introduction to Logic (Fall 2010, Summer 2012)

TEACHING ASSISTANT EXPERIENCE:

PHI 293: Ethics and Media Studies (Spring 2010)

PHI 251: Introduction to Logic (Spring 2008, Fall 2008, Spring 2009)

PHI 191: Ethics and Contemporary Issues (Fall 2007)

HONORS & AWARDS:

2011 Graduate Student Paper Award, Creighton Club: New York Philosophical Association Meetings

2011 Outstanding TA Award (Syracuse)

2011 Departmental Summer Research Fellowship (Syracuse)

2009 Departmental Summer Research Fellowship (Syracuse)

2008 Departmental Summer Research Fellowship (Syracuse)

2006 M. Holmes Hartshorne Memorial Award for Excellence in Philosophy (Colgate)

PROFESSIONAL ACTIVITY & SERVICE:

2013 Curriculum development, Onondaga Community College; consultant on grant proposal and syllabus construction for PHI 108 (Environmental Ethics), as part of the Sustainability Program (with Dave Bzdack and Patrick Denny)

2012 Graduate student organizer for SPAWN (Syracuse Philosophy Annual

	Workshop and Network), "Normative Realism," August 2012
2012	Chair at the Pacific APA for Stephen Morris, "Vargas-style Revisionism and the Problem of Desert," Comments by Joseph Keim Campbell
2011-2012	Philosophy Department Representative, Syracuse University Graduate Student Organization
2010	Co-organizer, Women in Philosophy Group, Syracuse University Philosophy Department (with Kirsten Egerstrom and Sarah Morales)
2009-2010	Graduate Student President, Syracuse University Philosophy Department
2009	Co-organizer, Syracuse University Philosophy Graduate Conference, April 2009 (with Aaron Wolf)

GRADUATE COURSEWORK:
(* indicates courses audited)

Syracuse University:

*Environmental Ethics	B. Bradley (Fall 2012)
*Concepts	K. Edwards (Fall 2011)
Contextualism	M. Heller (Fall 2009)
Independent Study: Responsibility & Identity	A. Gallois (Spring 2009)
Analytic Ethics	B. Bradley/K. McDaniel (Spring 2009)
Ancient Philosophy: Metaphysics	J. Roberston (Spring 2009)
Free Will	M. Heller (Fall 2008)
Natural Kinds	B. Nanay (Fall 2008)
Critical Theory	K. Baynes (Fall 2008)
Independent Study: Reasons	B. Bradley (Summer 2008)
Identity, Time, & Consciousness	A. Gallois (Spring 2008)
Epistemology	M. Heller (Spring 2008)
Early Modern: The Empiricists	A. Gallois (Spring 2008)
Kant's Ethics	E. Garcia (Fall 2007)
Logic & Language	M. Brown (Fall 2007)
Metaphysics of Death	B. Bradley (Fall 2007)

Cornell University:

*Free Will	D. Pereboom (Spring 2011)

University of Colorado at Boulder:

Proseminar: Ethics	C. Heathwood (Spring 2007)
Environmental Ethics	B. Hale (Spring 2007)
Kant's Critique of Pure Reason	B. Hanna (Spring 2007)
Proseminar: Causation	C. Cleland (Fall 2006)
Intentional Logic	D. Belcher (Fall 2006)

REFERENCES:

André Gallois
Professor of Philosophy
Syracuse University
agallois@syr.edu

Derk Pereboom
Professor of Philosophy
Cornell University
dp346@cornell.edu

Mark Heller
Professor of Philosophy
Syracuse University
heller@syr.edu

Manuel Vargas
Professor of Philosophy and Law
University of San Francisco
mvargas@usfca.edu

Ben Bradley
Associate Professor of Philosophy
Department Chair
Syracuse University
wbradley@syr.edu

Thomas McKay
Professor of Philosophy
Syracuse University
tjmckay@syr.edu